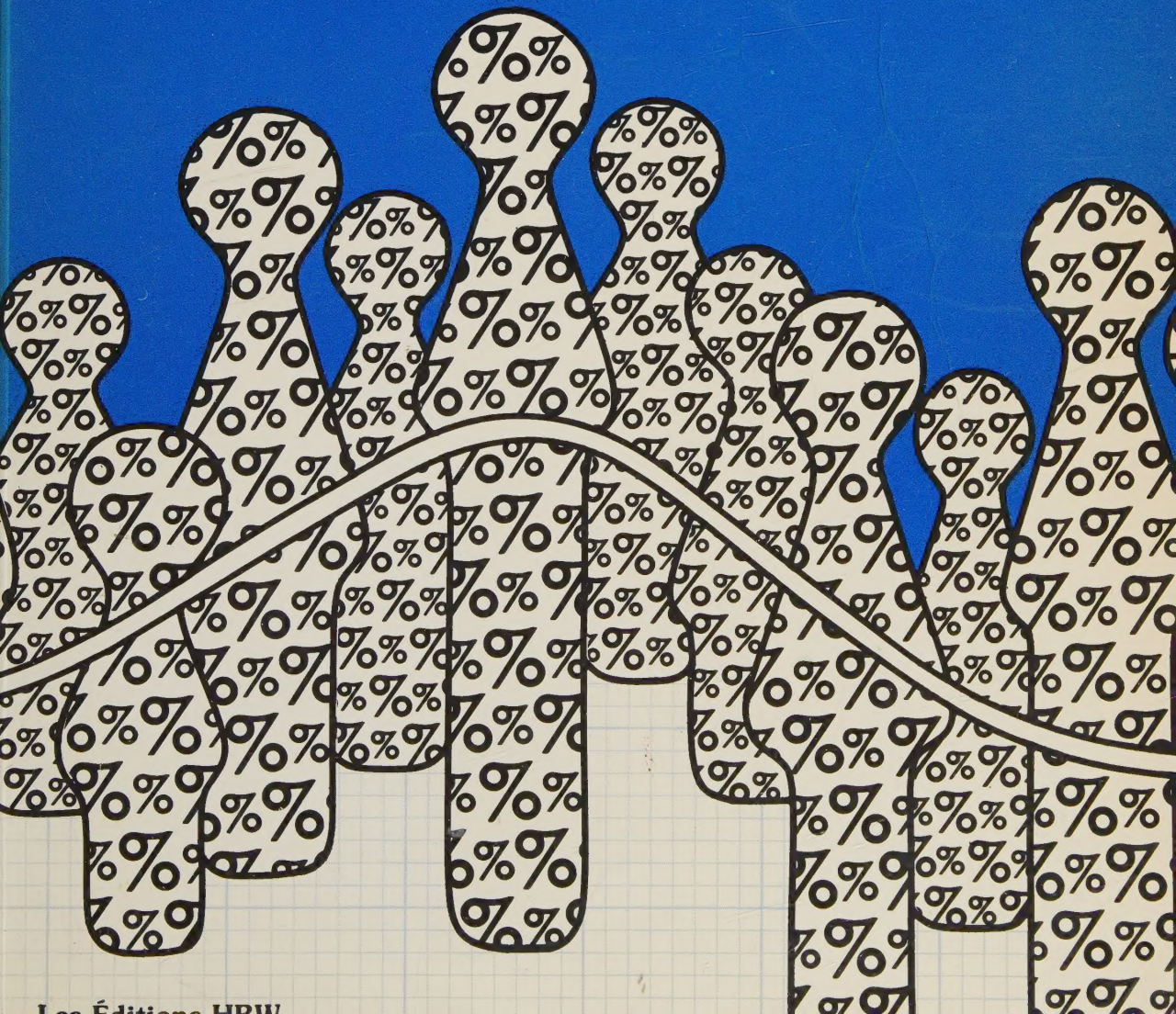
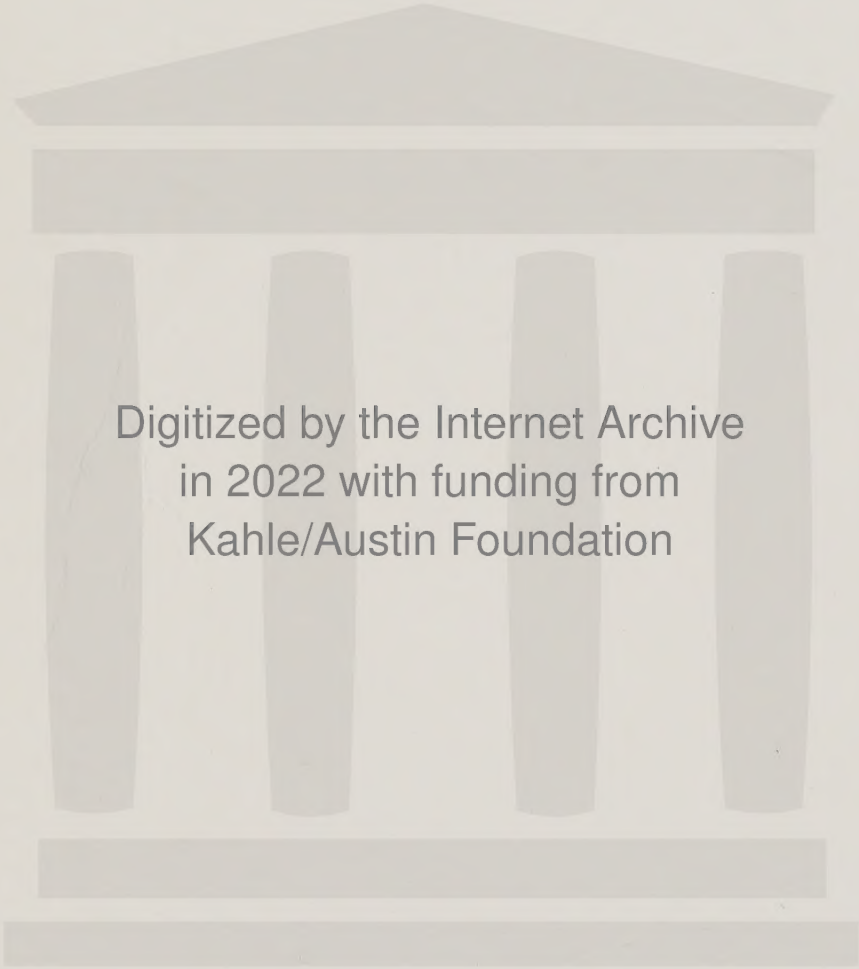
statistiques

gilbert

traduit et adapté par jean-guy savard



Les Éditions HRW


Digitized by the Internet Archive
in 2022 with funding from
Kahle/Austin Foundation

https://archive.org/details/statistiques0000gilb

statistiques



Norma Gilbert, Ph. D.
Département de mathématiques Université Drew

Traduit et adapté par
Jean-Guy Savard
Département de mathématiques Collège Marie-Victorin

Les Éditions HRW Ltée, Montréal

en collaboration avec
W.B. Saunders Company Philadelphie, Londres, Toronto

Statistiques

Traduction et adaptation de:
Statistics de Norma Gilbert, Ph. D.
Copyright © 1976 by the W.B. Saunders Company

Par:
Jean-Guy Savard
Département de mathématiques
Collège Marie-Victorin

Maquette de la couverture: Danielle Guy-Grandmont

Tous droits réservés
Copyright © 1978 by
Les Éditions HRW Ltée, Montréal



Les Éditions HRW Ltée
8035, rue Jarry est
MONTRÉAL, (Québec) H1J 1H6

ISBN: 0-03-929343-2

Dépôt légal 1er trimestre 1978
Bibliothèque nationale du Québec

Il est illégal de reproduire une partie quelconque de ce livre sans autorisation de la maison d'édition. La reproduction de cette publication, par n'importe quel procédé, sera considérée comme une violation du copyright.

Imprimé au Canada

5 6 7 8 LLC 88 87 86 85

Préface du traducteur

Une première précision s'impose: ce livre n'est pas une traduction, mais surtout et avant tout une adaptation du livre américain *Statistics* par Norma Gilbert. Tout en respectant la démarche pédagogique de l'auteur, j'ai tenu compte de la situation politico-socio-économique du Québec. De plus, toutes les mesures ont été transformées au système international d'unités.

Revenons à l'aspect pédagogique:

Pour les étudiants de **sciences humaines,** on trouve un très grand nombre de situations prises dans les sphères d'activités suivantes: psychologie, sociologie, criminologie, politique, éducation, justice.

Pour l'étudiant de **sciences administratives,** il y a beaucoup de problèmes à caractère économique. Pour l'étudiant de **sciences de la santé,** beaucoup de problèmes de médecine, de pharmacologie, de biologie, de chimie et d'agriculture.

Pour le futur ingénieur, l'ouvrage offre un peu de tout, mais surtout une ouverture vers un monde qui lui semblera peut-être un peu plus humain.

Ce livre se prête bien à différentes méthodologies: enseignement individualisé, ateliers, cours magistraux.

Ce livre n'exige pas de grandes connaissances mathématiques et peut être utilisé dès le début du collégial. Il ne fait pas appel au calcul différentiel et intégral. Dans la mesure du possible, il serait bon que l'étudiant puisse faire auparavant le cours de mathématiques 101; un étudiant dont le profil ne comporte pas de mathématiques peut très bien suivre ce cours à titre complémentaire.

Comme il est mentionné dans la préface de l'auteur, le cours de statistiques de 75 heures devrait comprendre normalement les chapitres 1 à 10; de plus, pour les étudiants de sciences humaines, on pourrait ajouter le chapitre 13 (test du χ^2). Pour les étudiants de sciences administratives et de sciences pures, on pourrait ajouter les chapitres 13 et 15 (régression et corrélation).

Ce livre est un premier cours de statistiques et ne sombre pas dans le formalisme; il contient seulement les démonstrations qui font appel à des transformations algébriques simples; celles qui

demandent une technique plus poussée sont faites intuitivement et justifiées par de nombreux exemples.

Ce livre insiste beaucoup sur la compréhension des situations, c'est sûrement sa principale qualité.

Jean-Guy Savard

Préface de l'auteur

Ce livre est destiné aux étudiants qui, sans avoir des aptitudes particulières pour les mathématiques, ont besoin de comprendre comment une décision statistique est prise. (Ce cours peut être suivi par un étudiant qui a complété un cours secondaire.) Plusieurs textes de statistiques sont présentés aux étudiants avec plus ou moins de bonheur; mis à part le plaisir que j'ai eu à écrire celui-ci, voici les principales caractéristiques de cet ouvrage.

Premièrement, un livre de statistiques contient beaucoup de formules. Plusieurs livres insistent sur leur utilisation. Il est important de savoir les utiliser, mais j'ai surtout essayé d'expliquer la provenance de la formule, pourquoi on doit l'utiliser et quand on ne doit pas l'utiliser. Souvent, les explications sont intuitives et non formelles, mais il est plus important qu'un étudiant sache pourquoi une décision est prise plutôt que d'apprendre seulement quoi chercher.

Deuxièmement, dans plusieurs ouvrages, la notation et le symbolisme sont très développés. Dans celui-ci, ce n'est pas le cas (sauf peut-être au chapitre 16; bien que les étudiants sont à ce moment familiers avec la statistique, certains professeurs préféreront ne pas voir ce chapitre, sans grand inconvénient d'ailleurs).

Troisièmement, ce livre est écrit pour permettre à la fois un enseignement individualisé ou un enseignement magistral: chaque titre de chapitre est suivi d'une présentation qui permet de situer le contexte et de préciser les points importants. Plus de 300 exemples sont faits en détail. À la fin de chaque chapitre, il y a une liste de «vocabulaire et symboles» qui permet une révision rapide puis une série de problèmes (plus de 425) dont les numéros impairs sont solutionnés à la fin de chaque chapitre, suivant le principe qu'il vaut mieux comprendre la solution que d'avoir la bonne réponse.

Finalement, certains textes omettent les probabilités, mais il est impossible de bien comprendre les statistiques sans leur utilisation. D'autres débutent par les probabilités ce qui pour plusieurs étudiants devient très difficile. Pour cette raison, j'ai débuté par la statistique descriptive et retardé l'étude de la probabilité jusqu'au chapitre 5; celle-ci est alors utilisée dans plusieurs des chapitres qui suivent.

Les chapitres *1 à 8, une partie du chapitre 9 et le chapitre 10 doivent normalement être vus dans l'ordre;* ce qui reste du chapitre 9 ainsi que les *chapitres 11 à 17 peuvent être choisis selon l'intérêt,* les aptitudes des étudiants et la longueur du cours.

Je suis particulièrement reconnaissante au professeur David Moore de l'Université Purdue qui a revisé le manuscrit à trois étapes différentes et qui m'a fourni des suggestions très valables. Un merci spécial à tous ceux qui ont lu et commenté le manuscrit: Stephen A. Book, California State College, Dominguez Hills, California; Albert Liberi, Westchester Community College, Valhalla, New York; Stanley M. Lukawecki, Clemson University, Clemson, South Carolina; Leon Gleser, Purdue University, West Lafayette, Indiana; Judith Tanur, State University of New York at Stony Brook, Stony Brook, New York.

Je suis également reconnaissante à Margo Hall (Drew University, Madison, New Jersey, 1974) qui a contribué à la rédaction, à John Miano (Drew University, 1975) qui m'a suggéré plusieurs problèmes intéressants, à tous les étudiants de Mathématique 3 de l'université Drew qui ont utilisé la première édition du manuscrit et à John Snyder de W.B. Saunders Company pour sa coopération et son encouragement.

Norma M. Gilbert

TABLE DES MATIÈRES

Préface du traducteur III
Préface de l'auteur IV

Chapitre 1

Quels sont les prérequis mathématiques? 1

1.1 Introduction 1
1.2 Pré-test 2
1.3 Priorité des opérations 3
1.4 La notation Σ 4
1.5 Formules 6
1.6 Carrés et racines carrées 8
1.7 Les lettres grecques 9
1.8 Vocabulaire et symboles 9
1.9 Exercice 9

Chapitre 2

Introduction à la statistique 13

2.1 Qu'est-ce que la statistique? 13
2.2 Populations et échantillons 14
2.3 Types de données 15
2.4 Tableau de données 17
2.5 Comment regrouper les résultats 19
2.6 Représentation graphique des données 21
2.7 Vocabulaire et symboles 24
2.8 Exercices 24

Chapitre 3

Mesures de tendance centrale 30

3.1 Introduction 30
3.2 Le mode 30

3.3 La médiane 31
3.4 La moyenne de données non groupées 32
3.5 La moyenne de données regroupées 35
3.6 Quartiles, centiles et rangs centiles 38
3.7 Vocabulaire et symboles 39
3.8 Exercices 40

Chapitre 4

Mesure de dispersion

4.1 Introduction 46
4.2 Le rapport de variation 46
4.3 L'étendue 47
4.4 L'intervalle semi-interquartile Q 48
4.5 Critères pour une mesure de dispersion 48
4.6 Variance et écart-type de données non groupées 49
4.7 Variance et écart-type de données groupées 53
4.8 Dissymétrie 56
4.9 Vocabulaire et symboles 57
4.10 Exercices 57

Chapitre 5

Probabilité 62

5.1 Pourquoi étudier la probabilité? 62
5.2 Définition de la probabilité 63
5.3 Espace échantillonnal et événement 65
5.4 Événements mutuellement exclusifs 70
5.5 Axiomes de probabilités 73
5.6 p (A ou B) si A et B ne sont pas mutuellement exclusifs 74
5.7 Événements indépendants p (A et B) 76
5.8 Vocabulaire et symboles 80
5.9 Exercices 80

Chapitre 6

Distributions de probabilités (cas discret) 87

6.1 Distributions de probabilités 87
6.2 Pré-test sur la permutation et les combinaisons 90
6.3 Permutations 90
6.4 La notation factorielle 92
6.5 Combinaisons 93
6.6 Hypothèses de la distribution binômiale 95
6.7 p (X succès en n tentatives) 95
6.8 Pourquoi l'expression «distribution binômiale»? 98
6.9 La moyenne et l'écart-type d'une distribution binômiale 98
6.10 Applications de la distribution binômiale 99
6.11 Vocabulaire et symboles 99
6.12 Exercices 99

Chapitre 7

Distributions normales et résultats standardisés 105

7.1 Histogramme d'une distribution binômiale 105
7.2 Courbes normales 108

7.3 La courbe normale centrée, réduite 109
7.4 L'aire sous la courbe normale centrée, réduite 110
7.5 Pourcentages des résultats compris entre 1; 2; 3 fois l'écart-type de la moyenne 115
7.6 Moyenne et écart-type d'une population 116
7.7 Variable centrée et réduite 118
7.8 Utilisation des résultats centrés et réduits 121
7.9 Distributions normales en général 122
7.10 Aire sous une courbe normale 123
7.11 Approximation d'une distribution binômiale par une loi normale 126
7.12 Vocabulaire et symboles 127
7.13 Exercices 127

Chapitre 8

Comment choisir un échantillon 133

8.1 Préparation de l'expérience statistique 133
8.2 Échantillons aléatoires 135
8.3 Autres méthodes d'échantillonnage 136
8.4 Vocabulaire 137
8.5 Exercices 137

Chapitre 9

Intervalle de confiance pour les moyennes 139

9.1 Échantillonnage avec et sans remise 139
9.2 Distribution d'échantillonnage des moyennes 140
9.3 Théorème central limite 144
9.4 Utilisation de la distribution d'échantillonnage des moyennes 144
9.5 Inférence statistique 146
9.6 Estimation de la moyenne de la population quand σ est connu 146
9.7 Estimateurs biaisés et efficaces 150
9.8 Intervalle de confiance pour μ quant σ est inconnu et $n \geqslant 30$ 151
9.9 Vocabulaire et symboles 151
9.10 Exercices 152

Chapitre 10

Intervalles de confiance de proportions, différences de moyennes 157

10.1 Distribution d'échantillonnage de proportions 158
10.2 Intervalle de confiance pour une proportion 161
10.3 Distribution d'échantillonnage des différences de moyennes 163
10.4 Utilisation de la distribution d'échantillonnage des différences de moyennes 165
10.5 Distribution d'échantillonnage de différences de proportions 168
10.6 Résumé des formules utilisées dans l'inférence statistique 171
10.7 Taille de l'échantillon 171
10.8 Taille de l'échantillon pour estimer la moyenne d'une population 172
10.9 Taille de l'échantillon avec des proportions 173
10.10 Taille de l'échantillon pour une différence de moyennes 175
10.11 Taille de l'échantillon pour une différence de proportions 175
10.12 Vocabulaire et symboles 177
10.13 Exercices 177

Chapitre 11

Vérification d'hypothèse 185

11.1 Introduction 185
11.2 Choix de l'hypothèse à vérifier (H_0) 186
11.3 Choix de l'hypothèse rivale 188
11.4 Règle de décision 189
11.5 Test unilatéral 195
11.6 Erreurs de 1re espèce et de 2e espèce 198
11.7 Avertissements 205
11.8 Vocabulaire et symboles 205
11.9 Exercices 205

Chapitre 12

Distribution *t* de Student 215

12.1 La nécessité de la distribution *t* 215
12.2 Caractéristiques de la distribution *t* 217
12.3 Quand doit-on utiliser une distribution *t*? 218
12.4 Degré de liberté 219
12.5 Table de la distribution *t* 219
12.6 Intervalle de confiance et test d'hypothèse concernant la moyenne 220
12.7 Différences de moyennes pour des échantillons indépendants 222
12.8 Démonstration de $\sigma_{(\bar{x} - \bar{y})}$ quand les populations ont des variances égales 225
12.9 Différence de moyennes de deux échantillons dépendants 226
12.10 Vocabulaire et symboles 228
12.11 Exercices 228

Chapitre 13

χ^2 (Khi — (carré) 235

13.1 Introduction 235
13.2 Distribution du χ^2 237
13.3 Test d'efficacité d'un ajustement 239
13.4 Relation et indépendance des variables; tableaux de contingences 241
13.5 Correction de Yates 248
13.6 Vocabulaire et symboles 249
13.7 Exercices 249

Chapitre 14

Analyse de variance 255

14.1 La distribution *F* 255
14.2 Utilisation de la table *F* 256
14.3 Utilisation de la distribution *F* dans les tests de variance 258
14.4 Introduction à l'analyse de variance 261
14.5 Méthode courte: soustraire une constance *A* de chaque donnée d'un échantillon 266
14.6* D'où viennent les formules? 267
14.7 Quelles moyennes sont différentes? 269
14.8 Vocabulaire et symboles 271
14.9 Exercices 271

Chapitre 15

Corrélation linéaire et régression **279**

15.1 Introduction 279
15.2 Graphique de la droite 280
15.3 La pente d'une droite 281
15.4 Équation d'une droite connaissant deux points 285
15.5 La droite de régression 286
15.6 Coefficient de corrélation: approche intuitive 292
15.7 Erreur type d'un estimé, variation totale et variation expliquée 294
15.8 Calcul de r 297
15.9 Quelques remarques 300
15.10 La relation entre la régression et la corrélation 302
15.11 Vocabulaire et symboles 302
15.12 Exercies 302

Chapitre 16

Inférence statistique: régression linéaire et corrélation **311**

16.1 Inférences sur la régression: hypothèses sur les populations 311
16.2 Vérifier H_0:B =0 312
16.3 Intervalles de confiance et de prédiction pour l'équation de régression 314
16.4 La distribution d'échantillonnage de r 320
16.5 Vérifier H_0 : ρ =0 321
16.6 Vérifier H_0 : $\rho = A$, $A \neq 0$ 323
16.7 Vocabulaire et symboles 325
16.8 Exercices 325

Chapitre 17

Tests non-paramétriques **329**

17.1 Introduction 329
17.2 Le test du signe 330
17.3 Le test des séquences «runs» 331
17.4 Le test de Mann-Whitney 333
17.5 Coefficient de corrélation de rang de Spearman 336
17.6 Vocabulaire et symboles 339
17.7 Exercices 339

Appendice A - Symboles 345
Appendice B - Formules 349
Appendice C - Tables 357
Index **382**

Chapitre 1

Quels sont les prérequis mathématiques?

*La plupart des connaissances contenues dans ce chapitre sont déjà acquises. Notez que le symbole * est utilisé de préférence aux symboles × ou •, ainsi: 2 * 3 = 6. Le symbole de sommation (Sections 1.4 et 1.5) est nouveau. Le contenu de ce chapitre sera très fréquemment utilisé tout au long du présent ouvrage, prenez donc soin de bien l'assimiler. Même si vous réussissez bien en mathématiques, il vous arrivera de douter de la priorité des opérations (l'ordre dans lequel doit être effectué les opérations numériques): (2 * 3 + 4 = 6 + 4 ou 2 * 7). Il en sera de même avec les parenthèses qui changeront l'ordre habituel de priorité.*

1.1 INTRODUCTION

Plusieurs étudiants abordent les statistiques avec appréhension, certains ont même pris la résolution de ne plus faire de mathématiques après leur cours secondaire. Ne perdez pas courage, et ce pour deux raisons. Les prérequis mathématiques dont vous aurez besoin sont remarquablement réduits. Durant vos études secondaires, on vous a appris à résoudre des équations quadratiques sans trop vous dire **à quoi** elles pouvaient bien servir. Beaucoup d'entre vous doivent poursuivre des études en mathématiques et plus spécialement en statistiques parce qu'elles sont un prérequis à la poursuite de vos études en psychologie, économie ou sociologie. Ceux-là trouveront — contrairement à leurs expériences passées — des applications immédiates de leurs nouvelles connaissances dans leur champ d'intérêt. Cette nouvelle et merveilleuse complicité avec les mathématiques vous les fera voir sous un nouvel angle.

Il faudra toujours essayer de comprendre la démarche qui produit un résultat et non accepter celui-ci aveuglément. Un ordinateur peut être programmé pour calculer le résultat d'une équation longue et complexe; il sera préférable de comprendre quand et pourquoi telle formule est utilisée. Occasionnellement, une explication ou une démonstration nécessitera l'utilisation du calcul différentiel; dans ce cas l'explication sera purement intuitive ou simplement supprimée. Une mise en garde sera donnée si de telles situations surviennent.

Il y aura très peu de symboles à mémoriser pour lire cet ouvrage. Une nouvelle notation nécessite des explications: la multiplication des nombres est désignée par: $*$

$$4 * 2 = 8$$

On utilise cette notation particulièrement parce qu'elle est utilisée sur les listes imprimées par ordinateur. La multiplication de x par y sera ordinairement indiquée par xy de préférence à $x * y$; également fd de préférence à $f * d$. Mais en aucun cas le point ($f \cdot d$) ou la croix ($f \times d$) ne seront utilisés pour indiquer la multiplication de f par d.

Noter attentivement que la multiplication d'un nombre par zéro donne 0:

$$7 * 0 = 0 \qquad 0 * 2/3 = 0$$
$$0 * 7 = 0 \qquad \sqrt{2} * 0 = 0$$
$$-3{,}8 * 0 = 0 \qquad 0 * x = 0$$

Un sigle spécial sera utilisé pour vous prévenir des pièges disposés à votre intention: difficultés particulières, simplifications spéciales, problèmes difficiles à résoudre. Ce sigle sera un gros point noir ● disposé près de la difficulté.

1.2 PRÉ-TEST

Si vous pouvez résoudre les problèmes suivants correctement, vous êtes très bien préparé pour ce cours. Si vous avez quelques difficultés avec certains problèmes, étudier ce chapitre avec attention. Les problèmes 13 à 20 nécessitent la connaissance de la notation de sommation Σ. Ne vous en faites pas si elle vous est inconnue, lire la section 1.4 page 4 avec attention. Les réponses sont fournies au bas de la page 3.

1. $4{,}38 + 0{,}12 + 0{,}02 =$
2. $48 - 51 =$
3. $6{,}21 * 0{,}01 =$
4. $5 * 0 =$
5. $\dfrac{27 * 75}{25 * 9} =$
6. $3 * 5^2 + 5 =$
7. $3 * (5^2 + 5) =$
8. $4 - 10^2/20 =$
9. Évaluer $x^2y + \dfrac{z}{y}$ pour $x = 2, y = 3, z = 6$
10. Trouver $\overline{X}$ pour $X = 10, s = 2, n = 25, z = 15$ si $z = \dfrac{X - \overline{X}}{\dfrac{s}{\sqrt{n}}}$.
11. $\sqrt{493} =$ (à 2 décimales près; utiliser la table 1, appendice C)
12. $\sqrt{0{,}00212} =$ (à 3 décimales près; table 1, appendice C)

Dans les problèmes 13 à 17, les valeurs de X sont 3; 5; 8; 2; 0; 3; 0; les valeurs correspondantes de Y sont 2; 1; 3; −1; 0; 2; 4.

13. $\Sigma X =$

14. $\Sigma(X - Y) =$

15. $\Sigma X^2 =$

16. $(\Sigma X)^2 =$

17. $\Sigma(X - 2) =$

18.

Y	f	$\Sigma fY =$
3	2	
4	7	
5	2	

19. Un ensemble de 5 valeurs de X est donné. Trouver une formule qui permettra de: «Soustraire 3 de chaque valeur de X, élever chaque nouvelle valeur au carré, additionner le nouvel ensemble et finalement prendre la racine carrée de la somme».

20. Un ensemble de 10 valeurs de Y est donné. Traduire littéralement les opérations contenues dans la formule $\dfrac{\Sigma(2Y + 5)^2}{10}$.

1.3 PRIORITÉ DES OPÉRATIONS

Le mot «opération» désigne ici: élévation à une puissance, multiplication ou division, addition ou soustraction. Les règles admises par les mathématiciens quand il y a plus d'une opération à effectuer, sont celles données ci-après. — Voici:

(a) Premièrement effectuer les puissances.
(b) Puis les multiplications ou divisions.
(c) Finalement les additions ou soustractions.

Exemple 1

$2 * 3 + 4 = ?$

$2^2 * 3 + 1 = ?$

$3 - 12/2^2 = ?$

$2 * 3 + 4 = 6 + 4 = 10.$

$2^2 * 3 + 1 = 4 * 3 + 1 = 12 + 1 = 13.$

$3 - 12/2^2 = 3 - 12/4 = 3 - 3 = 0.$

Réponses:

1) 4,52; 2) −3; 3) 0,0621; 4) 0; 5) $\dfrac{3 * 3}{1 * 1} = 9$; 6) 80; 7) 90; 8) −1; 9) 14; 10) 4; 11) 22,20; 12) 0,046; 13) 21; 14) $\Sigma X - \Sigma Y = 21 - (2 + 1 + 3 - 1 + 0 + 2 + 4) = 21 - 11 = 10$; 15) $9 + 25 + 64 + 4 + 0 + 9 + 0 = 111$; 16) $21^2 = 441$; 17) $\Sigma X - \Sigma 2 = 21 - 7 * 2 = 7$; 18) $6 + 28 + 10 = 44$; 19) $\sqrt{\Sigma (X - 3)^2}$; 20) Multiplier chaque valeur de Y par 2, ajouter 5 à chaque résultat. Élever au carré chaque valeur du nouvel ensemble puis additionner tous ces résultats. Finalement diviser cette somme par le nombre de données soit 10.

Supposons que l'on ne veut pas procéder dans cet ordre. Si par exemple 2 et 3 doivent être additionnés puis le résultat élevé au carré: les parenthèses ( ), les crochets [ ] et les accolades { } viennent à notre secours et une nouvelle règle est ajoutée.

> S'il y a des parenthèses, crochets ou accolades, on effectue ce qu'il y a à l'intérieur avant de continuer.

La notation $3\,(4 + 2)^2$ est utilisée de préférence à $3 * (4 + 2)^2$.

Exemple 2

a) $(2+3)^2 * 4 = ?$

b) $(2^2 * 3 - 2)^2 * 4 = ?$

c) $2(3^2 + 4 \div 2) - 3(2^3 - 4)^2 = ?$

a) $(2+3)^2 * 4 = 5^2 * 4 = 25 * 4 = 100.$

b) $(2^2 * 3 - 2)^2 * 4 = (4 * 3 - 2)^2 * 4 = (12 - 2)^2 * 4 = 10^2 * 4 = 100 * 4 = 400.$

c) $2(3^2 + 4 \div 2) - 3(2^3 - 4)^2 = 2(9 + 4 \div 2) - 3(8 - 4)^2$
$= 2(9 + 2) - 3 * 4^2$
$= 2 * 11 - 3 * 16 = 22 - 48 = -26.$

1.4 LA NOTATION Σ

Vous connaîtrez bientôt sept lettres grecques. La première est le *S* majuscule appelé **sigma** et s'écrit Σ. Nous utiliserons Σ comme une abréviation du mot **somme** ou encore de la phrase «La somme de...».

Si les valeurs de *X* sont 1; 2; 3; 5; alors $\Sigma X = 1 + 2 + 3 + 5 = 11$. On additionne **toutes** les valeurs de *X*.

Si les poids *P* de cinq étudiants sont 58; 63; 60; 59; 61 kg alors $\Sigma P = 58 + 63 + 60 + 59 + 61 = 301$ kg.

Ne pas oublier les règles régissant la priorité des opérations, même si Σ est utilisé dans des expressions comportant des exposants, multiplications ou parenthèses.

Exemple 1

Si les valeurs de *Y* sont 4; 3; 1; 6; 2; Trouver (a) ΣY, (b) ΣY^2, (c) $(\Sigma Y)^2$.

(a) $\Sigma Y = 4 + 3 + 1 + 6 + 2 = 16.$

(b) $\Sigma Y^2 = 4^2 + 3^2 + 1^2 + 6^2 + 2^2 = 16 + 9 + 1 + 36 + 4 = 66.$

(c) $(\Sigma Y)^2 = 16^2 = 256.$

Exemple 2

Deux colonnes de nombres sont données dans le cadre ci-dessous. Trouver.

a) $\Sigma fX = ?$

b) $\Sigma fX^2 = ?$

c) $(\Sigma fX)^2 = ?$

d) $\Sigma(fX)^2 = ?$

f	X	fX	fX^2
2	−1	−2	2
3	0	0	0
1	1	1	1
2	2	4	8
		+3	11

a) ΣfX veut dire «Multiplier chaque valeur de f par la valeur correspondante de X, puis additionner les résultats». Une nouvelle colonne appelée fX est ajoutée près du cadre. Ainsi:

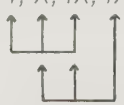
$$\Sigma fX = 2 * -1 + 3 * 0 + 1 * 1 + 2 * 2$$
$$= -2 + 0 + 1 + 4 = 3$$

b) $\Sigma fX^2 = \Sigma (fX) X = \Sigma X (fX)$, ainsi une autre colonne est ajoutée. On obtient celle-ci en multipliant les valeurs de la colonne X par les valeurs correspondantes de la colonne fX.

$$\Sigma fX^2 = \Sigma X(fX) = -1 * -2 + 0 * 0 + 1 * 1 + 2 * 4 = 2 + 0 + 1 + 8 = 11$$

Il aurait été possible d'élever X au carré puis de multiplier par f. Nous aurions obtenu les colonnes suivantes f, X, X^2, fX^2. Cette façon de procéder est à déconseiller. Premièrement parce que dans les calculs statistiques, on veut obtenir également fX. Le nouvel arrangement des colonnes serait alors f, X, fX, X^2, fX^2, ce qui nous oblige à multiplier des colonnes non-adjacentes, d'où possibilité d'erreurs. Nous préconisons donc f, X, fX, fX^2

c) $(\Sigma fX)^2 = ?$
$(\Sigma fX)^2 = (-2 + 0 + 1 + 4)^2 = 3^2 = 9.$



d) $\Sigma(fX)^2 = ?$
$\Sigma(fX)^2 = (-2)^2 + 0^2 + 1^2 + 4^2 = 4 + 0 + 1 + 16 = 21.$

Il y a trois règles à respecter quand on utilise le symbole de sommation.

Règle 1. $\Sigma(X + Y) = \Sigma X + \Sigma Y$. (Cette règle a un sens seulement si les valeurs affectées à X et Y sont en nombre égaux)

Règle 2. Si k est une constante, $\Sigma kX = k\Sigma X$.

Règle 3. Si A est une constante, $\Sigma A = nA$, où n est le nombre de valeurs à être additionnées.

Exemple 3

Valeurs de X: 4; 1; 6; 3; Valeurs correspondantes de Y: 2; 3; 7; -1. Trouver (a) $\Sigma (X + Y)$ (b) $\Sigma 2X$ (c) $\Sigma (X - 2)$ (d) $\Sigma 5$ (e) $\Sigma (X - Y)$

a) $\Sigma (X + Y) = \Sigma X + \Sigma Y = (4 + 1 + 6 + 3) + (2 + 3 + 7 - 1) = 14 + 11 = 25$ (Règle 1)

b) $\Sigma 2X = 8 + 2 + 12 + 6 = 28$
(par la règle 2, qui spécifie qu'une constante peut être mise en facteur)
$= 2 \Sigma X = 2 (4 + 1 + 6 + 3) = 2 * 14 = 28$

c) $\Sigma (X - 2) = (4 - 2) + (1 - 2) + (6 - 2) + (3 - 2) = 2 + (-1) + 4 + 1 = 6$ également $\Sigma (X - 2) = \Sigma X - \Sigma 2 = (4 + 1 + 6 + 3) - 4 * 2 = 14 - 8 = 6$

d) $\Sigma 5 = 4 * 5 = 20$ (Règle 3) le nombre de valeurs à additionner est donné par le contexte du problème, ici 4. Une notation plus rigoureuse, utilisant des indices pourrait être utilisée, mais ne le sera pas ici.

e) $$\begin{aligned}\Sigma(X - Y) &= \Sigma[X + (-1)Y] \\ &= \Sigma X + \Sigma(-1)Y \\ &= \Sigma X + (-1)\Sigma Y \\ &= \Sigma X - \Sigma Y \\ &= (4 + 1 + 6 + 3) - (2 + 3 + 7 - 1) \\ &= 14 - 11 \\ &= 3.\end{aligned}$$

1.5 FORMULES

Un certain nombre de formules seront utilisées dans ce cours; elles donnent une méthode pour effectuer des opérations. Étudier les exemples suivants pour apprendre comment traduire en français une expression mathématique et vice-versa.

Exemple 1

Traduire en français: $\Sigma(X - 4)$.

Soustraire 4 de chaque valeur de X; puis additionner les nouvelles valeurs.

Exemple 2

Écrire une formule permettant de soustraire 4 de chaque valeur de X, élever au carré chaque nouvelle valeur puis additionner ces carrés.

$$\Sigma(X - 4)^2$$

Exemple 3

Écrire une formule permettant de soustraire 4 de chaque valeur de X, d'additionner puis d'élever la somme au carré.

$$[\Sigma(X - 4)]^2$$

Essayer de résoudre les deux prochains exemples. Se rappeler de la priorité des opérations!

Exemple 4

Traduire $\Sigma(X + 3)$.

Exemple 5

Trouver une formule permettant de soustraire 5 de chaque valeur de X, élever au carré chaque nouvelle valeur ainsi obtenue, additionner les carrés puis diviser la somme par 3.

Réponses

4. Additionner 3 à chaque valeur de X puis additionner les nouvelles valeurs.

5. $\dfrac{\Sigma(X-5)^2}{3}$

S'il y a trois valeurs de X: 3; 5; 7 les opérations effectuées dans les exemples précédents sont les suivantes:

	Exemple 1	Exemple 2	Exemple 4	Exemple 5	Exemple 5
X	$X-4$	$(X-4)^2$	$X+3$	$X-5$	$(X-5)^2$
3	−1	1	6	−2	4
5	1	1	8	0	0
7	3	9	10	2	4
	3	11	24		8

Exemple 3: $[\Sigma(X-4)]^2 = 3^2 = 9$.

Exemple 5: $\dfrac{\Sigma(X-5)^2}{3} = \dfrac{8}{3}$.

Souvent vous aurez un certain nombre de valeurs X et un nombre correspondant de aleurs de f tel que dans le tableau ci-dessous. (f sera utilisé dans les chapitres suivants our désigner la fréquence avec laquelle les valeurs de X apparaissent.) Les quatre autres colonnes sont nécessaires dans les exemples 6 et 9.

(1)	(2)	(3)	(4)	(5)	(6)
X	f	fX	$X-5$	$f(X-5)$	$f(X-5)^2$
3	1	3	−2	−2	4
5	4	20	0	0	0
7	2	14	2	4	8
		37		2	12

Exemple 6

Traduire en français: $\Sigma fX = 37$.

Multiplier chaque valeur de X par le f correspondant puis additionner les produits; le résultat est 37 (colonne 3).

Exemple 7

Trouver une formule permettant de soustraire 5 de chaque valeur de X, multiplier le nouveau résultat par la valeur correspondante de f puis additionner les produits; le résultat est 2.

$$\Sigma f(X - 5) = 2$$

N.B.: le f utilisé ici représente la **fréquence** et ne doit en aucun cas être interprété comme la notation de fonction.

Essayer les deux exemples suivants:

Exemple 8

Traduire: $\Sigma f(X - 5)^2 = 12$.

Exemple 9

Trouver une formule permettant de soustraire 5 de chaque valeur de X, multiplier les nouvelles valeurs par le f correspondant, additionner les produits puis élever la somme au carré; le résultat est 4.

Réponses

8. Soustraire 5 de chaque valeur de X, élever au carré chaque nouveau résultat, multiplier par la valeur de f correspondante, additionner les produits; le résultat est 12 (colonne 6).
9. $[\Sigma f(X - 5)]^2 = 2^2 = 4$ (colonne 5)

1.6 CARRÉS ET RACINES CARRÉES

Dans les problèmes de statistique, il sera souvent nécessaire de trouver une racine carrée. Il faut se rappeler deux principes: (1) On ne peut extraire la racine carrée d'un nombre négatif. (2) La racine carrée n'est **jamais** négative; ainsi $\sqrt{4} = 2$ **non** ± 2 †.

Exemple 1

$\sqrt{34{,}5} =$ (2 décimales près)

Prendre 3,45 dans la colonne n de la table 1, appendice C; se déplacer horizontalement jusqu'à la colonne $\sqrt{10n}$, on trouve 5,873; arrondissant à 2 décimales on obtient $\sqrt{34{,}5} = 5{,}87$.

---

† *N.B.* Il ne faut pas confondre la solution de $x^2 = 4, x = \pm\sqrt{4} = \pm 2$ avec $\sqrt{4} = \pm 2$. Si vous n'avez pas de calculatrice, utiliser la table 1 appendice C pour trouver la racine carrée.

Exemple 2

$\sqrt{0{,}0126} =$ (3 décimales près)

Trouver 1,26 dans la colonne *n*. Comme 0,0126 est entre 0,0100 et 0,0999, lire le résultat dans la colonne $\sqrt{n}$, on obtient 1,122 en multipliant par 0,1 on obtient 0,1122 puis en arrondissant à 3 décimales: $\sqrt{0{,}0126} = 0{,}112$

Exemple 3

$134^2 = ?$ (table 1, appendice C)

Trouver 1,34 dans la colonne *n*, lire dans n^2, $1{,}34^2 = 1{,}7956$
comme $134 = 1{,}34 * 10^2$, on aura
$134^2 = 1{,}34^2 * (10^2)^2 = 1{,}34^2 * 10\,000$
$134^2 = 17\,956$

1.7 LES LETTRES GRECQUES

Vous avez déjà utilisé le S majuscule grec Σ. Les autres lettres que nous rencontrerons dans ce livre sont les minuscules suivantes

Lettre grecque	Nom	Équivalent français
μ	mu	m
σ	sigma	s
α	alpha	a
β	bêta	b
ρ	rho	r
χ	chi (prononcer kie)	--

1.8 VOCABULAIRE ET SYMBOLES

*

Sommation Σ
μ, σ, α, β, ρ, χ

1.9 EXERCICES

1. $\frac{1}{2} * \frac{3}{4} =$
2. $\frac{1}{2} + \frac{2}{3} =$
3. $26 - 34 =$
4. $3 - 4 - 6 + 2 + 4 - 3 + 6 =$
5. $4 * 3 + 5 =$
6. $6 + 2 * 3^2 =$
7. $(3^2 + 1)^2 * 0 =$
8. $(2^2 + 3)^2 * 2 =$
9. $(3^2 - 7)^3 * (4^2 - 7 * 2)^2 =$
10. $(\sqrt{16} - 3 * 2) * (2^2 - 5) =$

11. $(9 + 4^2) \cdot (3^2 - 5^2) =$

12. Évaluer $xy + z$, pour $x = 2{,}3$; $y = 3{,}0$ et $z = 4{,}1$

13. Évaluer $y\left(x + \frac{1}{yw}\right)$, si x = 2, y = 3 et w = 0

14. Évaluer $(-3)^2$

15. Évaluer -3^2

16. Soustraire + 4 de − 18

Dans les exercices 17-24, utiliser les valeurs suivantes pour X et Y;

X	Y
3	2
4	1
1	0
2	1
5	−3

17. $\Sigma X =$

18. $\Sigma X^2 =$

19. $(\Sigma X)^2 =$

20. $\Sigma (X - 2) =$

21. $\Sigma (X - Y) =$

22. $\sqrt{\frac{\Sigma(X-3)^2}{4}} =$ (1 décimale près)

23. $\Sigma XY =$

24. $(\Sigma X)(\Sigma Y) =$

Dans les exercices 25-29, utiliser les valeurs de y et f suivantes:

y	f
2	4
4	2
5	4

25. $\Sigma fy =$

26. $\Sigma fy^2 =$ additionner une colonne fy^2 au tableau précédent.

27. $(\Sigma fy)^2 =$

28. $\bar{y} = \frac{\Sigma fy}{\Sigma f} =$

29. Si $\overline{y} = 3{,}6$; $\Sigma f(y - \overline{y}) =$ (ajouter une colonne $y - \overline{y}$ à la table précédente.)

30. $\sqrt{844} =$ (1 décimale près)

31. $\sqrt{0{,}00132} =$ (3 décimales près)

32. Résoudre pour x: $3 = 2x - 3$

33. Résoudre pour x: $0{,}06x + 0{,}05(4\,000 - x) = 240$ (multiplier les deux membres de l'équation par 100)

34. Résoudre pour $\overline{X}$ si $z = 8$, $X = 3$, $s = 1$, et $n = 16$: $z = \frac{X - \overline{X}}{\frac{s}{\sqrt{n}}}$

35. Les résultats (X) de 10 étudiants à un test de personnalité sont les suivants: 22; 21; 16; 26; 23; 27; 23; 18; 23; 31.

a) Trouver une formule permettant d'additionner les résultats des 10 étudiants et de diviser la somme par le nombre d'étudiants; appeler le résultat $\overline{X}$ (qui se lit «X barre »).

b) Trouver une formule permettant de soustraire 23 de chaque résultat, élever au carré ces nouveaux résultats, additionner les carrés puis diviser la somme par le nombre de résultats moins 1. Prendre la racine carrée de l'expression puis appeler s le résultat.

c) Effectuer les calculs décrits en a et b à une décimale près.

36. Les grandeurs (H) en mètres de six étudiants sont 1,70; 1,60; 1,80; 1,65; 1,75; 1,85 m.

a) Trouver une formule permettant de soustraire 1,70 de chaque grandeur, additionner les nouvelles valeurs, diviser la somme par le nombre d'étudiants mesurés puis additionner 1,70 m au quotient, appeler le résultat $\overline{H}$.

b) Trouver une formule permettant de soustraire $\overline{H}$ de chaque grandeur, élever au carré chaque nouvelle valeur obtenue puis les additionner; diviser la somme par le nombre d'étudiants mesurés moins 1; prendre la racine carrée du résultat et appeler le résultat s.

c) Effectuer les calculs décrits en a et b à une décimale près.

37. Huit accusés ont été déclarés coupables et furent condamnés à des peines d'emprisonnement de 6; 10; 8; 1; 3; 1; 4 et 15 ans respectivement.

a) Traduire: $\overline{T} = \dfrac{\Sigma(T-7)}{8} + 7.$

b) Traduire: $s^2 = \dfrac{\Sigma(T-\overline{T})^2}{7}.$

c) Effectuer les calculs décrits en (a) et (b) à une décimale près.

Réponses

1. $\dfrac{3}{8}$
3. -8
5. 17
7. 0
9. $8 * 4 = 32$
11. $25 * -16 = -400$

13. Non-défini: $\dfrac{1}{0}$ en effet, on sait que $\dfrac{N}{D} = Q$ et $N = Q * D$ ainsi $\dfrac{8}{4} = 2$ car $8 = 2 * 4$ si on prend $\dfrac{1}{0} = 0$ par exemple, on aura $1 = 0 * 0$ ce qui n'a pas de sens $\dfrac{1}{0} = Q$ $1 = Q * 0 = 0$. On ne peut donc pas résoudre ce problème: $\dfrac{1}{0}$ n'est pas défini.

15. -9 17. 15

19. $15^2 = 225$.

21. $\Sigma X - \Sigma Y = 15 - 1 = 14$ or $1 + 3 + 1 + 1 + 8 = 14$

23. $6 + 4 + 0 + 2 - 15 = -3$ 25. $8 + 8 + 20 = 36$

27. $36^2 = 1296$.

29. $4 * -1{,}6 + 2 * 0{,}4 + 4 * 1{,}4 = 0$

31. 0,036 33) $X = 4\,000$

35. (a) $\overline{X} = \dfrac{\Sigma X}{10}$ (b) $s = \sqrt{\dfrac{\Sigma(X - 23)^2}{9}}$ (c) $\overline{X} = 23{,}0$; $s = \sqrt{\dfrac{168}{9}} = 4{,}3$

37. (a) On soustrait 7 de chaque valeur, on additionne les nouvelles valeurs, on divise la somme par 8 puis on ajoute 7 au quotient. Le résultat est appelé $\overline{T}$.

(b) $\overline{T}$ est soustrait de chaque valeur, la différence est élevée au carré, les carrés sont additionnés; la somme est divisée par 7 et le résultat appelé s^2.

(c) $\overline{T} = \dfrac{-8}{8} + 7 = 6.$

$$s^2 = \frac{0^2 + 4^2 + 2^2 + (-5)^2 + (-3)^2 + (-5)^2 + (-2)^2 + 9^2}{7} = \frac{164}{7} = 23{,}4$$

Chapitre 2

Introduction à la statistique

Un cours d'introduction à la statistique se divise en deux parties: la statistique descriptive et l'inférence statistique. Vous trouverez dans ce chapitre une brève description de ces deux parties; vous pourrez ainsi vous faire une bonne idée de chacune d'elles.

2.1 QU'EST-CE QUE LA STATISTIQUE?

Le mot statistique peut s'interpréter de deux façons:

1. **Les statistiques sont des faits classifiés (plus spécialement des faits numériques) concernant une catégorie particulière d'objets.** Un exemple: les statistiques sur les accidents automobiles (56 400 morts dans des accidents d'automobiles aux États-Unis en 1969). Au hockey, les Canadiens de Montréal ont conservé une moyenne offensive de 3,82 buts en 1975. Les statistiques montrent qu'à l'Université de Montréal, le taux d'échec s'élève à 38% en première année. C'est une interprétation de la statistique que les médias d'information utilisent le plus fréquemment.

L'expression «**statistique descriptive**» est utilisée dans la partie du cours qui se rapporte à la présentation des données numériques sous forme de tableaux ou graphiques. Sont aussi inclues dans cette partie les valeurs qui résumeront des données, en donnant le centre et l'éparpillement des données par rapport au centre, pour ne citer qu'un exemple. «Ma moyenne en économie est de 86% et mes résultats varient entre 74 et 96%.»

2. **La statistique est une partie de la science qui s'intéresse à généraliser à de grands ensembles (populations) les conclusions tirées des résultats obtenus avec des ensembles beaucoup plus restreints (échantillons).** C'est ce que l'on appelle l'inférence statistique. C'est le principe même du sondage d'opinion. Un exemple: en période électorale, on interroge 1 000 personnes sur leur intention de vote, puis à partir de ces résultats, on prévoit avec une certaine précision les résultats de l'élection. On tente de découvrir le comportement d'un ensemble de personnes en connaissant la réaction d'un petit groupe de cet ensemble. C'est l'inférence statistique.

En général, est-il possible de généraliser, en ne connaissant qu'une portion de l'information nécessaire pour répondre à une question? La réponse est oui, sinon peu de gens se pencheraient sur l'étude des statistiques, mais ce oui est soumis à plusieurs restrictions. En statistique descriptive (déf. 1), certaines erreurs peuvent survenir au moment de la compilation des résultats ou en effectuant certains calculs, mais sinon les résultats sont toujours exacts. En inférence statistique, la situation est bien différente. Comme Wallis et Roberts l'ont écrit dans *The Nature of Statistics*† «La statistique est un ensemble de méthodes permettant de prendre une bonne décision face à l'incertitude». Pour prendre une bonne décision, il faut connaître quel genre et quelle quantité d'information est nécessaire, quel genre de conclusion peut être donnée et avec quelle précision.

2.2 POPULATIONS ET ÉCHANTILLONS

Une des parties de la statistique consiste à apprendre comment on prend une bonne décision concernant un grand groupe (**population**) après avoir étudié les informations recueillies pour un petit groupe de cette population (**échantillon**).

Une population comprend **tous** les êtres ou objets d'un groupe bien défini sur lequel porte notre étude. Le terme population, tel que défini en statistique, ne se rapporte pas nécessairement à tous les individus d'une communauté particulière. Ce peut être le pourcentage de monoxide de carbone dans l'atmosphère pris à midi tous les jours de l'année dans toutes les stations météorologiques du Canada ou bien tous les citoyens du comté de Verchère inscrits sur la liste électorale lors des dernières élections provinciales.

Nous serons souvent intéressés à connaître les mesures ou résultats des différents membres d'une population. Par exemple, le nombre de soldats morts au Vietnam chaque année depuis 1964 ou encore, combien Mme Dubé a dépensé chaque semaine en 1976 pour son épicerie. Il peut arriver que des membres différents de la population obtiennent les mêmes résultats; un exemple: si la population est formée de cinq étudiants A; B; C; D; E; et que ceux-ci passent un test de Q.I., il peut arriver que les résultats obtenus soient 105; 120; 120; 115; 148. Ici deux résultats sont identiques bien que les cinq étudiants soient différents.

Un échantillon est un groupe relativement petit et choisi scientifiquement de manière à représenter le plus fidèlement possible une population. Il ne faut pas oublier que c'est toute la population qui préoccupe le statisticien. Nous verrons plus loin (chapitre 8) comment choisir scientifiquement un échantillon.

Si nous connaissons les Q.I. de tous les étudiants de la polyvalente Beloeil de la régionale Yamaska, avons-nous une population ou un échantillon? Il est impossible de répondre tant que l'on ne connaît pas la question qui nous intéresse. Nous aurons une population si la question est: «Quel est le Q.I. moyen de tous les étudiants de la polyvalente Beloeil?» Nous avons toutes les informations nécessaires pour répondre à cette question. Nous aurons un échantillon si la question posée est: «Quel est le Q.I. moyen de tous les étudiants de la régionale Yamaska?» Par contre, nous aurons une population si la question est: «Quel est le Q.I. moyen des étudiants de la polyvalente Beloeil qui suivent un cours de géographie?» En effet, nous connaissons le Q.I. de tous les étudiants de la polyvalente Beloeil.

---

† Wallis, W. A., et Robert, H. V., *The Nature of Statistics* (Glencoe, Illinois: The Free Press, 1965), p. 11.

Exemple 1

Quelle est la population et quel est l'échantillon dans les situations suivantes?

Un médecin vérifie un nouveau médicament sur 100 patients atteints de leucémie, choisis au hasard. Après six ans, 20 patients sont guéris. Quelle proportions de tous les leucémiques sera guérie après six ans de traitement avec ce nouveau médicament?

L'échantillon est formé des 100 patients sur lesquels le nouveau médicament est appliqué. La population est l'ensemble de tous les leucémiques identifiés.

Exemple 2

La compagnie Lampadaire Québec achète 10 000 ampoules électriques d'un fabricant qui affirme que ses ampoules ont une vie moyenne d'au moins 1 000 heures. La compagnie Lampadaire Québec vérifie 15 ampoules et, suite à ces résultats, doit décider si elle garde ou non les 10 000 ampoules. Décrire la population et l'échantillon.

Ici la population est l'ensemble des 10 000 ampoules et l'échantillon, les 15 ampoules vérifiées.

Exemple 3

Dans un échantillon de 1 000 libéraux choisis au hasard, 400 favorisent la peine de mort, 250 conservateurs d'un échantillon de 500 font de même. Trouver la différence dans les proportions de tous les libéraux et conservateurs qui favorisent la peine de mort. Décrire les populations et les échantillons.

Deux populations (tous les libéraux et tous les conservateurs) doivent être comparées après avoir étudié deux échantillons (1 000 libéraux et 500 conservateurs). Pour que la comparaison soit significative il faut que les termes libéral et conservateur soient bien définis.

2.3 TYPES DE DONNÉES

Les données peuvent être classées de la façon suivante: à caractères spécifiques, ordinales, quantitatives.†

Données à caractères spécifiques (nominales). Les individus sont simplement placés dans la bonne catégorie puis le nombre d'individus dans chaque catégorie est enregistré. Chaque individu doit entrer dans une catégorie seulement (les catégories sont dites mutuellement exclusives).

†*N.D.T.:* Certains auteurs peuvent également définir les données de la façon suivante: nominales, ordinales, cardinales.

Exemple 1

Sexe des étudiants de la polyvalente Beloeil en 1975

Sexe	Nombre
Masculin	706
Féminin	678

Dans l'exemple 1, les catégories sont classées en colonne. Quelquefois, les données sont classées de deux façons à la fois. Ce sont des tableaux à double entrée. Dans ce genre de tableau, il faut être prudent dans le dénombrement des effectifs de chaque catégorie.

Exemple 2

Couleur des yeux des étudiants de la polyvalente Beloeil par degré en 1976

Couleur	1er	2e	3e	4e
Bleu	124	86	82	98
Brun	150	170	135	136
Vert	15	4	7	2
Noisette	103	85	95	92

Dans le premier exemple, il y avait deux catégories; dans la table ci-dessus il y a 16 catégories [chaque étudiant occupe une case correspondant au couple (couleur: degré dans le tableau)]. Si un nouvel étudiant arrive à la polyvalente avec des yeux noirs, une nouvelle catégorie sera ajoutée à la colonne couleur; la nouvelle ligne sera Noir 1 0 0 0 et le tableau comportera alors 20 catégories.

Données ordinales. Les catégories possèdent un ordre donné. Par exemple, dans une course de chevaux, l'énoncé: «des paris furent pris sur les chevaux A, B et C mais pas sur D, E, F» représente deux catégories, les gagnants et les perdants.

L'énoncé: «A gagne suivi de B, C, D, E et F» dans cet ordre représente une donnée ordinale. Noter qu'il n'est pas nécessaire que les intervalles ou différences entre les catégories soient identiques. A peut gagner par une demi-longueur et E peut se classer dix longueurs devant F.

Données quantitatives (cardinales). Les données quantitatives impliquent une mesure. Dans la course précédente, le temps pris par les six chevaux est une donnée quantitative. Des unités sont utilisées (secondes ou minutes) avec le même intervalle entre les unités. Ceci rend possible certaines opérations arithmétiques. Avec des données ordinales, nous étions dans la possibilité de dire que A s'est classé premier et B second. Avec des données quantitatives, on peut dire que A a gagné avec un temps de 62,3 s et que B a pris 63,4 s, soit une différence de 1,1 s.

Quelques autres exemples de données quantitatives: les impôts payés par les résidents du Québec; temps pris par les cyclistes lors du Tour de France; les résultats des examens de mathématiques des étudiants de secondaire III dans une polyvalente, la grandeur de ces mêmes étudiants, leur poids.

En général, différentes techniques sont utilisées selon le type de données. Dans cet ouvrage, la plupart des données seront quantitatives. Les techniques développées pour les données quantitatives pourront être utilisées pour les données à caractères spécifiques et pour les données ordinales. Si nécessaire, nous développerons quelques techniques spécifiques pour les données à caractères spécifiques ou ordinales.

2.4 TABLEAU DE DONNÉES

Quatre-vingt-deux étudiants ont obtenu les résultats suivants à un test de personnalité:

22	22	20	27	30	23	29	21	26
21	23	25	29	18	22	31	30	28
16	28	33	25	23	31	23	18	24
26	25	17	22	25	28	19	24	20
23	26	21	31	25	24	33	29	20
27	21	25	28	24	23	25	30	27
23	26	22	24	17	33	26	24	19
18	33	25	28	31	29	27	28	24
26	24	22	26	24	18	21	29	22
31								

Ce sont des données brutes (non-classées). Une telle quantité de données est difficile à utiliser et la situation pourrait être pire, si nous avions 200 étudiants au lieu de 82. Un regroupement est fait en dressant la liste des différents résultats obtenus, en les écrivant verticalement en commençant par le plus petit. En reprenant les données du tableau précédent on inscrit un trait vertical vis-à-vis le résultat correspondant. Quand on a déjà quatre **traits** verticaux, le cinquième est fait horizontalement en coupant les quatre premiers. Ainsi, il est facile de compter les multiples de 5.

Résultat	Dénombrement	Fréquence
16	|	1
17	||	2
18	||||	4
19	||	2
20	|||	3
21	𝍸	5
22	𝍸 ||	7
23	𝍸 ||	7
24	𝍸 ||||	9
25	𝍸 |||	8
26	𝍸 ||	7
27	||||	4
28	𝍸 |	6
29	𝍸	5
30	|||	3
31	𝍸	5
32		0
33	||||	4
		82

Quand toutes les **données** du tableau ont été dénombrées, le nombre de données appartenant à chaque classe de résultat est enregistré dans une colonne appelée **fréquence.** Nous utiliserons habituellement *X* ou *Y* pour les classes de résultats et *f* pour la fréquence. Une distribution de fréquence est l'ensemble du nombre de données qu'il y a dans chaque classe de résultat. La **distribution de fréquence** obtenue pour le test de personnalité des 82 étudiants débutera ainsi:

classes X	f
16	1
17	2
18	4
.	.
.	.
.	.

Chaque distribution de fréquence doit comporter les classes d'observations (*X*) et la fréquence de chaque classe (*f*). Chaque résultat doit être présent dans une classe seulement. En examinant la distribution de fréquence, on peut aisément voir qu'elle résume le tableau des données brutes et qu'elle est par le fait même, plus informative que l'ensemble des données brutes. Il est facile de déterminer le plus bas résultat (16) le plus haut (33), combien d'étudiants ont eu 27, (4). Il y a encore 18 classes différentes. Il est souvent utile de **regrouper** les classes pour en diminuer le nombre.
On peut le faire de deux façons:

(a)		(b)	
classes X	f	classes X	f
16-17	3	16-18	7
18-19	6	19-21	10
20-21	8	22-24	23
22-23	14	25-27	19
24-25	17	28-30	14
26-27	11	31-33	9
28-29	11		82
30-31	8		
32-33	4		
	82		

Au fur et à mesure que le nombre de classes diminue, l'information, qu'elles fournissent, devient plus facile à assimiler. Ainsi en (b) on a six classes seulement. On y gagne en simplicité, mais d'un autre côté, on perd les détails dans le regroupement. En (b) on voit qu'il y a 7 résultats entre 16 et 18 mais la distribution de ces 7 résultats nous est inconnue; la distribution originale nous montre qu'il y a 1 résultat à 16, 2 à 17 et 4 à 18. Quand on regroupe, il faut toujours faire un compromis entre la clarté et la perte d'informations.

La **largeur de classe** est la différence entre le plus petit résultat qui entre dans une classe et le plus petit résultat qui entre dans la classe suivante (ce n'est pas la différence entre le plus grand et le plus petit résultat qui entrent dans une même classe). Dans (a) ci-dessus chaque classe a une largeur de 2 et dans (b) chaque classe a une largeur de 3.

Le **point milieu de la classe** est la moyenne du plus petit et du plus grand résultat qui entrent dans une classe. Le milieu de la classe 16-17 en (a) est $\frac{16+17}{2} = 16{,}5$; tandis que le milieu de la classe 16-18 en (b) est $\frac{16+18}{2} = 17$.

La **limite inférieure de la classe** est le plus petit résultat qui entre dans cette classe; la **limite supérieure** est le plus grand résultat qui entre dans cette classe, ainsi dans la première classe de (b): 16 est la limite inférieure et 18 la limite supérieure.

Les **bornes d'une classe** sont à mi-chemin entre la limite supérieure d'une classe et la limite inférieure de la classe suivante.

Exemple 1

classes X	Bornes des classes	classes Y	Bornes des classes
16–18	15,5; 18,5	0,1–0,3	0,05; 0,35
19–21	18,5; 21,5	0,4–0,6	0,35; 0,65
22–24	21,5; 24,5	0,7–0,9	0,65; 0,95

Noter que le dernier chiffre d'une borne est toujours 5 et la position de la virgule indique l'unité de mesure des résultats. Ainsi, 0,5 indique que les résultats sont enregistrés à l'unité près comme dans la distribution X ci-dessus; 0,05 indique une mesure au dixième d'unité près comme dans la distribution Y ci-dessus. Les bornes des classes se rejoignent, mais ceci ne veut pas dire qu'un résultat peut être affecté à deux classes.

Exemple 2

Salaire horaire des ouvriers du textile au Québec.

Salaire (en dollars)	Bornes des classes
0–2,99	0,00 ; 2,995
3–5,99	2,995; 5,995
6 et plus	5,995; ∞

Noter que 0,005 dans les bornes des classes indique que le salaire est évalué à un cent près. La borne supérieure de la dernière classe est l'infini (symbole ∞). Une telle classe est appelée une **classe ouverte**. Elle n'a ni limite supérieure ni milieu.

Il est à remarquer également que la borne inférieure de la première classe est 0,00 au lieu de −0,05; un salaire peut difficilement être négatif.

2.5 COMMENT REGROUPER LES RÉSULTATS

Voici quelques conseils pour le regroupement.

1. Chaque résultat doit être présent dans une classe seulement.
2. Utiliser entre 10 et 20 classes.
   Ce conseil est souvent ignoré, mais seulement pour de bonnes raisons. On peut

n'employer que 5 classes, si on veut gagner en simplicité en sacrifiant quelques informations. On peut également dépasser 20 classes si on désire présenter une masse d'informations à une audience de spécialistes.

3. Les classes doivent être de même largeur. Comme nous le verrons dans les chapitres 3 et 4, les calculs en seront grandement simplifiés; c'est aussi une méthode conventionnelle de regroupement.

Quelquefois,il est impossible d'avoir des classes de même largeur. Un psychologue veut enregistrer le temps pris par 40 rats pour parcourir un labyrinthe; il retire un rat qui ne l'a pas parcouru en 3 minutes ou plus. Voici les résultats obtenus.

Temps (min.)	f
0 et moins de 1	10
1 et moins de 2	18
2 et moins de 3	4
3 et plus	8

La largeur de la dernière classe (classe ouverte) ne peut être déterminée.

4. La largeur des classes doit être autant que possible un nombre impair. On peut également privilégier certaines valeurs.

Considérons les distributions suivantes:

(a)	(b)
100 – 124	100 – 122
125 – 149	123 – 145
150 – 174	146 – 168

En examinant les limites inférieures de la distribution (a), on voit qu'elle est plus pratique que la distribution (b). Les largeurs de classes sont respectivement 25 et 23, des nombres impairs. Dans certains cas, il est plus esthétique de choisir des largeurs de classe multiples de 5 telles que 5, 10, 15, ..., même si parmi ces nombres, on retrouve des largeurs de classe paires. On peut aussi prévilégier comme première limite inférieure un nombre multiple de la largeur de classe choisie.

Un ensemble de résultats peut être traité de la façon suivante:

1. Dénombrer les résultats (voir section 2.4).
2. Déterminer l'étendue (différence entre le plus grand et le plus petit résultat).
3. Choisir une largeur de classe et le nombre de classes en respectant les quatre conseils précédents et de sorte que (largeur de classe) · (nombre de classes) donne l'étendue. Pour y arriver on peut choisir le nombre de classes puis, diviser l'étendue par celui-ci pour trouver une approximation de la largeur de la classe.
4. Déterminer les limites de chaque classe, s'assurer que chaque résultat trouve une place dans une classe.
5. Construire un tableau montrant les classes avec leurs limites ainsi que la fréquence dans chaque classe. Ce tableau doit avoir un titre qui contient la provenance des données, la date, etc.

L'étendue est 45 − 10 = 35, si on prend 4 comme largeur on obtient 10 * 4 = 40 ce qui couvre l'étendue, mais 4 est un nombre pair et le conseil 4 n'est pas respecté. Alors on prendra 3 comme largeur, mais pour couvrir l'étendue il nous faudra 12 classes car 12 * 3 = 36 les classes seront alors 10 − 12; 13 − 15; . . . ; 43 − 45.

Exemple 2

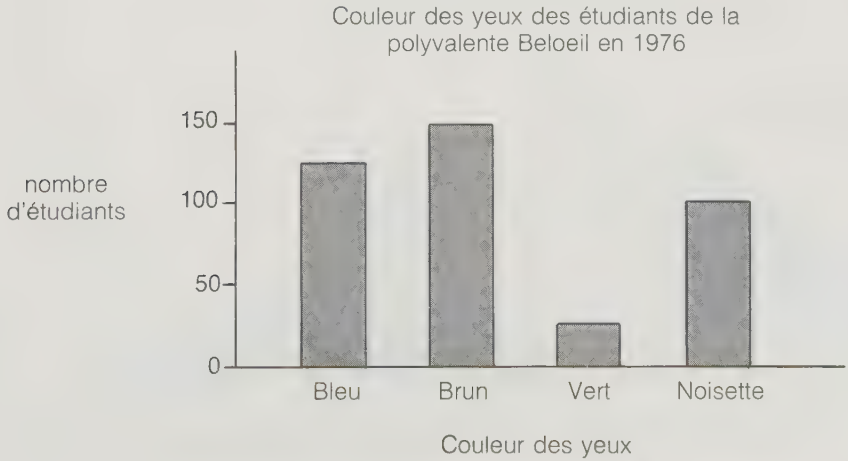
Environ 15 classes doivent servir à regrouper 300 résultats qui varient de 207 à 592. Comment choisir les classes?

L'étendue est 592 − 207 = 385. Largeur de classe: $\frac{\text{Étendue}}{\text{nombre de classes}} = \frac{385}{15} = 25$ qui est un nombre impair. Si on prend 15 classes: 15 * 25 = 375 (insuffisant). On prendra alors 16 classes: 16 * 25 = 400 (plus grand que 385). On peut former les classes suivantes: 207 − 231; 232 − 256; . . . ; 582 − 606. Un meilleur choix serait: 200 − 224; 225 − 249; . . . ; 575 − 599.

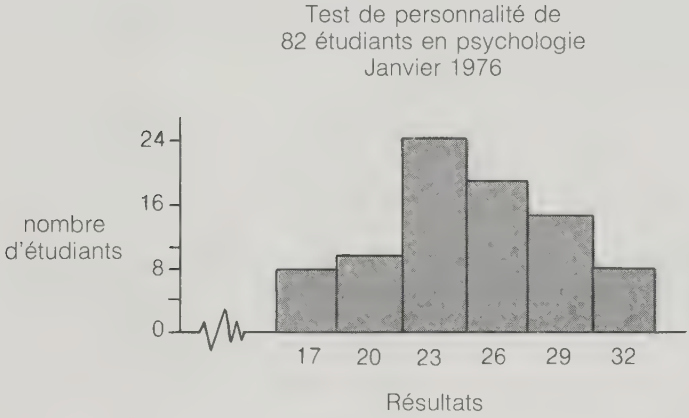
2.6 REPRÉSENTATION GRAPHIQUE DES DONNÉES

Pour plusieurs personnes la représentation graphique des données est plus facile à comprendre qu'une distribution de classes. Chaque graphique, tout comme un tableau, doit avoir un titre, une date et la provenance des données si nécessaire.

Diagramme en bâtonnets: Ce genre de graphique est utilisé pour des données à caractères spécifiques. Les caractères sont habituellement placés sur l'axe horizontal et la fréquence, la proportion ou le pourcentage est placé sur l'axe vertical. Les bâtonnets sont séparés les uns des autres pour bien faire ressortir les caractères. Les bâtonnets doivent avoir la même largeur (ainsi l'aire de chaque bâtonnet est proportionnelle à la fréquence de chaque caractère). Les axes doivent être identifiés.

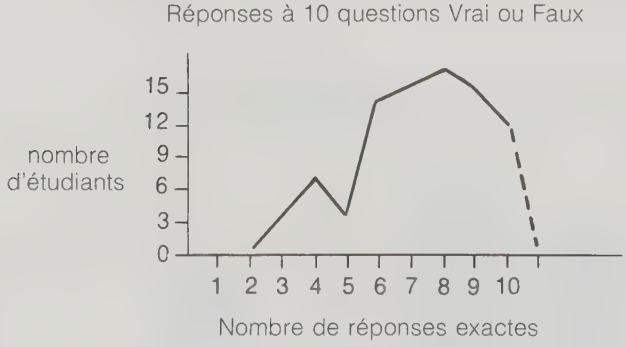


Histogramme: Un **histogramme** est utilisé quand on a des données quantitatives. Il ressemble au diagramme en bâtonnets, sauf que les bâtonnets ne sont pas séparés. Les classes sont placées sur l'axe horizontal dans un ordre croissant de gauche à droite. Les rectangles sont adjacents pour montrer que les données quantitatives ont une échelle continue. Le point milieu ou les bornes de chaque classe sont indiqués.

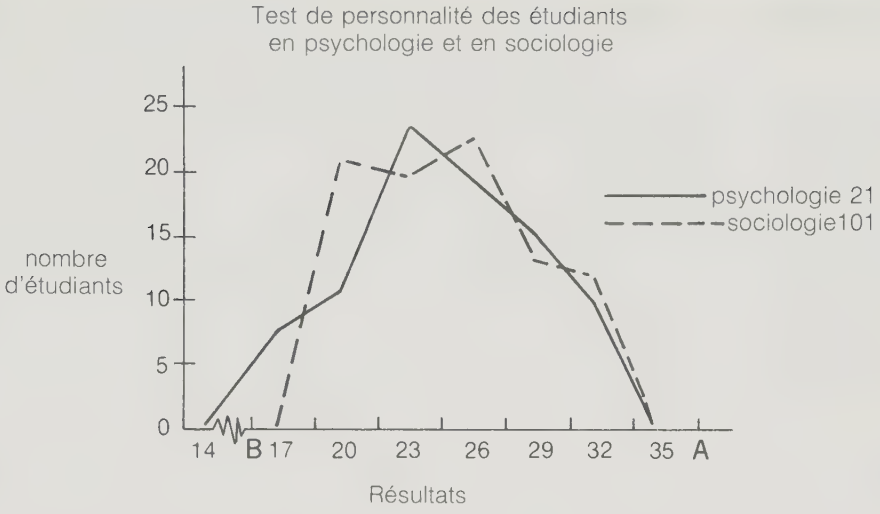


On ne peut utiliser un histogramme si une des classes est ouverte.

Polygone de fréquences: Un **polygone de fréquence** est une ligne obtenue en joignant les points milieux du sommet de chaque rectangle d'un histogramme. S'il n'y a pas d'inconvénients, une classe de fréquence zéro peut être ajoutée à chaque extrémité du graphique pour que le polygone parte de l'axe horizontal et se termine sur cet axe.

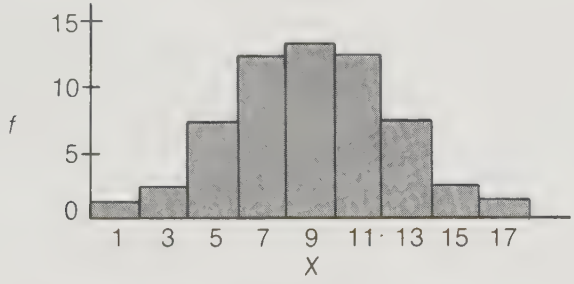


N.B. La ligne pointillée du graphique ne fait pas partie de celui-ci, car ceci laisserait sous-entendre qu'un étudiant peut avoir 11 bons résultats sur 10 questions.

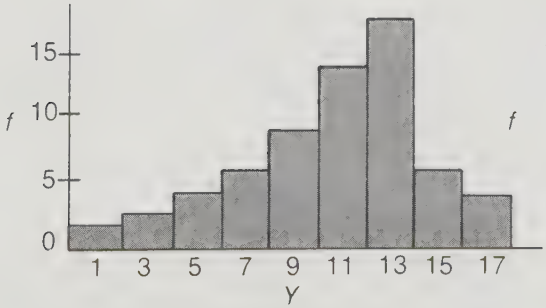


Un histogramme devient difficile à lire s'il comprend deux distributions différentes; un polygone de fréquence est ordinairement préférable.

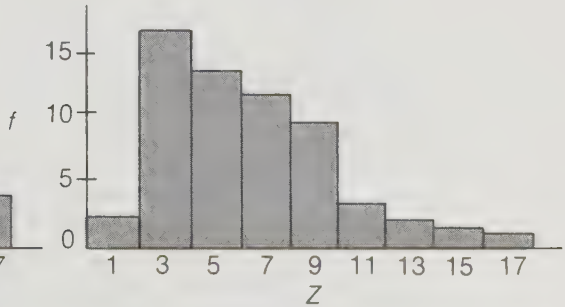
Symétrie et dissymétrie dans les graphiques: Un histogramme (ou un polygone de fréquence) est **symétrique** si en le pliant selon la verticale passant par le milieu de la distribution, les deux parties se superposent parfaitement. Si le graphique possède une «queue» plus longue à droite on dit qu'il est **dissymétrique à droite** ou encore que la distribution est étalée à droite. Si c'est l'inverse qui se présente, on dira qu'il est **dissymétrique à gauche** ou encore que la distribution est étalée à gauche.



Symétrique



dissymétrique à gauche



dissymétrique à droite

Variables discrètes et continues: Dans les différentes distributions déjà vues, la variable (le résultat du test de personnalité par exemple) utilisait un nombre fini de valeurs (16; 17; 18; . . . ; 33). Une telle variable est appelée **discrète**. Une variable est appelée **continue** si elle peut prendre toutes les valeurs d'une échelle continue. Par exemple, si X est n'importe quel nombre entre 1 et 4 comme 1,42 ou $\sqrt{3}$ ou 3,9117, alors X est une variable continue. La grandeur d'un enfant est une variable continue car elle passe par toutes les valeurs comprises entre 0,5 m et 1,5 m; ordinairement quand on mesure la grandeur d'un enfant on enregistre celle-ci au centimètre près, alors la grandeur devient une variable discrète. Il en est de même pour le poids d'une personne qu'on mesure au dixième de kilogramme près, le poids devient une variable discrète. Nous verrons plus loin que les distributions continues ont une importance théorique appréciable en statistique. Le graphique illustrant une distribution continue est une courbe au lieu d'un polygone de fréquences.

2.7 VOCABULAIRE ET SYMBOLES

statistique descriptive
inférence statistique
population
échantillon
données à caractères spécifiques
données ordinales
données quantitatives
tableau de données
dénombrement
résultat
diagramme en bâtonnets
histogramme
polygone de fréquence
distribution symétrique
fréquence
distribution de fréquences
largeur d'une classe
milieu d'une classe
limite inférieure d'une classe
limite supérieure d'une classe
bornes d'une classe
classe ouverte
étendue
$\approx$
dissymétrie à gauche
dissymétrie à droite
variable discrète
variable continue

2.8 EXERCICE

1. Le temps pris par chaque étudiant de la polyvalente pour se servir à la cafétéria est mesuré tous les midis entre le 1er et le 15 février. Pour les questions suivantes, préciser si l'information représente une population ou un échantillon (en d'autres mots avez-vous assez d'informations pour répondre à la question, si oui vous avez une population).

Quel est le temps moyen pris?

(a) Pour le dîner du 3 février?
(b) Pour les dîners du dimanche en février?
(c) Pour les dîners du second semestre?
(d) Pour tous les repas pris le 3 février?
(e) Pour les dîners les 1er et 7 février?

2. Pour les questions précédentes où l'on a un échantillon, peut-on suggérer une meilleure méthode pour choisir celui-ci?

3. Pour les cas suivants, décrire la population. Si le temps et le coût n'entrent pas en ligne de compte, peut-on recueillir les informations de toute la population?

(a) On veut vérifier si une pièce de monnaie est équilibrée. On la lance un grand nombre de fois et on enregistre le nombre de piles et le nombre de faces obtenues.

(b) Un médecin invente un nouveau traitement pour guérir le cancer de l'estomac. Il vérifie son traitement sur toutes les personnes atteintes de cette maladie en 1975.

(c) Un économiste veut compiler le nombre de sans-emploi ayant 17 ans ou plus au Canada et qui ont cherché un emploi en septembre 76.

4. Supposant que l'on connaît les revenus provenant des intérêts et dividendes de chaque citoyen de Trois-Rivières, donner des exemples de situations pour lesquels l'information représente (a) une population (b) un échantillon.

5. Décrire les populations et les échantillons.

(a) Un économiste du ministère du travail connaît le pourcentage des travailleurs sans-emploi à Montréal pour le mois dernier. Ce mois-ci, il vérifie 5 000 travailleurs pour savoir si le pourcentage de sans-emploi a changé.

(b) Un politicologue veut vérifier le pourcentage des gens qui appuient le parti Québécois à Montréal. Pour son étude, il choisit 1 000 votants dont le revenu annuel est supérieur à $16 000 et trouve que le pourcentage est de 25,3% et que celui-ci est de 22% pour 1 200 votants quand le revenu annuel est inférieur à $8 000.

(c) Quatre enveloppes de graines de tomate identiques sont semées avec des engrais différents et une cinquième est semée sans engrais et servira de témoin. On choisit 35 plants de chaque groupe puis on calcule la grandeur moyenne pour déterminer l'engrais le plus efficace.

(d) On sait qu'un chien esquimau d'un an gagne en moyenne 1 kg par mois. Une diète spéciale riche en protéines et vitamines est administrée à 35 chiens d'un an. Leur augmentation de poids est enregistrée pour savoir si la diète influence leur poids.

(e) Le régistraire de l'U.Q.A.M. sait que la moyenne des étudiants mariés est de 2,5% supérieure à ceux non-mariés et ce depuis les 4 dernières années. Ce semestre, il enregistre les résultats de 40 étudiants mariés et 50 étudiants non-mariés pour vérifier si la même différence persiste.

6. Déterminer le type de données (à caractères spécifiques, ordinales, quantitatives) .

(a) Dans une exposition d'art, un juge enregistre ses préférences.

(b) Un chimiste prépare 10 solutions en ajoutant 100 ml d'acide sulfurique à 10; 20; . . . ; 100 ml d'eau puis mesure le point d'ébullition de chaque solution.

(c) Des robes sont classées selon leur style et leur grandeur.

(d) 400 pompiers sont interrogés à propos de leurs antécédents ethnologiques.

(e) 10 tablettes de chocolat sont classées par un dégustateur.

(f) 30 étudiants passent un test de Q.I.

(g) La consommation d'alcool per capita en 1976 est déterminée pour chaque province du Canada.

7. (a) Dénombrer les résultats suivants:

126	132	121	149	130	139	127	136	138	129
121	134	139	135	128	123	133	136	124	130
127	136	132	126	145	139	131	133	142	131
134	130	141	144	136	124	136	136	133	128
123	125	139	145	148	141	126	145	138	139
133	147	136	134	132	142	149	122	131	139
130	139	136	148	132	147	121	124	148	133
139	127	147	124	148	135	142	142	133	142
121	146	145	148	127	136	130	144	143	124
148	140	136	136						

(b) Quel est le plus petit résultat? Le plus grand? Combien de fois a-t-on le résultat 139?

8. En utilisant le tableau ci-dessous, faire une classification à double entrée comportant le sexe et l'allégeance politique.

32 individus interrogés

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
L	x					x		x									x					x		x							x	
C									x		x				x				x	x	x					x		x		x		
I		x	x	x	x		x			x		x	x	x		x		x					x		x		x					x
H	x	x			x	x	x		x	x				x	x		x	x		x					x		x	x	x			
F			x	x				x			x	x	x			x			x		x	x	x	x		x				x	x	x

L = Libéral C = Conservateur
H = Homme F = Femme I = Autres N.P.D, C.S

9. Commenter les regroupements ci-dessous qui proviennent des résultats du numéro 7. Quels conseils ne sont pas respectés?

(a)

classes X	f
120–123	7
124–127	13
128–130	8
131–135	18
136–139	21
140–143	9
144–147	10
148–151	8

(b)

classes X	f
121–126	16
126–130	15
131–135	18
136–140	22
141–145	14

(c)

classes X	f
120–122	6
123–125	8
126–128	9
129–131	9
132–134	13
135–137	13
138–140	11
141–143	8
144–146	7
147–149	11

10. (a) Comment les résultats de l'exercice numéro 7 peuvent-ils être regroupés dans les classes de largeur 5?

(b) Quelles sont les bornes et le milieu de la classe contenant le plus petit résultat?

11. Suivre les conseils pour déterminer la largeur de classe, le nombre de classes de chacune des distributions suivantes. Quelles seront les limites de la classe contenant le plus petit résultat?

	Nombre de résultats	Plus grand résultat	Plus petit résultat	Nombre approximatif de classes
(a)	10 000	208	22	20
(b)	600	737	112	14
(c)	120	114	52	10

12. (a) Les diamètres d'un échantillon de 120 écrous varient de 12,5 mm à 12,8 mm. Comment le groupement des données peut-il être fait?

(b) Quelles sont les bornes de deux classes contenant les plus petites valeurs?

(c) Quels sont les milieux de ces deux classes?

13. Le registraire d'un collège enregistre les résultats avec la pondération suivante: A = 4,00 B = 3,00 C = 2,00 D = 1,00 F = 0,00. Les résultats des 490 étudiants varient de 0,42 à 4,00. Comment regrouper ces résultats en 15 classes environ? (*N.B.* il est impossible qu'un étudiant ait un résultat supérieur à 4,00.)

14. Le meilleur résultat obtenu lors d'un examen d'une heure en statistique est 97 et le plus bas est 62; 100 étudiants ont passé cet examen. Comment regrouper les résultats en classes de largeur 5? Combien de classes seront nécessaires?

15. Trouver les bornes, la largeur et le point milieu des deux premières classes des distributions suivantes. Soit X le point milieu de chaque classe; trouver ΣfX si possible (ajouter des colonnes pour X et fX aux distributions suivantes).

(a)

Nombre d'essais pour apprendre à un rat à reconnaître des objets

Nombre d'essais	Nombre de rats
1–3	7
4–6	5
7–9	3
10–12	3

(b)

Prix de vente des automobiles américaines en 1969

Prix	Pourcentage
$1 500–2 499	0,28
$2 500–3 499	0,49
$3 500–4 499	0,17
$4 500 et plus	0,06

(c)

Grandeur de 288 Canadiens

Grandeur en mètres	Nombre
1,50 – 1,62	10
1,63 – 1,75	150
1,76 – 1,88	120
1,89 – 2,01	8

16. Tracer un histogramme et un polygone de fréquences (si possible) pour chacune des distributions du problème précédent.

17. Illustrer chacune des distributions suivantes avec le meilleur diagramme possible en choisissant parmi un diagramme en bâtonnets, un histogramme ou un polygone de fréquences.

(a)

Habits loués dans des agences de la rue Ste-Catherine à Montréal en 1976

Types d'habits	Nombre
Flanelle grise	1800
Tweed gris	1200
Serge grise	400
Autres	200

(b)

Salaire horaire des travailleurs de la construction du Québec

Salaire	Pourcentage	
	1960	1970
\$1 moins de \$3	20	5
\$3 moins de \$5	70	60
\$5 moins de \$7	10	35

RÉPONSES

1. Échantillon: b, c, d. Population a, e

3. (a) La population est l'ensemble de **tous** les résultats des jets de la pièce et ne peut être déterminée. (b) La population est l'ensemble des malades ayant un cancer de l'estomac. On peut obtenir toutes les informations nécessaires si on peut faire un dépistage adéquat de cette maladie. (c) Population: tous les sans-emplois ayant 17 ans ou plus au Canada qui ont cherché un emploi en septembre 1976. Toutes les informations peuvent être obtenues du bureau de la main-d'oeuvre du Canada.

5. (a) Population: l'ensemble des travailleurs de Montréal ce mois-ci. Échantillon: 5 000 travailleurs choisis ce mois-ci. (b) Deux populations: ensemble des votants montréalais gagnant plus de \$16 000 et la deuxième population, ceux gagnant moins de \$8 000. Il est bon de souligner que cette façon de choisir les populations est douteuse, car une famille de 5 enfants dont le père gagne \$16 500 peut être plus défavorisée qu'un célibataire gagnant \$7 800. Les échantillons, 1 000 votants du premier groupe et 1 200 du second. (c) 5 populations: chaque paquet de graines. Les échantillons: 35 plants de chaque population. (d) Population: l'ensemble des chiens esquimaux d'un an. Échantil-

lon: 35 chiens d'un an soumis à une diète spéciale. (e) Deux populations: les étudiants mariés et les étudiants célibataires à l'U.Q.A.M. Échantillons: 40 étudiants mariés, 50 étudiants célibataires.

7. (a) Colonne X: 121 à 149; colonne f: 4; 1; 2; 5; 1; 3; 4; 2; 1; 5; 3; 4; 6; 3; 2; 11; 0; 2; 8; 1; 2; 5; 1; 2; 4; 1; 3; 6; 2. (b) 121; 149; 8.

9. (a) Pas assez de classes, pas la même largeur (128 – 130) et (131 – 135), le conseil 4 n'est pas respecté (la plupart des classes ont une largeur de 4 qui est pair). (b) 126 apparaît dans 2 classes, les résultats de 146 à 149 n'apparaissent pas, pas assez de classes, le conseil 4 n'est pas respecté 120 – 124; 125 – 129; . . . . . serait préférable. (c) Parfait, sauf que la fréquence de la classe 120 – 122 est 5 et non 6.

11. (a) 10; 19; 20 – 29. (b) 50; 13; 100 – 149. (c) 5; 13; 50 – 54 (préférable à 51 – 55).

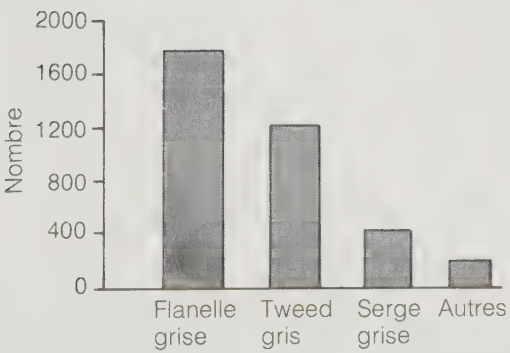
13. Quinze classes de largeur 0,25; 0,26 – 0,50; 0,51 – 0,75; . . . ; 3,76 – 4,00. On peut également écrire (0,25 – 0,49; . . . ; 3,5 – 3,74; 3,75 – 4,00. Cette dernière classe étant plus large que les précédentes).

15. (a) (0,5 – 3,5); 3; 2; (3,5 – 6,5); 3; 5. $\Sigma fX = 14 + 25 + 24 + 33 = 96$ (b) ($1499,50 – $2499,50); 1000; 1999,50 ($2499,50 – $3499,50); 1000; 2999,50. ΣfX ne peut être calculé car le point milieu de la dernière classe n'est pas connu. (c) (1,495 – 1,625) 0,13; 1,56. (1,625 – 1,755); 0,13; 1,69. $\Sigma fX = 10 * 1,56 + 150 * 1,69 + 120 * 1,82 + 8 * 1,95 = 503,1$.

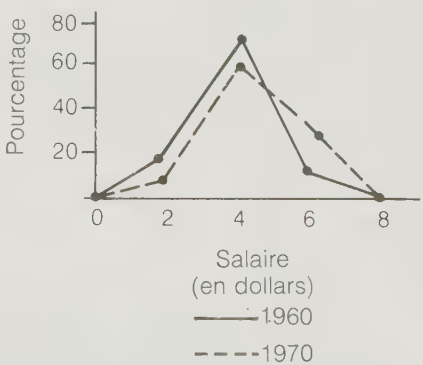
17. (a) Données à caractères spécifiques, on utilise un diagramme en bâtonnets. (b) Données quantitatives, on peut utiliser un histogramme, mais on a deux groupes de données, un polygone de fréquences est préférable.

(a) Habits loués sur la rue Ste-Catherine

30 novembre



(b) Salaire horaire des travailleurs de la construction



Chapitre 3

Mesures de tendance centrale.

Dans ce chapitre, nous nous efforcerons de trouver un nombre qui représente au mieux l'ensemble d'une collection. Le mode, la médiane et la moyenne sont les trois mesures de tendance centrale que nous étudierons. La moyenne est la mesure de tendance centrale la plus connue et la plus utilisée. Les quartiles, les déciles et les rangs centiles seront également étudiés. Enfin, nous étudierons la mesure des mesures de tendance centrale à savoir...

3.1 INTRODUCTION

Nous chercherons un nombre qui représente le mieux un ensemble de données. Il y a trois façons d'y arriver : le **mode**, la **médiane** et la **moyenne**. Ces trois termes sont appelées des «mesures de tendance centrale» parce que chacune d'elles donne une idée du résultat qui représente le centre de l'ensemble des données.

Si on jette un coup d'œil sur la manière dont on a bien coché et que l'on dit « a **plus** de 100 que tout autre chose» on utilise à ce moment le mode. L'expression «mes résultats en linguistique sont bien au-dessus de 50%» utilise la médiane. «La **moyenne** de mes résultats d'examens est 80%» Utilise la moyenne. Dans la vie courante, la moyenne est la plus largement utilisée.

3.2 LE MODE

Le mode est le résultat qui apparaît le plus souvent dans une distribution. L'abréviation en est *Mo* (certains auteurs utilisent également la notation $\hat{x}$). Le mode est la seule mesure centrale qui peut être utilisée pour des données à caractères spécifiques.

Exemple 1

X : 1 1 1 2 2 2 2 2 3 3 3 3 3 3 4 4 5 6 8

Le résultat 3 apparaît 6 fois. Tous les autres résultats apparaissent moins souvent. Alors $Mo = 3$ ($\hat{x} = 3$).

Exemple 2

Y: 1; 1; 1; 2; 2; 2; 2; 2; 3; 3; 4; 4; 4; 4; 4; 5; 6; 6.

Les résultats 2 et 4 apparaissent chacun 5 fois donc $Mo = 2$ et 4. Cette distribution est appelée **bimodale.**

Exemple 3

Z: 1; 7; 8; 10; 12; 14; 17; 23; 24.

Chaque résultat apparaît le même nombre de fois (une fois dans ce cas); le mode est une mesure non-significative. C'est souvent le cas quand on a une distribution qui contient peu de résultats. Dans le cas des données groupées, la classe modale est celle qui possède la plus grande fréquence; le point milieu de cette classe est appelé le mode brut.

Exemple 4

classes *X*	*f*
1–4	2
5–8	7
9–12	5
13–16	3

La classe modale est 5-8; le mode brut est 6,5 (milieu de la classe modale).

Le mode est la mesure la plus simple des trois mesures de tendance centrale. Tout ce dont on a besoin pour trouver le mode, c'est la table de fréquences d'une distribution, que celle-ci soit classée ou non. Le mode n'est pas affecté par les résultats placés aux extrémités; ainsi les données suivantes: 1; 3; 5; 5; 5; 6; et 1; 3; 5; 5; 5; 213 ont le même mode. Ceci peut dans certains cas présenter un inconvénient: un exemple: un rat non entraîné prendra un temps énorme a parcourir un labyrinthe par rapport aux autres rats et ceci n'affectera pas le mode. Le mode est rarement utilisé, sauf peut-être quand on a des données à caractères spécifiques.

3.3 LA MÉDIANE

La médiane est le résultat du milieu **après** que les résultats de la distribution furent ordonnés dans un ordre croissant ou décroissant. L'abbréviation utilisée est *Md* (certains auteurs utilisent $\tilde{\tilde{X}}$). La médiane peut être utilisée pour des données ordinales ou quantitatives mais pas pour des données à caractères spécifiques.

Exemple 1

X: 8; 1; 10; 7; 17; 23; 12; 15; 24.

En les ordonnant on obtient: 1; 7; 8; 10; 12; 15; 17; 23; 24. Le résultat du milieu est 12 donc: $Md = (\tilde{X} = 12)$

Exemple 2

Y: 1; 7; 8; 10; 12; 15; 17; 23; 24; 29.

Ici on a 10 résultats; le résultat du milieu se trouve à mi-chemin entre 12 et 15 soit $Md = 13{,}5$. Il est à noter que la médiane d'une distribution n'est pas nécessairement un résultat de celle-ci.

Quand on a seulement quelques résultats, il est mieux de savoir ce que veut dire la médiane et d'utiliser le bon sens pour la trouver, comme on vient de le faire dans les deux exemples précédents. Quand on possède un grand nombre de résultats, il est bon d'avoir une loi pour nous aider. Soit n le nombre de résultats ordonnés. Si $n/2$ est un entier, la médiane est à mi-chemin entre le $(n/2)^{e}$ résultat et le suivant $\left[(n/2) + 1\right]^{e}$. Si $(n/2)$ n'est pas un entier, arrondir vers le haut pour trouver le rang du résultat médian.

Exemple 3

n résultats ont été classés en ordre croissant. Quelle est la médiane si (*a*) $n = 200$ (*b*) $n = 201$?

a) $200/2 = 100$. La médiane est à mi-chemin entre le 100^{e} et le 101^{e} terme.

b) $201/2 = 100{,}5$. La médiane est le 101^{e} terme.

La médiane pour des données groupées est trouvée par interpolation. Cette méthode ne sera pas étudiée car la médiane est très peu utilisée pour des données groupées. La médiane est aussi facile à comprendre que le mode mais implique plus de travail. On doit écrire tous les résultats dans un ordre croissant et les compter, puis en appliquant la loi précédente, on trouve la médiane. Tout comme le mode, la médiane n'est pas affectée par les résultats placés aux extrémités. Ainsi les distributions 1; 3; 4; 6; 9; 12 et 1; 3; 4; 6; 9; 213 ont la même médiane (5). On ne peut trouver la médiane de données à caractères spécifiques.

3.4 LA MOYENNE DES DONNÉES NON GROUPÉES

La moyenne est la somme des résultats divisée par le nombre de résultats. Elle est utilisée seulement pour des données quantitatives. Le symbole utilisé pour désigner la **moyenne d'un échantillon** est $\overline{X}$ (lire: *X* barre) si les résultats sont appelés *X*. Il en va de même si les résultats sont appelés *Y*, alors leur moyenne sera $\overline{Y}$ (*Y* barre). Soit n le nombre de résultats d'un échantillon, alors

$$\overline{X} = \frac{\text{somme des résultats}}{\text{Nombre de résultats}} = \frac{\Sigma X}{n}$$

Exemple 1

Résultats d'examens: 70; 75; 80; 80; 85; 90.

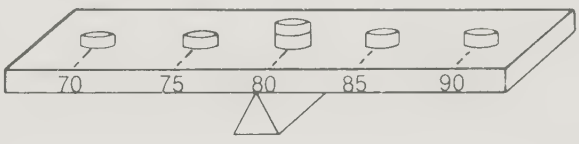
$$\overline{X} = \frac{70 + 75 + 80 + 80 + 85 + 90}{6} = \frac{480}{6} = 80$$

Exemple 2

Nombre de centimètres de neige (X) tombé chaque jour de la première semaine de janvier 1976: 3; 0; 2; 5; 1; 1; 2. Trouver la chute moyenne pour cette semaine.

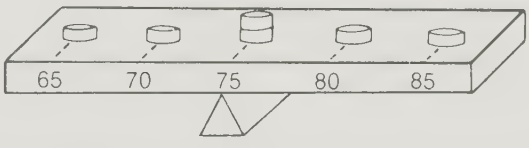
$$\bar{X} = \frac{3+0+2+5+1+1+2}{7} = \frac{14}{7} = 2{,}0$$

Pour mieux comprendre ce qu'est la moyenne, on peut faire une analogie avec le point d'équilibre d'une balance formée d'une règle uniforme. Par exemple les résultats 70; 75; 80; 80; 85; 90 seront représentés par la balance suivante:



Le point d'équilibre est 80 qui est la moyenne des résultats.

Qu'arrive-t-il si chaque résultat de l'échantillon est diminué d'une même quantité? Reprenons votre balance et essayons de voir ce qui se passe si on déplace chaque masse de 5 unités vers la gauche.



Il est évident pour que la règle soit en équilibre, qu'il faut déplacer le support de 5 unités vers la gauche. Mais, si on connaît le point d'équilibre après que les poids furent déplacés de 5 unités vers la gauche, on peut alors dire où celui-ci était originalement. Similairement, une constante A peut être soustraite de chaque résultat pour faciliter les calculs. Après que la moyenne du nouvel ensemble est trouvée, la constante A est ajoutée à celle-ci et on obtient ainsi la moyenne de la distribution originale.

Exemple 3

Trouver la moyenne des résultats ci-dessous:

	$A = 5$	$A = 20$	$A = 70$	$A = 80$
X	$X - 5$	$X - 20$	$X - 70$	$X - 80$
70	65	50	0	−10
75	70	55	5	− 5
80	75	60	10	0
80	75	60	10	0
85	80	65	15	5
90	85	70	20	10
Somme: 480	450	360	60	0
Moyenne: 80	75	60	10	0
$= \bar{X}$	$\bar{X} = 75 + 5$	$\bar{X} = 60 + 20$	$\bar{X} = 10 + 70$	$\bar{X} = 80 + 0$

Il est à noter que l'on peut choisir un nombre A quelconque. Toutefois, il est plus utile d'estimer la moyenne et de donner à A une valeur près de cet estimé.

Exemple 4

Trouver la moyenne des résultats Y ci-dessous en posant (a) $A = 40$ (b) $A = 200$ (c) $A = 204$

	(a)	(b)	(c)
Y	$Y - 40$	$Y - 200$	$Y - 204$
197	157	− 3	− 7
198	158	− 2	− 6
199	159	− 1	− 5
201	161	1	− 3
203	163	3	− 1
207	167	7	3
212	172	12	8
215	175	15	11
	1312	32	0

$$\bar{Y} = \frac{1312}{8} + 40 \qquad \bar{Y} = \frac{32}{8} + 200 \qquad \bar{Y} = \frac{0}{8} + 204$$
$$= 164 + 40 = 204 \qquad = 204 \qquad = 204$$

Il est évident que 40 est un mauvais choix bien que l'on obtienne quand même un bon résultat. $A = 200$ est certainement le meilleur choix. La question (c) a été posée dans le but de vous montrer que si la constante soustraite de chaque résultat est égale à la moyenne, alors $\Sigma (X - \bar{X}) = 0$. **Pour tout ensemble de résultats, la somme des différences en utilisant la moyenne est égale à zéro. Si la somme des différences d'une certaine constante est zéro, alors celle-ci est la moyenne** (voir problème 23 à la fin du chapitre).

Quand on a peu de résultats, la moyenne est relativement simple à calculer. La situation se complique passablement si on possède 1 000 résultats. Heureusement, pour de tels problèmes on peut utiliser un ordinateur ou une calculatrice.

La moyenne peut ne pas exister. Supposons qu'un psychologue veut enregistrer les temps pris par 5 rats pour franchir un labyrinthe, et qu'il décide de retirer un rat qui prend plus de 5 minutes. Quel est le temps moyen pris par les rats si les résultats sont les suivants: 2; 3; 3; 4; et plus de 5 minutes? Les quatre premiers résultats sont connus mais le cinquième ne l'est pas; ce temps peut être aussi bien 6 ou 17 minutes. On ne peut donc pas calculer la moyenne.

Contrairement au mode et à la médiane, la moyenne dépend de tous les résultats. Quelquefois, c'est un inconvénient. Par exemple, pour les résultats 7; 9; 9; 31, la médiane est 9 et la moyenne 14. Laquelle de ces deux valeurs est la plus représentative? Si on s'intéresse au salaire des ouvriers d'une petite usine exprimé en milliers de dollars, il est évident que la médiane (9) est plus représentative, la moyenne (14) étant trop affectée par

le résultat 31 qui représente probablement le salaire du propriétaire.

Jusqu'à présent, la moyenne ne semble pas avoir beaucoup plus d'avantages que le mode ou la médiane, bien qu'elle soit très utilisée en statistique. La première raison est que la moyenne dépend de tous les résultats, ce qui est un grand avantage. Un exemple: si vos résultats d'examens sont 50; 50; 50; 90; 100, vous préférerez la moyenne (68) à la médiane (50), pas seulement parce que celle-ci est plus haute, mais surtout parce qu'elles vous semble plus juste.

La seconde raison est que la moyenne est déterminée algébriquement. Nous verrons que la moyenne sera très utile lorsque nous ferons de l'inférence statistique entre un échantillon et la population d'où il provient.

On peut voir un certain avantage de la moyenne dans l'exemple qui suit. Considérons le rendement de deux champs de haricots exprimé en boisseaux.

Champ X: 1; 3; 3; 3; 4; 4; 5; 5; 6; 6. $Mo = 3$, $Md = 4$, $\overline{X} = 4$, $n = 10$.

Champ Y: 1; 1; 2; 2; 4; 4; 4; 6; 6; 6; 6; 7; 7; 9; 10. $Mo = 6$, $Md = 6$,

$\overline{Y} = 5$, $n = 15$.

Supposons qu'on décide de considérer la production totale des deux champs. Pour trouver les valeurs du mode et de la médiane, il nous faudra réécrire en ordre croissant les 25 données ainsi:

Champs X et Y: 1; 1; 1; 2; 2; 3; 3; 3; 4; 4; 4; 4; 4; 5; 5; 6; 6; 6; 6; 6; 6; 7; 7; 9; 10.

On obtient: $Mo = 6$, $Md = 4$.

Pour trouver la moyenne de la production totale, on n'est pas obligé de tout réécrire, on peut raisonner comme suit:

$$\overline{X} = \frac{\Sigma X}{10} = 4 \text{ et } \Sigma X = 4 * 10 = 40, \text{ de même,}$$

$$\overline{Y} = \frac{\Sigma Y}{15} = 5 \text{ et } \Sigma Y = 5 * 15 = 75$$

par suite la somme des 25 résultats sera $\Sigma X + \Sigma Y = 40 + 75 = 115$, la moyenne de la production totale est $\frac{115}{25} = 4{,}6$. Cet exemple ne vous a peut-être pas convaincu, car avec si peu de données, on peut toujours trouver le mode et la médiane. On peut facilement s'imaginer le travail nécessaire pour résoudre le même problème avec plus de 100 données.

3.5 LA MOYENNE DE DONNÉES REGROUPÉES

Pour trouver la moyenne de données regroupées, on agit comme si tous les résultats d'une classe étaient situés au point milieu de celle-ci. Pour trouver la moyenne de la distribution suivante, il suffit de trouver la moyenne d'une distribution en prenant les points

milieux de chaque classe comme résultat. Ainsi on aura 15 résultats 2; 10 résultats 5 et 5 résultats 8.

classes	f
1 – 3	15
4 – 6	10
7 – 9	5
	30

En additionnant 15 résultats 2, on obtient $15 * 2$. Il en sera de même pour les autres classes. Nous obtiendrons la moyenne

$$\frac{15 * 2 + 10 * 5 + 5 * 8}{30} = \frac{30 + 50 + 40}{30} = \frac{120}{30} = 4.$$

Il sera plus facile de former une colonne appelée fX et d'additionner celle-ci (f représente la fréquence et X le point milieu de la classe).

classes	f	X	fX
1 – 3	15	2	30
4 – 6	10	5	50
7 – 9	5	8	40
	30		120

$$\overline{X} = \frac{120}{30} = 4.$$

En trouvant la somme des colonnes f et fX on obtient $\Sigma f = n$ et la formule pour trouver la moyenne de données regroupées devient

$$\overline{X} = \frac{\Sigma fX}{\Sigma f} = \frac{\Sigma fX}{n}$$

Exemple 1

Trouver la moyenne de la distribution

classes	f	X	fX
10 – 19	14	14,5	203,0
20 – 29	14	24,5	343,0
30 – 39	22	34,5	759,0
	50		1 305,0

$$\overline{X} = \frac{1\,305{,}0}{50} = 26{,}1$$

Nous aurons à calculer fréquemment la moyenne de données regroupées. Comme nous l'avons fait pour les données non groupées, nous simplifierons les calculs. Un nombre A peut être soustrait de chaque valeur de X (point milieu de la classe), la moyenne de ces nouveaux résultats est calculée puis le nombre A est ajouté à celle-ci pour trouver la moyenne de la distribution originale.

Exemple 2

Trouver la moyenne de la distribution précédente en posant $A = 24{,}5$

X	f	$X - A$ $X - 24{,}5$	$f(X - A)$
14,5	14	−10	−140
24,5	14	0	0
34,5	22	+10	220
	50		80

$$\overline{X} = \frac{\Sigma f(X - A)}{n} + A$$

$$= \frac{80}{50} + 24{,}5 = 1{,}6 + 24{,}5 = 26{,}1$$

Si A est judicieusement choisi, les calculs sont considérablement simplifiés. Habituellement, on pose A égal au point milieu d'une des classes de la distribution.

Exemple 3

Trouver la moyenne de la distribution ci-dessous en posant (a) $A = 185$ (b) $A = 175$.

X	f
175	12
180	12
185	13
190	14
195	9
	60

(a)		(b)	
$X - 185$	$f(X - 185)$	$X - 175$	$f(X - 175)$
−10	−120	0	0
− 5	− 60	5	60
0	0	10	130
5	70	15	210
10	90	20	180
	− 20		580

(a) $\overline{X} = \frac{-20}{60} + 185 = 184{,}7.$

(b) $\overline{X} = \frac{580}{60} + 175 = 9{,}7 + 175 = 184{,}7$

Il est à noter que le choix (a) est meilleur que celui de (b). En général, on prend pour A le point milieu d'une classe qui se situe vers le centre de la distribution. Également, la moyenne est calculée à une décimale de plus que les données. La décimale 5 est arrondie au chiffre pair le plus près. Ainsi, 15,35 devient 15,4 mais 15,25 devient 15,2. Il est bon également de se faire une bonne idée de la valeur de la moyenne d'une distribution, on pourra ainsi voir si le résultat calculé a du sens et déceler des erreurs de calcul.

Exemple 4

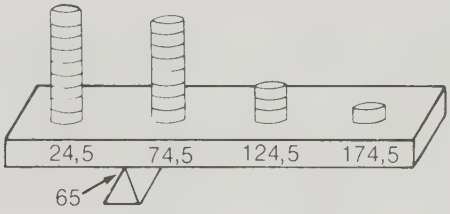
Parmi les résultats suivants, quels sont ceux qui ne peuvent être la moyenne de la distribution suivante?

(a) 204 (b) 12 (c) 23 (d) 100 (e) 70

classes	f
0–49	9
50–99	8
100–149	3
150–199	1

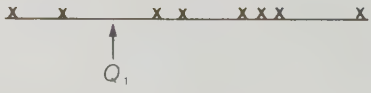
Le résultat 204 est en dehors de la distribution; 12 et 23 sont tous les deux impossibles puisqu'ils sont tous les deux dans la même classe; 100 est trop grand, car il y a seulement 4 résultats qui lui sont supérieurs et 17 lui sont inférieurs. 70 est possible, se situant approximativement au milieu de la distribution; la valeur véritable de la moyenne est 65,0.

Pour bien comprendre la moyenne, on peut utiliser l'exemple de la balance en plaçant les résultats au point milieu de chacune des classes.



3.6 QUARTILES, CENTILES ET RANGS CENTILES

Le raisonnement qui a été utilisé pour trouver la médiane, peut être réutilisé pour trouver les quartiles et les centiles. Le premier quartile (noté Q_1) est le nombre qui est plus grand que le quart des données (et plus petit que les trois quarts). Le deuxième quartile Q_2 est la médiane. Le troisième quartile, Q_3, est plus grand que les trois quarts des données. Le soixantième centile (noté P_{60}) est plus grand que 60% des données (et plus petit que 40%). Noter que $P_{25} = Q_1$; $P_{50} = Q_2$; $P_{75} = Q_3$. Les quartiles et les centiles sont appelés des mesures de position.



Si huit résultats sont écrits en ordre, Q_1 sera plus grand que deux résultats et plus petit que les six autres; le premier quartile sera donc à mi-chemin entre le 2e et le 3e résultat. Si $\frac{1}{4} * n$ n'est pas entier, on l'arrondit vers le haut pour trouver le rang de Q_1. Ainsi, si $n = 50$, Q_1 sera le 13e résultat après que ceux-ci seront écrits en ordre croissant. En effet, 50/4 = 12,5 et en arrondissant, on obtient 13, ce qui veut dire que 12 résultats sont plus petits et 37 sont plus grands.

Q_3 et P_a sont déterminés de la même façon, en multipliant n par 3/4 et par $a/100$ respectivement. Si le résultat est un entier K, on choisit le nombre qui est à mi-chemin entre le K^e et le $(K + 1)^e$ résultat; si le résultat n'est pas un entier, on l'arrondit vers le haut pour trouver le rang de Q_3 ou P_a.

Exemple 1

Quatre-vingt-dix résultats sont écrits en ordre croissant. Trouver (a) Q_1 (b) Q_3 (c) P_{60} (d) P_{16}

(a) $\frac{1}{4} * 90 = 22{,}5$; Q_1 est le 23e résultat.

(b) $\frac{3}{4} * 90 = 67{,}5$; Q_3 est le 68e résultat.

(c) $\frac{60}{100} * 90 = 54$; P_{60} est à mi-chemin entre le 54e et le 55e.

(d) $\frac{16}{100} * 90 = 14{,}4$; P_{16} est le 15e résultat.

Le rang centile d'un résultat est le pourcentage de résultats qui lui sont inférieurs. Si 75 est le 349e résultat d'une liste de 400 résultats écrits dans un ordre croissant, alors il y a 348 résultats inférieurs à 75 et $\frac{348}{400} * 100 = 87$ pour cent des résultats sont plus petits que 75; le rang centile de 75 est 87.

Exemple 2

Un étudiant de CEGEP a une moyenne de 92,5; il est le 16e plus fort d'une classe de 300. Quel est son rang centile?

Il est supérieur à $300 - 16 = 284$ étudiants de sa classe. Son rang centile sera: $\frac{284}{300} * 100 = 94{,}7$.

3.7 VOCABULAIRE ET SYMBOLES

mode	Mo, $\hat{X}$	moyenne	$\overline{X}$, $\overline{Y}$
mode brut		quartiles	Q_1, Q_3
médiane	Md, $\widetilde{X}$	centiles	P_a
		rang centile	

3.8 EXERCICES

1. Traduire en termes mathématiques:

(a) Pour trouver la moyenne $\bar{X}$ de n résultats X d'un ensemble, soustraire une constante A de chaque résultat, trouver la moyenne du nouvel ensemble puis ajouter A; appeler le résultat $\bar{X}$.

(b) Pour trouver la moyenne $\bar{X}$ d'une distribution regroupée, soustraire A de chaque point milieu des classes, multiplier cette différence par la fréquence correspondante de chaque classe, additionner, diviser par le nombre de résultats puis additionner A.

2. 12 étudiants ont obtenu les résultats suivants à deux examens.

Examen 1: 72; 75; 75; 97; 54; 72; 86; 72; 63; 79; 82; 91.
Examen 2: 78; 42; 72; 88; 86; 97; 91; 79; 82; 86; 91; 74.

(a) Trouver le mode, la médiane et la moyenne de chaque examen.

(b) Trouver le mode, la médiane et la moyenne de l'ensemble des deux examens (24 résultats).

(c) Laquelle des mesures centrales représente le mieux les résultats du 2e examen.

(d) Trouver Q_1 et Q_3 pour le 2e examen.

(e) Quel est le rang centile du résultat 88 dans le 2e examen?

3. (a) Trouver le mode, la médiane et la moyenne de chacune des distributions suivantes, si possibles; sinon, expliquer.

(I) Salaire hebdomadaire (en dollars) des employés de cafétéria: 70; 60; 70; 70; 6; 4; 90; 30; 16; 80; 12; 70.

(II) Résultats d'examens de statistiques: 50; 100; 90; 100; 50; 60; 80; 70; 40; 100; 90; 95; 75; 50; 74; 81; 76; 69; 75.

(III) Poids de lettres et colis postaux enregistrés par une balance ayant une échelle variant de 0 à 10 g: 0,7; 0,9; 0,8; 1,1; 0,5; 0,8; 0,5; 1,1; 0,7; 0,6; 0,4; 0,9; 1,3; 0,4; 0,6; 0,5; 1,2; 0,4; 0,3; 0,5; plus de 10; 3,0; 0,2; 0,4.

(IV) Nombre de voitures stationnées dans la cour de l'école à midi entre les 17 et 23 avril: 212; 650; 435; 672; 510; 621; 287.

(V) Marques des voitures stationnées dans la cour de l'école le 22 avril à midi: 131 Ford — 142 Chevrolet — 183 VW — 165 autres.

(b) Pour chaque distribution donnée en (a), quelle mesure de tendance centrale semble la plus représentative (le mode, la médiane ou la moyenne?)

4. Les salaires annuels de 20 travailleurs sont donnés ci-dessous.

$8 000	$ 9 800	$11 200	$11 700	$12 400
8 500	10 000	11 400	11 800	12 800
8 700	10 400	11 600	11 900	13 500
9 200	10 800	11 700	11 900	33 000

(a) Trouver la médiane et le salaire moyen.

(b) Laquelle entre la médiane ou la moyenne est la plus représentative? Pourquoi?

(c) Trouver Q_1, Q_3 et P_{40}.

5. Laquelle parmi les mesures de tendance centrales sera la plus représentative dans les situations suivantes?

(a) Le gérant d'un magasin de chaussures veut classer ses chaussures selon leur pointure.

(b) Un économiste veut étudier les revenus des chefs de famille habitant un complexe à loyer modique de Montréal.

(c) Un ingénieur veut vérifier la charge maximum que peut supporter chacun des câbles produits par sa compagnie.

6 Trouver la moyenne des résultats suivants en donnant a A une valeur appropriée.

(a) X: 15; 17; 18; 19; 26; 22; 24; 16; 21; 17.

(b) Y: 514; 514; 520; 536; 518; 504; 498; 537; 521.

(c) Z: 0,0021; 0,0027; 0,0019; 0,0018; 0,0023; 0,0026; 0,0029; 0,0018.

7 Sans la calculer, déterminer laquelle des valeurs suivantes de $\overline{X}$ serait la moyenne des distributions suivantes.

	Distributions	Valeurs de $\overline{X}$
(a)	15; 17; 24; 56; 63; 81	14; 16; 30; 43; 63; 89
(b)	190; 230; 112; 200; 180; 140	100; 120; 170; 200; 291
(c)	24; 12; 18; 36; 10; 12	4; 10; 12; 20; 25

8 La chute de pluie mensuelle a Miami, Floride est donnée par la table suivante. Trouver (a) la chute moyenne mensuelle (b) la médiane.

Mois	Jan	Fév	Mars	Avril	Mai	Juin	Juil	Août	Sep	Oct	Nov	Déc
pluie (cm)	5	4	6.2	11.4	22.4	25	20	23.4	26	20	6.5	4

9 On donne le nombre d'enfants pour 30 familles habitant une maison de rapport de la rue Papineau à Montréal.

3	1	2	0	4	6	1	0	1	5	2	3	2	0	7
2	2	0	3	1	3	2	4	2	4	0	4	3	1	4

Trouver (a) la moyenne (b) la médiane (c) le mode.

10. Les résultats d'un examen de statistique d'une classe d'étudiants du collège Maisonneuve.

24	30	32	40	48	50	58	60	60	62	63	64	64	67	67	67
70	70	71	72	74	74	75	75	76	77	78	78	78	78	79	80
81	83	83	83	84	84	84	84	85	86	87	88	88	88	89	90
92	92	93	95	95	97	98	99	100	100						

Trouver (a) le mode, (b) la médiane (c) Q_1 et P_{10} (d) le rang centile du résultat 93.

11. Discuter les énoncés suivants:

(a) Un principal d'école dit à l'un de ses professeurs: «Toutes les classes de l'école ont subi le même examen et votre classe a un résultat sous la moyenne. Je ne veux qu'aucune classe soit sous la moyenne.»

(b) René déclare: «Je suis dans le premier quartile de ma classe.»

(c) Anne déclare: «Mon rang centile dans une classe de 90 étudiants est 100.»

12. Dans chacun des cas suivants, dire si la mesure de tendance centrale donnée est possible compte tenu des restrictions imposées. Sinon dire pourquoi. Si oui donner un ensemble de résultats répondant à toutes les conditions.

(a) $n = 5$; le plus petit résultat = 4; étendue = 10; $\overline{X} = 14$
(b) $n = 5$; le plus petit résultat = 4; étendue = 10; $Md = 14$
(c) $n = 5$; le plus petit résultat = 50; le plus grand résultat = 100; $\overline{X} = 55$
(d) $n = 8$; étendue = 6; le mode n'est pas représentatif, car aucun résultat n'est répété.

13. Dans chacun des cas suivants, donner deux exemples de résultats qui répondent aux exigences suivantes. (Exemple: $n = 6$, étendue = 10, mode = 4, étendue = différence entre le plus grand et le plus petit résultat... Réponses: 3; 4; 4; 4; 7; 13; ou 0; 2; 4; 4; 4; 10.)

(a) $n = 5$; étendue = 7; $Md = 54$
(b) $n = 3$; étendue = 10; $\overline{X} = 90$
(c) $n = 8$; étendue = 20; $Mo = 42$; $Md = 45$
(d) $n = 4$; $\overline{X} = 10$; $Md = 8$

● 14. Trouver la moyenne de chacune des distributions (a) en utilisant la définition $\left(\overline{X} = \frac{\Sigma fX}{n}\right)$; (b) faire un choix judicieux pour A.

(i) classes (X)	f
1–5	4
6–10	4
11–15	2

(ii) classes (Y)	f
30–39	10
40–49	12
50–59	14
60–69	24

(iii) classes (Z)	f
0–20	9
21–40	5
41–60	6
61–80	10

15. (a) Parmi les valeurs suivantes possibles pour A dans le calcul de la moyenne des distributions suivantes, choisir celle qui vous semble la plus appropriée. 59,5; 34,5; 14; 40; 24,5; 60.

(i) classes (X)	f
10–12	4
13–15	7
16–18	3
19–21	2

(ii) classes (Y)	f
10–29	2
30–49	3
50–69	8
70–89	7

(iii) classes (Z)	f
40–49	2
30–39	5
20–29	4
10–19	3

(b) Utiliser la méthode courte pour trouver la moyenne de chaque distribution (soustraire A de chaque point milieu).

16. Le temps écoulé entre deux réparations pour une flotte de voitures de location a été enregistré de la façon suivante:

Temps (jours)	Nombre de voitures
0–14	8
15–29	6
30–44	32
45–59	16
60–74	8
75–89	10

Trouver le temps moyen entre deux réparations.

17. Deux méthodes différentes ont été utilisées pour enseigner le tir à l'arc à des groupes de débutants. Leur résultat individuel a été enregistré pour une distance de 30 mètres.

Méthode I	Homme	100	95	110	140	80	100	
	Femme	110	60	120	100	90	70	130	110
Méthode II	Homme	100	70	110	130	90	100	130	120
	Femme	90	100	110	100	150	60	120	120

(a) Comparer le résultat moyen des deux méthodes en prenant (1) les hommes (2) les femmes (3) tous les étudiants.

(b) Trouver le résultat moyen de (1) tous les hommes indépendamment de la méthode (2) toutes les femmes.

18. Les élèves d'un cours d'introduction à la musique participent à l'évaluation du cours en utilisant la pondération suivante: (échelle de 1 à 6) où 1 indique une insastisfaction totale et 6 une satisfaction totale. À la question «Votre professeur se rend-il disponible aux étudiants intéressés ou à ceux qui ont des problèmes?», voici les résultats obtenus:

Pondération:	1	2	3	4	5	6
Fréquences:	12	4	8	15	24	17

Quelle est la pondération moyenne à cette question? Que peut-on conclure face à la disponibilité du professeur?

19. Trouver la moyenne et le mode brut (point milieu de la classe modale) des distributions suivantes:

(a) classes X	f	(b) classes Y	f	(c) classes Z	f	(d) classes W	f
100–102	30	10–14	2	0,5055—0,5059	1	10–19	6
103–105	20	15–19	5	0,5050—0,5054	2	20–29	0
106–108	60	20–24	8	0,5045—0,5049	8	30–39	8
109–111	40	25–29	0	0,5040—0,5044	17	40–49	4
112–114	10	30–34	5	0,5035—0,5039	8	50–59	7
				0,5030—0,5034	0	60 et plus	3
				0,5025—0,5029	4		

20. Pendant six ans la compagnie Whitney a acheté des obligations de la ville de Montréal comme suit:

Année	Nombre d'obligations	Prix total
1970	140	$1400
1971	100	1100
1972	300	3800
1973	200	2900
1974	160	1800
1975	700	1500

Quel est le coût moyen d'une obligation?

21. (a) Trouver le résultat moyen de: 1; 1; 1; 1; 2; 3; 4; 4; 4; 4; 5; 6.
(b) Regrouper les résultats précédents dans les classes 1—3 4—6, puis trouver la moyenne.
(c) Expliquer la différence entre les deux moyennes.

● 22. Pendant une période de deux semaines, un dépanneur a vendu le nombre suivant de litres de lait: 78; 59; 47; 56; 76; 84; 76; 62; 50; 54; 74; 86.
(a) Quel est le nombre moyen de litres de lait vendus par le dépanneur?
(b) Combien de litres le gérant doit-il commander chaque jour, s'il désire n'avoir un surplus que dans un maximum de 10% des jours?

23. Étudier la preuve suivante:

$$\begin{aligned}\Sigma(X-\overline{X}) &= \Sigma X - \Sigma\overline{X} \quad \text{(règle 1, sec. 1,4)}\\ &= \Sigma X - n\overline{X} \quad \text{(règle 3, sec. 1,4)}\\ &= \Sigma X - n\left(\frac{\Sigma X}{n}\right) \qquad \left(\overline{X} = \frac{\Sigma X}{n}\right)\\ &= 0\end{aligned}$$

Puis prouver que si $\Sigma(X - A) = 0$ où A est une constante, alors $A = \overline{X}$

24. La moyenne des résultats de 1 000 étudiants de la faculté des arts à la Faculté des Arts à l'Université de Laval était au dernier semestre 2,85; elle était de 2,90 pour 200 étudiants de la faculté de commerce, et de 3,10 pour 400 étudiants de la faculté des sciences. Quel est le résultat moyen de l'ensemble des étudiants des trois facultés?

25. Les cotes d'écoute (en pourcentage) pour l'émission «Debout c'est l'heure» au poste RWJT durant une période de 8 semaines furent: 12,4; 14,5; 10,6; 18,6; 16,1; 15,2; 13,3; 10,0. Quelle est la cote (a) médiane (b) moyenne (c) le rang centile de la cote 12,4?

RÉPONSES

1. (a) $\overline{X} = \frac{\Sigma(X-A)}{n} + A$. (b) $\overline{X} = \frac{\Sigma f(X-A)}{n} + A$.

3. (a) (I) Mo = \$70; Md = \$65; $\overline{X}$ = \$48,20
(II) Mo = 50 et 100 (bimodale); Md = 75; $\overline{X}$ = 75,0
(III) Mo = 0,4 et 0,5 (bimodale); Md = 0,65; Pas de moyenne, classe ouverte
(IV) Mo = pas de signification; Md = 510; $\overline{X}$ = 483,9
(V) Mo = vw; Pas de médiane, ni moyenne (données à caractères spécifiques)
(b) (I) La médiane car les travailleurs occasionnels affectent trop la moyenne
(II) La moyenne ou la médiane
(III) La médiane (IV) La médiane
(V) Le mode

5. (a) Mo (b) $\overline{X}$ (c) $\overline{X}$
7. (a) 43 (b) 170 (c) 20
9. (a) 2,4 (b) 2 (c) 2

11. (a) Toutes les classes ne peuvent être au-dessus de la moyenne; la seule possibilité serait que toutes les classes aient la même moyenne, ce qui est peu probable.

(b) Il peut être au premier quartile et non dans le premier quartile. Il veut certainement dire aussi qu'il est parmi les 25% plus forts de sa classe.

(c) Même si elle est première de sa classe, son rang centile sera $\frac{89}{90} * 100 = 99$, et non 100.

13. (Il y a plusieurs réponses possibles pour chaque cas.) À titre indicatif on donne les résultats suivants: (a) 48; 50; 54; 54; 55 ou 54; 54; 54; 57; 61 (b) 85; 90; 95 ou 86; 88; 96 (c) 42; 42; 42; 44: 46; 48; 50; 62 ou 40; 42; 42; 42; 48; 50; 55; 60 (d) 7; 8; 8; 17 ou 4; 6; 10; 20

15. (a) (I) 14 (II) 59,5 (III) 34,5 ou 24,5

(b) (I) $\overline{X} = 9/16 + 14 = 14{,}6$ (II) $\overline{Y} = -0{,}20 + 59{,}5 = 59{,}5$
(III) $\overline{Z} = -80/14 + 34{,}5 = 28{,}8$

17. (a)

	Hommes	Femmes	Tous les étudiants
Méthode I	$25/6 + 100 = 104{,}2$	$-10/8 + 100 = 98{,}8$	$(25 - 10)/14 + 100 = 101{,}1$
Méthode II	$50/8 + 100 = 106{,}2$	$50/8 + 100 = 106{,}2$	$106{,}2$

(b) Hommes: $(25 + 50)/14 + 100 = 105{,}4$
Femmes: $(-10 + 50)/16 + 100 = 102{,}5$

19. (a) $\overline{X} = -60/100 + 107 = 106{,}6$; $Mo = 107$
(b) $\overline{Y} = 5/20 + 22 = 22{,}2$; Mo = 22
(c) $\overline{Z} = -0{,}0025/40 + 0{,}5042 = 0{,}50414$; $Mo = 0{,}5042$
(d) Pas de moyenne (classe ouverte) $Mo = 34{,}5$

21. (a) $36/12 = 3$ (b) $(2 * 6 + 5 * 6)/12 = 3{,}5$ (c) Elles sont différentes, car dans le calcul de la moyenne par classe, chaque résultat est considéré comme étant au point milieu de celle-ci; et dans ce cas, dans chaque classe il y a les 2/3 des résultats qui sont inférieurs au point milieu.

23. $\Sigma(X - A) = 0$ (Donné)
$\Sigma X - \Sigma A = 0$ (règle n° 1, art. 1.4)
$\Sigma X - nA = 0$ (règle n° 3, art. 1.4)
$A = \frac{\Sigma X}{n}$
$A = \overline{X}$

25. (a) 13,9 (b) $-1.3/8 + 14 = 13{,}84$
(c) $(2/8) * 100 = 25$

Chapitre 4

Mesure de dispersion

Pour résumer un ensemble de résultats, il est nécessaire de donner deux mesures: une qui traduit le centre de la distribution (la moyenne, la médiane ou le mode) et une autre qui traduit la dispersion ou l'éparpillement des données. Nous examinerons cinq mesures de dispersion: le rapport de dispersion, l'étendue, l'intervalle semi-interquartile, la variance et l'écart-type. Nous porterons plus d'attention à la variance et à l'écart-type compte tenu de leur utilité dans les chapitres suivants. Ce chapitre se termine par une brève présentation sur la dissymétrie.

4.1 INTRODUCTION

Oklahoma et les îles Canaries ont les mêmes températures moyennes annuelles. Faut-il en conclure que nous serons aussi confortables durant toute l'année aux deux endroits? La température à Oklahoma peut varier entre −10°C et 35°C et dans les îles Canaries entre 16°C et 30°C, donc pendant la majeure partie de l'année les deux climats sont très différents.

Les distributions suivantes ont le même mode, médiane, et le même nombre de résultats.

X: 40; 50; 50; 50; 50; 60
Y: 0; 30; 50; 50; 70; 100

Ces distributions diffèrent par l'éparpillement des données par rapport au centre.

4.2 LE RAPPORT DE VARIATION

Le **rapport de variation** (symbole: V) est la proportion de résultats non-modaux. C'est la seule mesure de dispersion qui peut être utilisée pour des données à caractères spécifiques.

$$V = \frac{\text{(Nombre de résultats) — (fréquence modale)}}{\text{nombre de résultats}}$$

$$V = \frac{n - f_{Mo}}{n}$$

V est un nombre entre 0 et 1. Si V est près de 0, la plupart des résultats sont égaux au mode; si V est près de 1, le mode est moins représentatif.

Exemple 1

Calculer le rapport de variation de chacune des distributions suivantes:

Catégories	Fréquence		
	(a)	*(b)*	*(c)*
A	1	1	2
B	1	3	5
C	9	7	3
D	1	1	2

(a) $V = \frac{12 - 9}{12} = 0{,}25$

(b) $V = \frac{12 - 7}{12} = 0{,}42$

(e) $V = \frac{12 - 5}{12} = 0{,}58$

4.3 L'ÉTENDUE

L'Étendue d'une distribution est la différence entre le plus grand résultat et le plus petit.

Exemple 1

X: 5; 12; 13; 13; 14; 15; 15; 15; 18; 20.

L'étendue est 20 − 5 = 15.

Exemple 2

Y: 5; 5; 11; 11; 11; 19; 19; 19; 20; 20.

L'étendue est 20 − 5 = 15.

Exemple 3

classes	f
1–3	4
4–6	7
7–9	3
10–12	2

L'étendue est 12,5 – 0,5 = 12. (*N.B.* On utilise ici la borne supérieure de la dernière classe et la borne inférieure de la première classe.)

L'étendue est facile à comprendre et à calculer. Elle dépend seulement des résultats extrêmes. Un exemple: deux échantillons de même grandeur d'une même population peuvent avoir des étendues très différentes. Par contre, les résultats des exemples 1 et 2 ont la même moyenne (14) et la même étendue (15) mais semblent être très différents.

4.4 L'INTERVALLE SEMI-INTERQUARTILE Q.

L'intervalle semi-interquartile (symbole Q) est la moitié de la différence entre le troisième et le premier quartile des données.

$$Q = \frac{Q_3 - Q_1}{2}$$

Exemple 1

X: 5; 12; 13; 13; 14; 15; 15; 15; 18; 20.

$$Q_3 = 15,\ Q_1 = 13,\ Q = \frac{15 - 13}{2} = 1{,}0$$

Exemple 2

Y: 5; 5; 11; 11; 11; 19; 19; 19; 20; 20.

$$Q_3 = 19,\ Q_1 = 11,\ Q = \frac{19 - 11}{2} = 4{,}0$$

L'intervalle semi-interquartile (aussi appelé **écart des quartiles)** est plus difficile à comprendre et à calculer que l'étendue. Il n'est pas affecté par les données extrêmes puisqu'il dépend de Q_3 et Q_1. Les deux mesures suivantes de dispersion dépendent de la moyenne; on utilise l'intervalle semi-interquartile quand la moyenne ne peut être calculée (par exemple, quand on a une classe ouverte). Souvent on utilise l'intervalle semi-interquartile comme mesure de dispersion, quand la médiane est utilisée comme mesure de tendance centrale.

4.5 CRITÈRES POUR UNE MESURE DE DISPERSION

L'étendue dépend seulement des résultats extrêmes, l'intervalle semi-interquartile

dépend de 50% des données situées au milieu de la distribution. Recherchons maintenant une mesure qui dépendra de tous les résultats. Cette mesure devra être petite si les résultats sont près les uns des autres et grande si les résultats sont très éparpillés; elle ne devra pas être nécessairement grande même si le nombre de données est grand.

Que penser de la somme des déviations de la moyenne? La somme $\Sigma(X - \overline{X})$ est toujours égale à zéro tel que démontré dans le problème 23 section 3.8. Un exemple:

$$X\text{: } 1;\ 2;\ 6;\ \overline{X} = 3;\ \Sigma(X - \overline{X}) = (1 - 3) + (2 - 3) + (6 - 3) = 0$$

Maintenant, nous pouvons ignorer le signe moins et additionner les valeurs absolues des déviations de la moyenne. Ainsi dans l'exemple précédent nous aurons $|1 - 3| + |2 - 3| + |6 - 3| = 2 + 1 + 3 = 6$. Si nous avons 1 000 résultats, cette somme sera grande même si chacun des résultats est compris entre 1 et 6. Si nous divisons la somme des valeurs absolues des déviations de la moyenne par le nombre de résultats, nous obtenons une mesure de dispersion appelée **l'écart-moyen.** Elle satisfait les critères établis précédemment mais est peu utilisée parce que les valeurs absolues sont peu propices aux calculs algébriques. De plus $\dfrac{\Sigma|X-A|}{n}$ n'est pas minimum quand A est la moyenne, mais quand A est la médiane. (Voir problème 20 section 4.10)

4.6 VARIANCE ET ECART-TYPE DE DONNÉES NON GROUPÉES

Tout nombre, positif ou négatif, devient positif quand il est élevé au carré: $(+4)^2 = +16$; $(-2)^2 = +4$; $(-1{,}2)^2 = +1{,}44$. Au lieu de prendre la moyenne des valeurs absolues des déviations de la moyenne, nous pouvons prendre la moyenne des carrés de ces déviations de la moyenne;nous obtenons ainsi la moyenne des carrés des déviations $\dfrac{\Sigma(X - \overline{X})^2}{n}$. C'est la **variance.** Comme nous le verrons au chapitre 7, dans le cas d'un échantillon de taille n la variance (symbole s^2) est obtenue par la formule.

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

Pourquoi avons-nous remplacé n par $n - 1$ au dénominateur? Nous voudrons plus tard évaluer la variance d'une population à partir de la variance d'un échantillon. Si «n» est le dénominateur de la variance d'un échantillon, l'estimation sera trop faible. Il est à noter qu'une fraction augmente si le dénominateur diminue ($^1/_2$ est plus grand que $^1/_3$). En changeant n par $n - 1$ au dénominateur, la fraction augmente juste assez pour que la variance de l'échantillon devienne une bonne estimation de la variance d'une population.

La variance satisfait tous les critères énoncés à la section 4.5 concernant la mesure de dispersion. Il y a toutefois un inconvénient, en effet: son résultat est en unités carrées. Si les données X sont mesurées en centimètres, $\overline{X}$ sera en centimètres et $X - \overline{X}$ également, alors que $(X - \overline{X})^2$ et la variance seront en centimètres carrés. Pour revenir à l'unité initiale, on prend la partie positive de la racine carrée de la variance. Cette valeur est appelée **l'écart-type**. C'est une mesure de dispersion qui a les mêmes unités que les données. L'abréviation pour l'écart-type d'un échantillon est s.

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}}$$

Il est à noter que pour trouver l'écart-type, il faut en premier trouver la moyenne $\overline{X}$ puis trouver $(X - \overline{X})$, $(X - \overline{X})^2$ finalement on additionne $\Sigma\ (X - \overline{X})^2$. On utilisera la table 1 appendice C pour extraire la racine carrée.

Exemple 1

Trouver la variance et l'écart-type des résultats X:1; 2; 6.

$\overline{X} = 9/3 = 3.$

X	$X - \overline{X}$	$(X - \overline{X})^2$
1	−2	4
2	−1	1
6	3	9
9		14

$$s^2 = \frac{\Sigma\ (X - \overline{X})^2}{n - 1} = \frac{14}{2} = 7;\ s = \sqrt{7} = 2{,}6$$

Exemple 2

Trouver la variance et l'écart-type de Y: 10; 12; 14; 18.

$\overline{Y} = 54/4 = 13{,}5$

Y	$Y - \overline{Y}$	$(Y - \overline{Y})^2$
10	−3,5	12,25
12	−1,5	2,25
14	0,5	0,25
18	4,5	20,25
54		35,00

$$s^2 = \frac{\Sigma\ (Y - \overline{Y})^2}{n - 1} = \frac{35}{3} = 11{,}67;\ s = \sqrt{11{,}67} = 3{,}4$$

Attention, ne pas arrondir trop tôt. Si on veut trouver s à une décimale près, il faut exprimer s^2 en conservant deux décimales. Ce dernier exemple implique plus de calculs que le premier; il ne serait pas difficile de trouver des exemples où les calculs seraient longs. Heureusement il existe une autre forme de formule pour calculer plus facilement la variance et l'écart-type. Cette nouvelle formule est:

$$s^2 = \frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}}{n - 1}$$

$$s = \sqrt{\frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}}{n - 1}}$$

Les carrés des nombres sont fournis dans la table 1 appendice C; ces formules favorisent l'utilisation d'une calculatrice. Il faut bien distinguer entre ΣX^2 (on élève au carré puis on additionne) et $(\Sigma X)^2$ (on additionne puis on élève au carré).

Examinons maintenant d'où viennent ces formules en se rappelant que $\bar{X}$ est une constante et que $\Sigma \bar{X}^2 = n\bar{X}^2$ (selon la règle 3, section 1.4).

$$\begin{aligned}
\Sigma(X-\bar{X})^2 &= \Sigma(X^2 - 2X\bar{X} + \bar{X}^2) \\
&= \Sigma X^2 - \Sigma 2X\bar{X} + \Sigma\bar{X}^2 \\
&= \Sigma X^2 - 2\bar{X}\Sigma X + n\bar{X}^2 \\
&= \Sigma X^2 - 2\left(\frac{\Sigma X}{n}\right)\Sigma X + n\left(\frac{\Sigma X}{n}\right)^2 \qquad \left(\bar{X} = \frac{\Sigma X}{n}\right) \\
&= \Sigma X^2 - 2\frac{(\Sigma X)^2}{n} + \frac{(\Sigma X)^2}{n} \\
&= \Sigma X^2 - \frac{(\Sigma X)^2}{n}
\end{aligned}$$

En divisant par n−1, on obtient la variance;
en effet:

$$s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}}{n-1}$$

Exemple 3

Trouver la moyenne et l'écart-type de l'échantillon Z: 1; 2; 7

Z	Z^2
1	1
2	4
7	49
10	54

$$\bar{Z} = \frac{\Sigma Z}{n} = \frac{10}{3} = 3{,}3$$

$$s = \sqrt{\frac{\Sigma Z^2 - \dfrac{(\Sigma Z)^2}{n}}{n-1}} = \sqrt{\frac{54 - 10^2/3}{2}} = \sqrt{10{,}33} = 3{,}2$$

Exemple 4

Comparer les deux méthodes pour calculer la variance de l'échantillon X: 1; 2; 3; 5; 10.

Définition:

X	$X-\overline{X}$	$(X-\overline{X})^2$
1	−3,2	10,24
2	−2,2	4,84
3	−1,2	1,44
5	0,8	0,64
10	5,8	33,64
21		50,80

$$\overline{X}=\frac{21}{5}=4,2$$

$$s^2=\frac{50,80}{4}=12,7$$

Transformation:

X	X^2
1	1
2	4
3	9
5	25
10	100
21	139

$$s^2=\frac{139-21^2/5}{4}=\frac{139-88,2}{4}$$

$$s^2=12,7$$

Dans le calcul de la moyenne, nous avons trouvé que le fait de soustraire une constante de chaque donnée simplifie grandement les opérations. Le même principe peut être utilisé dans le calcul de la variance. On peut remarquer que contrairement à la moyenne, la variance n'est pas affectée si on soustrait une constante à chaque donnée. En effet chaque donnée est diminuée d'une constante A, mais l'éparpillement de celles-ci demeure inchangé. On doit choisir A de manière à simplifier les calculs.

Exemple 5

Trouver l'écart-type de X: 15,5; 17,5; 21,5; 23,5. a) directement b) en soustrayant 17,5 de chaque résultat.

(a)

X	X^2
15,5	240,25
17,5	306,25
21,5	462,25
23,5	552,25
78,0	1561,00

(b)

$X-17,5=Y$	Y^2
−2,0	4,00
0,0	0,00
4,0	16,00
6,0	36,00
8,0	56,00

$$s_X^2=\frac{1561-(78)^2/4}{3}=\frac{40}{3}=13,3333 \qquad s_X=3,65$$

$$s_Y^2=\frac{56-8^2/4}{3}=\frac{40}{3}=13,3333 \qquad s_Y=3,65$$

Noter l'utilisation des indices pour distinguer l'écart-type de l'échantillon X s_X de celui de l'échantillon Y s_Y.

Exemple 6

Trouver la variance de la distribution Z: 42; 47; 54; 53; 48; 51; 50; 49; 52.

Z	$Z - 50 = X$	X^2
42	−8	64
47	−3	9
54	+4	16
53	+3	9
48	−2	4
51	1	1
50	0	0
49	−1	1
52	2	4
	−4	108

$$s^2{}_z = s^2{}_x = \frac{108 - (-4)^2/9}{8} = \frac{108 - 1{,}78}{8} = 13{,}28$$

4.7 VARIANCE ET ÉCART-TYPE DE DONNÉES GROUPÉES

Considérer les données de chaque classe comme si les valeurs étaient équivalentes à la valeur du point milieu (cette façon de procéder a été utilisée auparavant pour trouver la moyenne de données groupées).

Exemple 1

Trouver la variance et l'écart-type de la distribution suivante:

classes	f	X
4–6	2	5
7–9	5	8
10–12	3	11
	10	

Cette distribution aura la même variance et écart-type que les résultats 5; 5; 8; 8; 8; 8; 8; 11; 11; 11. La moyenne est 8,3; nous devons trouver la somme de:

$$\underbrace{(5 - 8{,}3)^2 + (5 - 8{,}3)^2}_{\text{2 termes}} + \underbrace{(8 - 8{,}3)^2 + \ldots + (8 - 8{,}3)^2}_{\text{5 termes}} +$$

$$\underbrace{(11 - 8{,}3)^2 + (11 - 8{,}3)^2 + (11 - 8{,}3)^2}_{\text{3 termes}}$$

puis la diviser par 9 pour trouver la variance. L'addition répétée est remplacée par une multiplication. Pour faciliter le calcul, on ajoute des colonnes supplémentaires.

classes	f	X	fX	$X-\overline{X}$	$(X-\overline{X})^2$	$f(X-\overline{X})^2$
4–6	2	5	10	−3,3	10,89	21,78
7–9	5	8	40	−0,3	0,09	0,45
10–12	3	11	33	2,7	7,29	21,87
	10		83			44,10

$$\overline{X} = 83/10 = 8,3$$

$$s^2 = 44,10/9 = 4,90$$

$$s = \sqrt{4,90} = 2,2$$

Noter que les résultats de la colonne $f\ (X - \overline{X})^2$ sont obtenus par $2\,(5 - 8,3)^2$, $5\,(8 - 8,3)^2$ et $3\,(11 - 8,3)^2$. Ces résultats sont additionnés, puis la somme divisée par 9 pour obtenir la variance.

Les formules générales pour obtenir la variance et l'écart-type de données groupées sont:

$$s^2 = \frac{\Sigma f(X-\overline{X})^2}{n-1}, \quad s = \sqrt{\frac{\Sigma f(X-\overline{X})^2}{n-1}}$$

Note: dans ces formules X représente le milieu de la classe.

Exemple 2

Si $\overline{X} = 8,3$; trouver s^2 de la distribution suivante:

classes	f	X
4–6	2	5
7–9	5	8
10–12	3	11
	10	

Réponse: $n = 10$; colonne X: 5; 8; 11; colonne $X - \overline{X}$: −3,3; −0,3; 2,7;
colonne $f(X - \overline{X})$: −6,6; −1,5; 8,1;
colonne $f(X - \overline{X})^2$: 21,78; 0,45; 21,87; $s^2 = \frac{44,10}{9} = 4,90$

La formule pour des données classées devient:

$$s^2 = \frac{\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}}{n-1}$$

où X est le milieu de la classe. Utiliser la table des carrés (Appendice C) ou une calculatrice.

Exemple 3

Trouver l'écart-type de la distribution suivante:

classes	f	X	fX	fX^2
4–6	2	5	10	50
7–9	5	8	40	320
10–12	3	11	33	363
	10		83	733

$s^2 = \dfrac{733 - 688{,}9}{9} = 4{,}90$; $s = 2{,}2$.

Exemple 4

Trouver $s = \sqrt{\dfrac{\Sigma fX^2 - (\Sigma fX)^2/n}{n-1}}$.

classes	f	X	fX	fX^2
0–4	4			
5–9	10			
10–14	6			

Réponses

X: 2; 7; 12; fX: 8; 70; 72; $\Sigma fX = 150$; fX^2: 16; 490; 864; $\Sigma fX^2 = 1370$

$$s = \sqrt{\frac{1370 - (150)^2/20}{19}} = \sqrt{\frac{1370 - 1125}{19}} = \sqrt{12{,}89} = 3{,}6.$$

Comme nous l'avons déjà fait, la soustraction d'une constante A de chaque X n'affecte pas l'écart-type de la distribution. Les calculs peuvent être passablement simplifiés si A est choisi judicieusement. Évaluer la moyenne; un bon choix pour A est souvent le milieu de la classe ayant la plus grande fréquence ou contenant la moyenne.

Exemple 5

Trouver l'écart-type de la distribution suivante (a) directement (b) en soustrayant 124,5 de chaque X.

(a)

classes	f	X	fX	fX^2
100–109	3	104,5	313,5	32 760,75
110–119	5	114,5	572,5	65 551,25
120–129	7	124,5	871,5	108 501,75
130–139	2	134,5	269,0	36 180,50
140–149	6	144,5	867,0	125 281,50
	23		2893,5	368 275,75

$$s_x^2 = \frac{368\,275{,}75 - (2893{,}5)^2/23}{22} = \frac{4260{,}87}{22} = 193{,}68;\ s_x = 13{,}9$$

Soyez prudent en arrondissant aux étapes intermédiaires; si on arrondit 2893,5 à 2894,0 on trouvera $s_X^2 = 4135{,}05/22 = 187{,}96$ et $s_X = 13{,}7$.

(b)

$X - 124{,}5 = Y$	fY	fY^2
−20	−60	1200
−10	−50	500
0	0	0
10	20	200
20	120	2400
	30	4300

$$s_Y^2 = \frac{4300 - (30)^2/23}{22} = \frac{4260{,}87}{22} = 193{,}68; \quad s_Y = 13{,}9$$

Exemple 6

Refaire l'exemple 4, en soustrayant 7 de chaque X. Trouver s.

classes	f	X	$X - 7 = Y$	fY	fY^2
0–4	4				
5–9	10				
10–14	6				

$s_X = s_Y = ?$

Réponses

Colonne Y: −5; 0; 5; fY: −20; 0; 30; ΣfY: 10; fY^2: 100; 0; 150

$$s^2 = \frac{250 - (10)^2/20}{19} = \frac{245}{19} = 12{,}89 \qquad s = 3{,}6.$$

4.8 DISSYMÉTRIE

Quelquefois, une distribution présente une dissymétrie à droite si celle-ci a une «queue» plus longue qu'à gauche, réciproquement elle peut présenter une dissymétrie à gauche. Dans ce cas, la moyenne et la médiane peuvent être très différentes, car les résultats dans la «queue» la plus longue, affecteront la moyenne, mais pas la médiane. Cette différence est exprimée dans la mesure de dissymétrie suivante:

$$\text{Dissymétrie} = \frac{3(\overline{X} - Md)}{s}$$

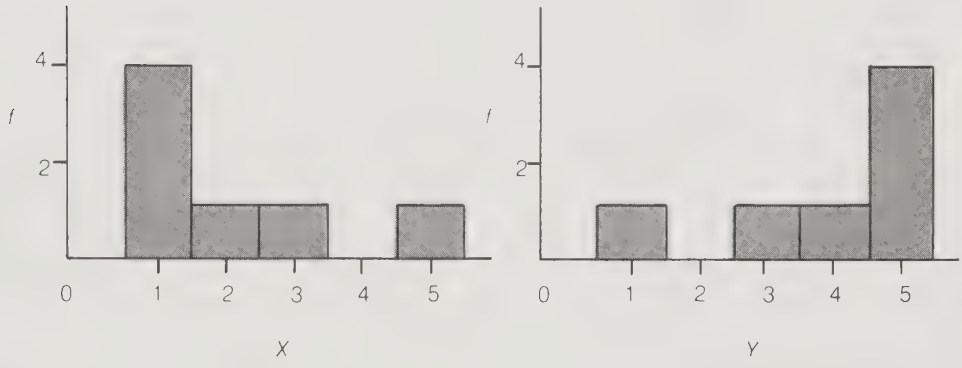
En divisant par s, la dissymétrie devient une mesure sans dimension.

Exemple 1

Trouver la mesure de dissymétrie et tracer un histogramme de chacune des distributions suivantes.

X: 1; 1; 1; 1; 2; 3; 5.

Y: 1; 3; 4; 5; 5; 5; 5.



$\bar{X} = 2{,}0 \quad Md = 1 \quad s = 1{,}5$ (à vérifier); Dissymétrie $= \dfrac{3\,(2 - 1)}{1{,}5} = 2{,}0$

$\bar{Y} = 4{,}0 \quad Md = 5 \quad s = 1{,}5$ (à vérifier); Dissymétrie $= \dfrac{3\,(4 - 5)}{1{,}5} = -2{,}0$

Il est à noter qu'une dissymétrie à droite est positive et une à gauche est négative.

4.9 VOCABULAIRE ET SYMBOLES

rapport de variation	V, f_{Mo}	intervalle semi-interquartile
étendue	Q, S	écart type
carrés des écarts à la moyenne		écart moyen
variance	s^2	mesure de dissymétrie

4.10 EXERCICES

1. Trouver le rapport de variation des résultats de l'exemple 1 section 3.2.

2. Trouver le rapport de variation des résultats du test de personnalité tel que groupés a) en 9 classes b) en 6 classes dans la section 2.4.

●3. Trouver a) l'étendue b) l'intervalle semi-interquartile c) la moyenne d) l'écart-type à partir de la définition e) l'écart-type par la méthode courte de chacune des distributions suivantes:

(i) X: 1; 4; 9; 18.
(ii) Y: 51; 54; 59; 62; 63; 63; 67; 69.
(iii) Z: 30; 24; 21; 25; 20; 30; 22; 28.
(iv) W: 14; 16; 17; plus de 18.

4. Trouver a) l'étendue b) l'intervalle semi-interquartile c) l'écart-type d) la mesure de dissymétrie des points de fusion et d'ébullition des alcools suivants:

Alcools	Fusion °C	Ébullition °C
Méthylique	− 97	64,7
Éthylique	−114	78,3
Isopropylique	− 89	82,3
Allylique	−129	97,0
Isobutylique	−108	107,9
Isoamylique	−117	131,5

5. Trouver a) l'étendue b) l'intervalle semi-interquartile c) l'écart-type d) la mesure de dissymétrie des résultats d'un examen de statistiques donnés dans l'exercice 3a (II) section 3.8.

6. Dénombrer puis trouver s pour le nombre d'enfants des 30 familles de l'exercice 9 article 3.8.

7. Les pressions artérielles systolique et diastolique furent prises sur 10 adultes et la différence d des deux pressions pour chaque individu fut enregistrée. On a trouvé que $\Sigma d = 300$ et $\Sigma d^2 = 9\,200$. Trouver la moyenne et la variance de ces différences.

8. La Upton Calculator Co. a reçu 5 livraisons de 1 000 piles électriques. Le nombre de piles défectueuses dans chaque livraison fut: 4; 12; 6; 4; 8. Quel est le nombre moyen de piles défectueuses par livraison et l'écart-type?

9. Un professeur de calcul a fait passer en début de session un test sur l'algèbre élémentaire à ses étudiants. Voici les résultats:

78	72	84	81	78	76
83	89	81	71	72	78
79	83	81	72	84	68
86	82				

Trouver l'écart-type de ces résultats.

10. Dans une expérience en psychologie, on a fait passer le même test à 10 sujets. Préalablement, on a averti 5 d'entre eux que ce test était très difficile et les 5 autres qu'il était très facile. Voici les résultats obtenus:

Très difficile	Très facile
38	30
52	38
46	40
42	24
38	32

Trouver l'écart-type de chacun des groupes puis pour l'ensemble des 10 sujets.

11. Trouver 3 résultats dont la moyenne est 10 et la variance est 1.

12. Trouver 2 résultats dont la moyenne est 40 et la variance 8.

13. La vitesse du son en m/sec dans différents métaux est donnée.

Métal	Vitesse (m/sec)
Aluminium	5000
Cadmium	2300
Cuivre	3600
Fer	5100
Plomb	1200
Nickel	5000
Platine	2700
Argent	2600
Étain	2500

Trouver s; conseil: si un ensemble de résultats est divisé par 100 et que s est calculé pour ce nouvel ensemble, on doit multiplier par 100 cette valeur de s pour obtenir l'écart-type de l'ensemble original.

14. (a) Trouver l'étendue, l'intervalle semi-interquartile et l'écart-type des salaires de l'exercice 4 article 3.8. Conseil: pour trouver s, poser $Y = \frac{X}{100}$ (X = salaires) puis $Z = Y - 80$ alors $s_Z = s_Y = s_X/100$. (b) Laquelle des mesures de dispersion trouvées en (a) représente le mieux la dispersion des salaires.

●15. Trouver la variance de chacune des distributions suivantes en utilisant $s^2 = \frac{\Sigma f(X - \overline{X})^2}{n - 1}$.

(a)

X	f
5–9	2
10–14	5
15–19	2

(b)

Y	f
6–10	2
11–15	4
16–20	3

(c)

Z	f
10, < 20	4
20, < 30	8
30, < 40	6
40, < 50	3
50, < 60	4

(d)

W	f
10–19	5
20–29	10
30–39	12
40 et plus	4

16. Trouver s de chacune des distributions précédentes en utilisant la méthode courte (en soustrayant une constante de chaque point milieu).

17. Trouver la variance de chaque distribution de l'exercice 14 section 3.8, choisir A adéquatement dans chaque distribution.

18. Six pièces de monnaie sont lancées 100 fois; à chaque fois le nombre de pièces présentant le côté face est noté. Voici les résultats.

Nombre de faces	0	1	2	3	4	5	6
Fréquences	3	7	24	34	21	10	1

Trouver le nombre moyen de faces et l'écart-type.

19. Supposons qu'une population est formée des 5 résultats suivants: 1; 2; 4; 6; 7. (a) Trouver (à deux décimales près) l'écart-type de la population (il faut se rappeler que dans le calcul de la variance d'une population, il faut diviser par 5 et non par 5 − 1).

(b) Calculer (à deux décimales près) l'écart-type de tous les échantillons de 2 éléments pris dans la population en utilisant $s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$ ou $s = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}}$.

(*N.B.* Il y a 10 échantillons possibles.)

(c) Calculer l'écart-type de tous les échantillons de 2 éléments en utilisant $s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n}}$ ou $s = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/n}{n}}$.

Résultats de l'échantillon	(b) $s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$	(c) $s = \sqrt{\frac{\Sigma (X-\bar{X})^2}{n}}$
1;2		
1;4		
1;6		
1;7		
2;4		
2;6		
2;7		
4;6		
4;7		
6;7		

(d) Calculer la moyenne des écarts-types trouvés en (b) et (c) et la comparer avec l'écart-type de la population trouvé en (a). Quelle formule donne la meilleure approximation pour le calcul de l'écart-type d'une population à partir d'un échantillon.

Réponses

1. $\frac{19 - 6}{19} = 0{,}68$

3. (i) (a) 17 (b) $(13{,}5 - 2{,}5)/2 = 5{,}5$ (c) 8 (d) $\sqrt{166/3} = 7{,}4$
(e) $\sqrt{(422 - 32^2/4)/3} = 7{,}4$
(ii) (a) 18 (b) $(65 - 56{,}5)/2 = 4{,}2$ (c) $8{,}0/8 + 60 = 61{,}0$
(d) $\sqrt{26}\ 2/7 = 6{,}1$ (e) (soustraire 60) $\sqrt{(270 - 8^2/8)/7} = 6{,}1$
(iii) (a) 10 (b) $(29 - 21{,}5)/2 = 3{,}8$ (c) 25,0 (d) $\sqrt{110/7} = 4{,}0$
(e) $\sqrt{(110 - 0^2/8)/7} = 4{,}0$
(iv) Non-déterminé (classe ouverte)

5 (a) 60 (b) $\frac{90 - 60}{2} = 15$ (c) 18,5 (d) $3\frac{(75 - 75)}{18{,}5} = 0$

7. Moyenne $= 300/10 = 30$; $s^2 = (9200 - 300^2/10)/9 = 22{,}22$

9. Soustraire 80 de chaque résultat $s = \sqrt{(620 - (-22)^2/20)/19} = 5{,}60$

11. 9; 10; 11.

13. $s = 1410$ m/s (1414,21 m/s)

15. (a) $X = 12$, $s^2 = 100/8 = 12{,}5$ (b) $Y = 13{,}6$; $s^2 = 122{,}24/8 = 15{,}28$ (c) $Z = 33{,}0$; $s^2 = 4200/24 = 175$ (d) Impossible classe ouverte.

17. (i) $A = 8; s^2 = \dfrac{150 - (-10)^2/10}{9} = 15{,}56$

(ii) $A = 54{,}5; s^2 = \dfrac{7600 - (-80)^2/60}{59} = 127$

(iii) $A = 50{,}5; s^2 = \dfrac{20762{,}25 - \dfrac{(-264{,}5)^2}{30}}{29} = 635{,}53$

19. (a) L'écart-type de la population est 2,28.
(b) 0,71; 2,12; 3,54; 4,24; 1,41; 2,83; 3,54; 1,41; 2,12; 0,71.
(c) 0,50; 1,50; 2,50; 3,00; 1,00; 2,00; 2,50; 1,00; 1,50; 0,50.
(d) La moyenne des s en (b) est 2,26. La moyenne des s en (c) est 1,60. Pour la population on avait trouvé en (a) 2,28 donc si on se base sur des échantillons pour trouver s de toute une population, il est évident que (b) donne un meilleur résultat. On a obtenu (b) en utilisant la formule $s = \sqrt{\dfrac{\Sigma(X - \bar{X})^2}{n - 1}}$ donc ce n'est pas une preuve, mais une confirmation de l'hypothèse qui veut que pour calculer s à partir d'un échantillon il vaut mieux diviser par $n - 1$.

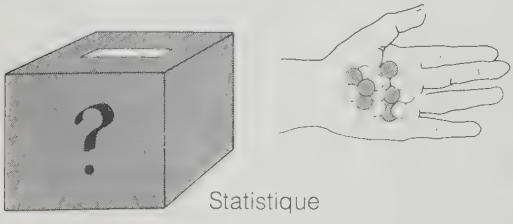
Chapitre 5

Probabilité

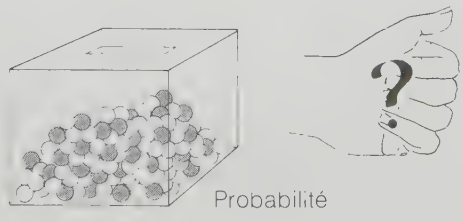
Depuis le début, nous avons étudié la statistique descriptive et maintenant nous sommes prêts à aborder l'étude de l'inférence statistique. À partir de données recueillies pour un échantillon, nous apprendrons comment se comporte la population d'où cet échantillon a été prélevé. La probabilité est le fondement de l'inférence statistique et nous l'utiliserons souvent dans les chapitres qui suivent. L'étude de la probabilité peut être très vaste et peut faire à elle seule l'objet d'un cours. Pour ce cours, nous n'aurons besoin que de quelques notions de base. Nous nous limiterons à l'étude des probabilités des événements A et B quand ceux-ci ne peuvent se produire en même temps et quand la réalisation de *A* n'a pas d'influence sur la réalisation de *B* et vice-versa.

5.1 POURQUOI ÉTUDIER LA PROBABILITÉ?

La probabilité et la statistique sont en étroite relation. Un exemple: en inférence statistique, on peut choisir 100 votants au hasard de la liste électorale et leur demander pour qui ils voteront à la prochaine élection; puis à partir de ces résultats, on prédira qui sera élu lors de cette élection. Après l'élection, on connaît exactement le nombre de votes qu'a eu chaque candidat A, B, C. Si on choisit 100 votants au hasard, on peut dire combien parmi eux ont probablement voté pour les candidats A, B et C. Un autre exemple: en statistique, si on pige d'une boîte opaque contenant des billes blanches ou rouges, en examinant le résultat de la pige, on tentera de découvrir ce qu'il y a dans la boîte. En probabilité, on comptera le nombre de boules blanches et le nombre de boules rouges que contient la boîte puis en pigeant au hasard un certain nombre de boules, sans regarder, on tentera de prédire combien de billes de chaque couleur ont été pigées. Nous apprendrons comment prédire le résultat des élections ou prédire le contenu d'une boîte, seulement après avoir étudié la probabilité. Nous pourrons alors choisir un échantillon, faire une estimation du comportement de la population et décider le degré de certitude de cette estimation.



Statistique



Probabilité

5.2 Définition de la probabilité

Si un événement A consiste en n expériences également probables et que parmi ces n cas possibles, il y a s cas favorables à la réalisation de A, alors la probabilité que A se produise est s/n. Soit $p\ (A)$ cette probabilité.

$$p\ (A) = \frac{s}{n} = \frac{\text{nombre de cas favorables}}{\text{nombre de cas total}}$$

Il est à noter que s est un nombre compris entre 0 et n et que $p\ (A)$ est un nombre compris entre 0 et 1. Si $s = 0$, l'événement A ne se produit pas et $p\ (A) = 0$ (événement impossible). Si A se produit toujours $s = n$ et $p\ (A) = 1$ (événement certain).

> $p\ (A)$ est un nombre compris entre 0 et 1. Si A est un événement certain $p\ (A) = 1$; si A est un événement **impossible** $p\ (A) = 0$.

Exemple 1

Si une pièce de monnaie est lancée, elle peut présenter un côté face (F) ou pile (P); ces deux résultats sont également probables si la pièce est régulière. Ainsi:

$p\ (\text{F}) = 1/2$ et $p\ (\text{P}) = 1/2$

Exemple 2

La population de Beloeil est de 15 000 habitants et celle de St-Hilaire de 10 000 habitants. Si une personne est choisie au hasard parmi tous les citoyens pour représenter ces deux villes à un congrès régional, quelle est la probabilité qu'elle habite à St-Hilaire?

$$p\ (\text{St-Hilaire}) = \frac{\text{nombre de citoyens de St-Hilaire}}{\text{nombre total de citoyens}} = \frac{10\,000}{15\,000 + 10\,000}$$

$$\frac{10\,000}{25\,000} = 0.40$$

> En général, une probabilité sera exprimée à deux décimales près.

La définition de la probabilité, $p\ (A) = s/n$ pour n événements également probables, est classique. La limitation de cette définition provient des mots «événements également probables». Si un homme de 40 ans demande à une compagnie d'assurance de l'assurer pour un an est-ce que celle-ci lui répondra: «Dans un an, vous serez mort ou vivant, la probabilité que vous soyez mort est 0,50; alors votre prime sera la moitié du montant assuré». C'est absolument absurde, car les deux événements ne sont pas également

probables. En effet, l'expérience antérieure prouve que sur 100 000 hommes âgés de 40 ans, 99 382 sont vivants un an plus tard alors que seulement 618 sont décédés. Alors la compagnie d'assurance conclut que la probabilité qu'un homme de 40 ans pris au hasard soit vivant un an plus tard est $\frac{99\,382}{100\,000} = 0{,}99$ et non 0,50 comme précédemment. Il y a une seconde façon de définir la probabilité; la probabilité d'un événement est la fréquence relative à laquelle l'événement se produit. Elle est basée sur l'expérience antérieure comme dans le cas précédent ou sur l'expérimentation comme dans l'exemple suivant.

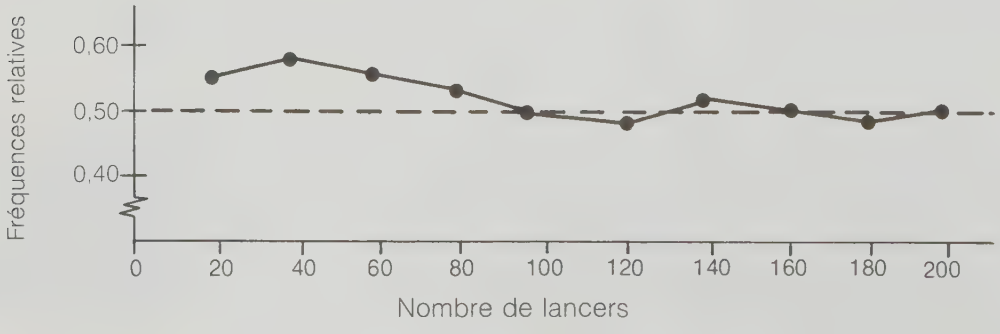
Exemple 3

On soupçonne une pièce de monnaie de ne pas être équilibrée (régulière). Elle est lancée 200 fois et le nombre de faces est compilé à tous les 20 lancers dans le tableau ci-dessous. Quelle est la probabilité d'obtenir une face quand la pièce est lancée?

Nombre de lancers	Nombre de faces dans les 20 derniers lancers	Nombre cumulatif de faces	Fréquence relative cumulées
20	11	11	0,55
40	12	23	0,58
60	10	33	0,55
80	9	42	0,52
100	7	49	0,49
120	9	58	0,48
140	14	72	0,51
160	7	79	0,49
180	8	87	0,48
200	11	98	0,49

En se basant sur **l'expérience précédente,** le meilleur résultat pour la probabilité d'obtenir une face en lançant cette pièce de monnaie est 98/200 = 0,49.

Le graphique ci-dessous donne la fréquence relative en fonction du nombre de lancers. Noter la fluctuation de la fréquence relative autour de 0,50 (valeur théorique) dans le cas d'une pièce régulière.



Quand on utilise la fréquence relative dans le calcul d'une probabilité, on voudrait avoir théoriquement un nombre infini d'essais. La réalisation d'un nombre fini d'essais produit une certaine erreur mais celle-ci diminue à mesure que le nombre d'essais croît.

Une troisième façon de définir la probabilité est la **probabilité subjective:** en se basant sur l'information ou la connaissance des circonstances d'une expérience, un nombre entre 0 (représentant une certitude que l'événement ne se produira pas) et 1 (représentant la certitude que l'événement se produira) est donné à la probabilité de réalisation à un événement. Par exemple, un agronome qui a produit un nouvel insecticide peut dire: «Le produit *A* tue les insectes en 30 secondes. Je lui ai ajouté un groupement sulfate: le nouveau produit agit plus rapidement et la probabilité que les insectes meurent en 10 secondes ou moins est de 0,80.» Le nouveau produit n'a jamais été expérimenté, on ne peut donc pas le comparer. Son expérience sur les insecticides lui permet de ne pas trop se tromper. Après expérimentation, il pourra toujours réajuster sa probabilité.

Dans la théorie moderne des probabilités, on les associe aux événements en respectant certains axiomes et règles que ce soit pour des probabilités d'événements également probables, probabilités provenant de fréquences relatives ou probabilités subjectives. La plupart des problèmes de probabilité qu'on aura à résoudre pour comprendre la statistique utilise la définition classique $p(A) = \frac{n}{N}$, mais d'autres principes pourront être utilisés pour certains problèmes plus complexes tout en élargissant le champ de nos connaissances.

5.3 ESPACE ÉCHANTILLONNAL ET ÉVÉNEMENT

Nous sommes familiers avec différents ensembles: l'ensemble de golf, les membres d'un club de tennis, etc. Un **ensemble** est une collection d'objets ou êtres possédant des caractéristiques bien définies. On doit pouvoir dire si un objet quelconque appartient ou non à un ensemble donné. Un ensemble peut être défini en énumérant les **éléments** qui lui appartiennent et en écrivant ceux-ci entre accolades. Ainsi si on a «*A* est l'ensemble des entiers compris entre 4 et 8 inclusivement» on peut écrire A = {entiers compris entre 4 et 8 inclusivement} $A = \{x \in N \rightarrow x \in [4; 8]\} = \{4; 5; 6; 7; 8\}$. B est un **sous-ensemble** de A si tous les éléments de B sont compris dans A. Ainsi $B = \{4; 5; 7\}$ est sous-ensemble de A; on écrira également $B \subset A$.

L'ensemble de tous les résultats d'une expérience s'appelle l'**espace échantillonnal** S. Un événement est un sous-ensemble de l'espace échantillonnal. Les **diagrammes de Venn** utilisés dans la théorie des ensembles pourront également servir ici puisque l'espace échantillonnal S joue le rôle d'**ensemble universel ou de référence** et les événements celui de sous-ensemble.

Exemple 1

Si un dé est lancé une fois, $S = \{1; 2; 3; 4; 5; 6\}$. $A = \{1 ; 2; 3\}$ est l'événement (obtenir 1; 2; ou 3). $B = \{5\}$ est l'événement «Obtenir un 5». (Noter qu'un événement peut contenir un ou plusieurs résultats élémentaires.)



Exemple 2

Un homme est accusé d'un vol de banque. Voici l'ensemble des événements possibles: A: {il a volé et les jurés l'ont trouvé coupable}. B: {il a volé et les jurés l'ont acquitté}. C: {il n'a pas volé et les jurés l'ont trouvé coupable}. D: {il n'a pas volé et les jurés l'ont acquitté}.

		Les jurés l'ont trouvé coupable/non coupable	
L'homme a volé la banque ?	Oui	A	B
	Non	C	D

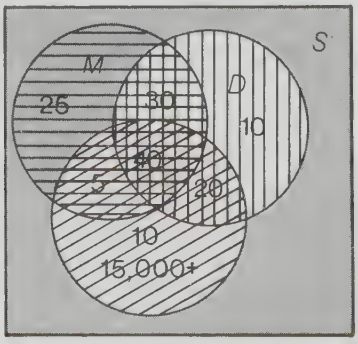
Exemple 3

Un club social de Montréal est formé de la façon suivante: il y a 200 membres dont 40 sont des diplômés d'université, mariés, gagnant plus de \$15 000 par année; 70 sont des diplômés mariés; 60 sont diplômés gagnant plus de \$15 000 par année; 100 sont des diplômés; 45 sont mariés gagnant plus \$15 000 par année; 100 sont mariés et 75 gagnent plus de \$15 000 par année.

(a) Combien de membres ne sont pas mariés, ne sont pas diplômés et gagnent moins de \$15 000 par année?

(b) Combien sont mariés mais ne sont pas diplômés?

Traçons un diagramme de Venn en posant S = {membres du club}, M = {membres mariés}, D ={-------- diplômés} 15 000 + = {gagnant plus de \$15 000}. Dans un carré S, traçons trois cercles qui se coupent représentant chacun de ces ensembles puis, débutons par la partie centrale. 40 membres sont dans l'intersection des trois cercles. 70 sont dans l'intersection des cercles M et D mais 40 y sont déjà inscrits il en reste donc 30 à inscrire dans l'autre portion. 60 sont dans l'intersection de D et 15 000 +, il en reste donc 20 à inscrire dans l'autre portion. En continuant ainsi pour les autres cas, on peut compléter le diagramme de Venn.



(a) $200 - (25 + 30 + 10 + 5 + 40 + 20 + 10) = 60$.
(b) $25 + 5 = 30$.

Exemple 4

Si un dé régulier est lancé, $S = \{1, 2, 3, 4, 5, 6\}$ il y a 6 résultats également possibles et des probalités égales doivent être données à chacun d'eux de sorte que leur somme donne 1, ainsi on obtient:

1	2	3	4	5	6	S
1/6	1/6	1/6	1/6	1/6	1/6	

On utilise souvent l'expression «au hasard» en parlant du choix d'un objet, d'un ensemble. Quelquefois, cette expression indique que chaque groupe possède une probabilité connue d'être choisi. Dans ce cours, elle sera utilisée dans un sens bien particulier: **un objet choisi au hasard implique que chaque membre du groupe aura la même probabilité d'être choisi.** «Une bille est choisie au hasard d'un sac contenant 4 billes rouges et 6 billes blanches.» Les billes ont la même grosseur et le même poids, elles ne diffèrent que par la couleur et elles ont été bien mélangées. Ce ne serait pas le cas si les 4 billes blanches étaient plus grosses que les rouges.

Exemple 5

Dans un échantillon de 100 000 Américains âgés de 40 ans, une étude montre que 99 382 seront vivants dans un an et 618 décéderont durant l'année. Comment peut-on utiliser cette information pour déterminer la probabilité des événements « vivant; mort» pour un Américain âgé de 40 ans choisi au hasard?

S = {Vivant; Mort}. On voit que la somme des probabilités doit égaler 1. Comme il y a 99 382 vivants sur un total de 100 000, on aura p (Vivant) $= \dfrac{99\,382}{100\,000} = 0{,}99382$ et p (Mort) $= \dfrac{618}{100\,000} = 0{,}00618$.

Vivant	Mort	S
0,993 82	0,006 18	

Exemple 6

Au palais de justice de St-Jérôme, il y eut 150 comparutions au mois d'octobre. 90 ont reçu une amende et ont été mis en probation et 60 reçurent une peine d'emprisonnement d'un an ou moins. Si un des inculpés est choisi au hasard, quelle est la probabilité a) qu'il reçut une amende b) qu'il reçut une peine d'emprisonnement d'un an ou moins c) qu'il fut condamné à plus d'un an?

Soient A, B et C les événements {amende} {prison 1 an ou moins} et {prison plus d'un an}.

A	B	C	S
$\frac{90}{150}$	$\frac{60}{150}$	$\frac{0}{150}$	

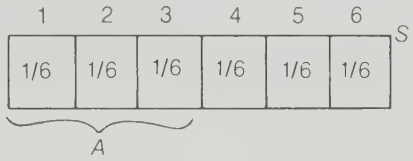
(a) $p(A) = 90/150 = 0{,}60$ (b) $p(B) = 60/150 = 0{,}40$

(c) $p(C) = 0/150 = 0{,}00$

Pour trouver la probabilité d'un événement A, $p(A)$, il faut additionner tous les nombres assignés aux éléments de A. Si une expérience comprend S éléments alors $p(S) = 1$.

Exemple 7

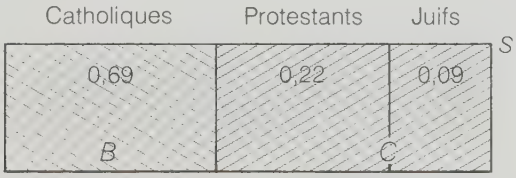
Si un dé est lancé, quelle est la probabilité d'obtenir 1, 2 ou 3? $S = \{1, 2, 3, 4, 5, 6\}$ $A = \{1, 2, 3\}$. Chaque élément de S a une probabilité de 1/6. La probabilité d'obtenir 1, 2 ou 3 sera $p(A) = 1/6 + 1/6 + 1/6 = 1/2$.



Exemple 8

Dans une petite localité, il y a 4 112 catholiques, 1 332 protestants et 556 juifs. Si une personne est choisie au hasard, quelle est la probabilité qu'elle soit a) Catholique b) Protestante ou Juive?

Soit B = {Catholique} C = {Protestant ou Juif}



(a) $p(B) = \dfrac{4\,112}{4\,112 + 1\,332 + 556} = \dfrac{4\,112}{6\,000} = 0{,}69$

(b) $p(C) = \dfrac{1\,332 + 556}{6\,000} = \dfrac{1\,888}{6\,000} = 0{,}31$

Notation: Dans les deux derniers exemples la notation $p(A)$ a été utilisée pour indiquer que A est un événement de l'espace échantillonnal S. Dans les cas des exemples 7 et 8 (a) on aurait dû écrire $p(\{1; 2; 3\}) = 0{,}50$ et $p(\{\text{catholique}\}) = 0{,}69$; en pratique, les accolades ne sont pas utilisées et on écrit $p(1; 2; 3) = 0{,}50$ et $p(\text{catholiques}) = 0{,}69$.

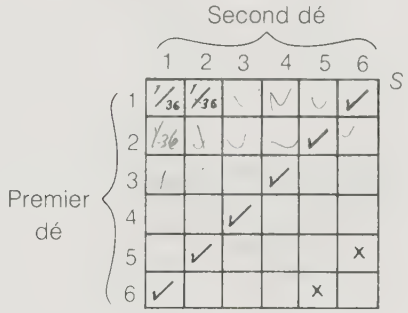
Exemple 9

Si deux dés sont lancés, quelle est la probabilité d'obtenir (a) un total de 11 avec les deux dés (b) un total de 7?

(a) Ici il est plus difficile de donner la liste des éléments de l'espace échantillonnal. Chaque élément de S est un couple comme (4;3) où 4 indique le résultat du premier dé et 3 le résultat du second. Donc $S = \{(1; 1)\,(1; 2) \ldots .\,(1; 6)$

(2; 1) . . . . (2; 6) . . . . (6; 6)}. Il est facile de voir avec un diagramme de Venn qu'il y a 36 éléments dans S; la probabilité de chacun est 1/36. L'événement «obtenir un total de 11» comprend les éléments indiqués par une croix dans le diagramme. Il y en a deux donc $p\ (11) = p\ [(6; 5)\ (5; 6)] = 1/36 + 1/36 = 1/18$.

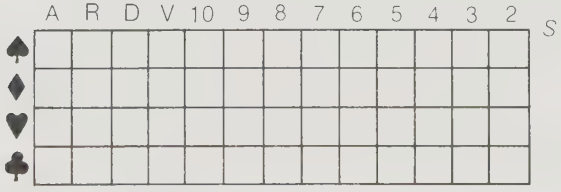
(b) Les éléments de l'événement «obtenir un total de 7» sont indiqués par un crochet $p\ (7) = p\ [(6; 1)\ (5; 2)\ (4; 3)\ (3; 4)\ (2; 5)\ (1; 6)] = 6 * 1/36 = 1/6$.



Exemple 10

Une carte est prise au hasard d'un jeu régulier de 52 cartes. Quelle est la probabilité d'obtenir (a) un roi (b) un coeur (c) un pique ou un coeur? Quelle est la probabilité de chaque élément de S?

(a) $p(R) = ?$
(b) $p(\heartsuit) = ?$
(c) $p(\spadesuit \text{ ou } \heartsuit) = ?$



Réponses

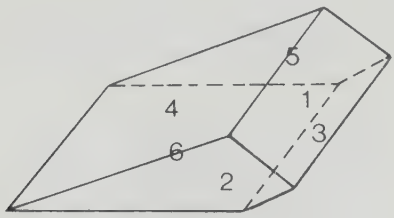
Chaque élément de S a une probabilité de 1/52;

(a) $p\ (R) = 4/52 = 1/13$; (b) $p\ (\heartsuit) = 13/52 = 1/4$; $p\ (\spadesuit \text{ou} \heartsuit) = 26/52 = 1/2$.

Exemple 11

Un drôle de dé est fabriqué de la façon suivante: il a 6 faces et la somme des nombres de chaque paire de faces opposées est 7. Dans la figure ci-dessous, la face 1 est en haut à droite, la face 6 sur la table, la face 2 en avant verticalement, la face 5 en arrière, la face 4 en haut à gauche et la face 3 inclinée à droite. Dû à cette construction, ce dé ne peut reposer sur la face 3 (il retombe immédiate-

ment sur la face 6) donc on ne peut obtenir la face 4 et il ne peut reposer sur la face 1 (il retombe sur le côté 4) on ne peut donc obtenir 6.



Le dé est lancé 1 000 fois. Voici les résultats obtenus:

Face	Nombre de fois
1	470
2	110
3	310
4	0
5	110
6	0
	1000

Quelle est la probabilité que si le dé est lancé une fois (a) on obtienne un nombre pair (b) La face 1 ou la face 3 (c) la face 2 ou la face 5.

$S = \{1; 2; 3; 4; 5; 6\}$. Les probabilités des éléments de l'espace échantillonnal peuvent être déterminées en se basant sur l'expérience précédente. 4 et 6 sont des évènements impossibles donc $p(4) = p(6) = 0$. 2 est apparu 110 fois en 1 000 tentatives donc $p(2) = \frac{110}{1000} = 0{,}110$ et ainsi de suite (il faut vérifier si $p(1) + p(2) + \ldots + p(6) = 1$).

La solution devient alors simple:

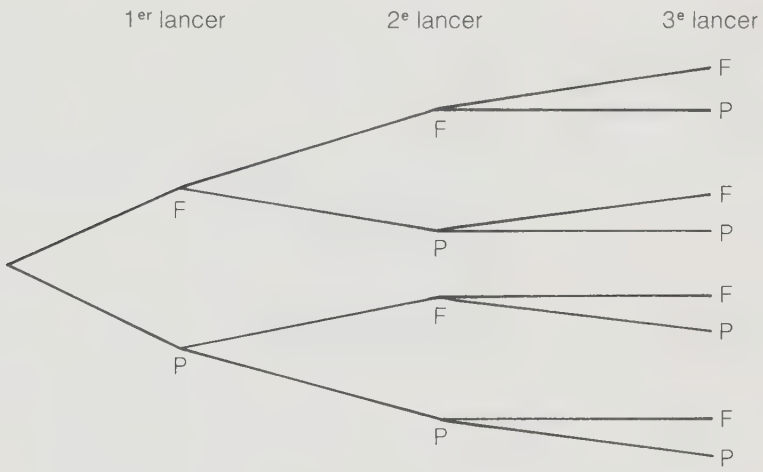
(a) p (2; 4 ou 6) = 0,11 + 0 + 0 = 0,11
(b) p (1 ou 3) = 0,47 + 0,31 = 0,78
(c) p (2 ou 5) = 0,11 + 0,11 = 0,22

1	2	3	4	5	6
0,47	0,11	0,31	0	0,11	0

5.4 ÉVÉNEMENTS MUTUELLEMENT EXCLUSIFS

Souvent, il sera plus facile de déterminer la probabilité d'un événement en trouvant la probabilité des autres événements. Nous avons déjà utilisé ce principe en additionnant les probabilités des éléments d'un événement. Si un dé est lancé, nous avons trouvé (exemple 7 précédent) que $p(1, 2 \text{ ou } 3) = p(1) + p(2) + p(3) = 1/6 + 1/6 + 1/6 = 1/2$. Voici une

nouvelle situation que nous voulons résoudre: quelle est la probabilité d'obtenir deux faces quand on lance une pièce de monnaie trois fois? Pour représenter cette situation, il nous faudrait un diagramme de Venn à trois dimensions. Heureusement, on peut utiliser un arbre pour illustrer les différentes possibilités.



On peut abréger «face à chacun des trois lancers» par FFF, face aux deux premiers et pile au dernier par FFP et ainsi de suite. On peut maintenant écrire les éléments de S: FFF; FFP; FPF; FPP; PFF; PFP; PPF; PPP. Si la pièce est lancée sept fois, cette méthode devient vite laborieuse, il nous faudra trouver quelques racourcis.

> ***Définition:*** Deux événements A et B sont **mutuellement exclusifs** s'ils n'ont pas d'éléments en commun. Si A et B ont une intersection vide dans un diagramme de Venn, ceci implique que A et B ne peuvent se produire en même temps.

Exemple 1

Un dé est lancé. $S = \{1, 2, 3, 4, 5, 6\}$.

Posons A: obtenir un nombre impair; $A = \{1; 3; 5\}$

B: obtenir les nombres 1, 2 ou 3; $B = \{1; 2; 3\}$

C: obtenir un 2; $C = \{2\}$

Les événements A et B ont les éléments 1 et 3 en commun donc A et B ne sont pas mutuellement exclusifs.

Il est impossible d'obtenir un nombre impair et un 2 si un dé est lancé, donc A et C n'ont pas d'élément en commun et sont mutuellement exclusifs. B et C ont l'élément 2 en commun, donc ne sont pas mutuellement exclusifs.

Exemple 2

Dans une élection partielle d'un comté où il y a 180 000 votants dont 100 000 citadins et 80 000 ruraux, on sait que, si la température est mauvaise,

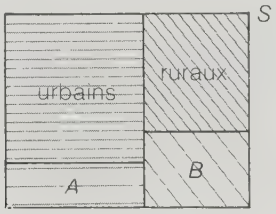
10% des citadins et 30% des ruraux ne voteront pas. Quelle est la probabilité qu'un des votants choisi au hasard reste à la maison si la température est mauvaise?

A: {Votants citadins qui resteront à la maison si la température est mauvaise}

B: {Votants ruraux qui resteront à la maison si la température est mauvaise}

A et *B* sont mutuellement exclusifs.

Si *C*: {Votants citadins} alors *A* et *C* ne sont pas mutuellement exclusifs mais *B* et *C* le sont.



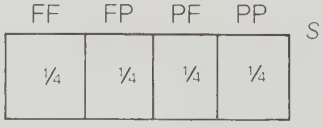
Si *A* et *B* sont mutuellement exclusifs alors p (*A* ou *B*) = p (*A*) + p (*B*).

Note: En général, ***A ou B*** , veut dire que *A* se produira ou *B* se produira ou les deux se produiront en même temps. Si *A* et *B* sont mutuellement exclusifs alors ***A ou B*** veut dire *A* se produira ou *B* se produira mais pas les deux en même temps.

Exemple 3

Si une pièce de monnaie est lancée deux fois, quelle est la probabilité d'obtenir une «face» et un «pile»?

S = {FP; PF; PP; FF}. (Se rappeler que FP veut dire obtenir F au premier jet et P au second jet.) Si la pièce est régulière, chaque élément de *S* est également probable donc aura une probabilité de 1/4. Soit *A* = {FP} et *B* = {PF} alors *A* et *B* sont mutuellement exclusifs.



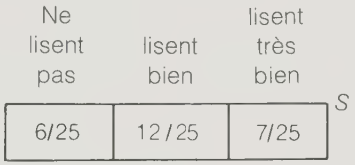
p (obtenir une face ou un pile) = p (FP ou PF) = p (A ou B) = p (FP) + p (PF) = 1/4 + 1/4 = 1/2.

Exemple 4

Si un enfant est choisi au hasard dans une classe de 25 enfants dont 6 lisent moins bien que la moyenne, 12 lisent moyennement et 7 lisent au-dessus de la moyenne, quelle est la probabilité qu'il lise moyennement ou moins.

S = {ne savent pas lire, lisent bien, lisent très bien}

p (que l'enfant choisi lise moyennement ou moins) $= p$ (lit moins bien que la moyenne) $+ p$ (lit moyennement) $= 6/25 + 12/25 = 18/25 = 0{,}72$



Si A, B et C sont mutuellement exclusifs, A n'a pas d'élément en commun avec B ou C, B n'a pas d'élément en commun avec A ou C. Donc $p\ (A$ ou B ou $C) = p\ (A$ ou $[B$ ou $C]) = p\ (A) + p\ (B$ ou $C) = p\ (A) + p\ (B) + p\ (C)$. De la même manière, les probabilités de plusieurs événements mutuellement exclusifs peuvent être additionnées.

Exemple 5

Si un résultat est choisi au hasard de la distribution suivante, quelle est la probabilité que celui-ci soit a) 4 ou moins b) 14 ou moins? Nous pouvons écrire a) $p\ (X \leqslant 4) = \frac{3}{25}$ b) $p\ (X \leqslant 14) = \frac{14}{25}$

X	f
0 – 4	3
5 – 9	4
10 – 14	7
15 – 19	2
20 – 24	9
	25

5.5 Axiomes de probabilités

Tous les résultats de la théorie des probabilités découlent des axiomes ou énoncés suivants:

1) $p\ (S) = 1$
2) $0 \leqslant p\ (A) \leqslant 1$ «$p(A)$ est compris entre 0 et 1 inclusivement»
3) Si A et B sont mutuellement exclusifs alors $p\ (A$ ou $B) = p\ (A) + p\ (B)$

Il faut se rappeler que si un événement A se produit toujours, A: est un événement certain et $p\ (A) = 1$. De même si A est un événement impossible alors $p\ (A) = 0$. Dans un problème, si on trouve une probabilité comme 2,30 ou $-0{,}5$, il faut revoir et vérifier ses calculs, car le 2e axiome affirme qu'une probabilité est toujours comprise entre 0 et 1 inclusivement.

Si A est un événement d'un espace échantillonnal S, appelons $\overline{A}$ l'événement «A ne se produit pas ». Il découle des axiomes précédents que: A et $\overline{A}$ sont des événements mutuellement exclusifs et $A \cup \overline{A} = S$, $A \cap \overline{A} = \emptyset$.

Donc $p\ (S) = 1$, $p\ (A \text{ ou } \overline{A}) = 1$(axiome 1), $p\ (A) + p\ (\overline{A}) = 1$(axiome 3)

$$p\ (A) = 1 - p\ (\overline{A})$$

Si la probabilité qu'il pleuve aujourd'hui est de 0,40, alors la probabilité qu'il ne pleuve pas est de $(1 - 0{,}40) = 0{,}60$.

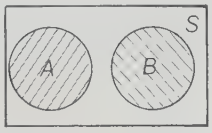
Exemple 1

Si deux dés sont lancés, quelle est la probabilité que la somme des résultats soit 4 ou plus (voir diagramme de Venn de l'exemple 9, section 5.3).

Soit X: la somme $p\ (X \geqslant 4) = ?$; $p\ (X \geqslant 4) = 1 - p\ (X < 4) = 1 - p\ (2 \text{ ou } 3) = 1 - [p\ (2) + p\ (3)] = 1 - (1/36 + 2/36) = 11/12$. Noter que les événements «la somme est 2» et «la somme est 3» sont mutuellement exclusifs, donc $p\ (2 \text{ ou } 3) = p\ (2) + p\ (3)$.

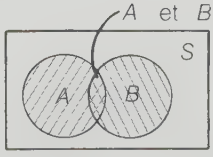
5.6 p (A OU B) SI A ET B NE SONT PAS MUTUELLEMENT EXCLUSIFS

Si A et B sont mutuellement exclusifs alors $p\ (A \text{ ou } B) = p\ (A) + p\ (B)$.



$p(A \text{ ou } B) = p(A) + p(B)$

Que se passe-t-il si A et B ont une partie commune tel que montré dans le diagramme de Venn suivant? La probabilité d'un événement est la somme des probabilités de chaque partie de l'événement.



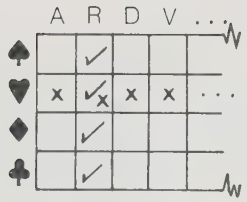
$p(A \text{ ou } B) = ?$

Si on additionne $p\ (A)$ et $p\ (B)$ quand A et B se recoupent, les éléments de la région commune ont été additionnés deux fois, on doit donc soustraire une fois ces éléments.

$$p(A \text{ ou } B) = p(A) + p(B) - p(A \text{ et } B)$$

Exemple 1

Si une carte est choisie au hasard d'un jeu de cartes, quelle est la probabilité que l'on pige un roi ou un coeur?



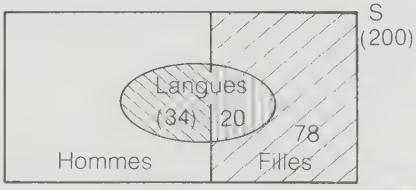
$p(R) = 4/52$; $p(\heartsuit) = 13/52$.

$p(R$ et $\heartsuit)$ = La probabilité que le roi de coeur soit tiré
$= 1/52$.

$p(R$ ou $\heartsuit) = p(R) + p(\heartsuit) - p(R$ et $\heartsuit)$
$= 4/52 + 13/52 - 1/52 = 16/52$
$= 4/13$.

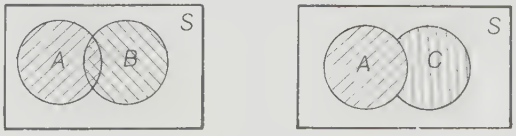
Exemple 2

Sur 200 finissants dans un collège il y a 98 filles. 34 hommes ont une concentration langues ainsi que 20 filles. Si un étudiant est choisi au hasard dans le groupe de finissants, quelle est la probabilité qu'il ait une concentration langue ou qu'il soit une fille ou les deux?



p (langues ou fille) $= p$ (langues) $+ p$ (fille) $- p$ (langues et fille)
$= 54/200 + 98/200 - 20/200 = 132/200$
$= 0{,}66$

Il est à noter qu'on peut résoudre certains problèmes sans l'utilisation de cette dernière formule. Chaque élément qui est dans A ou B est aussi dans A ou C si C est l'ensemble des éléments qui sont dans B, mais non dans A. Ainsi A et C sont adjacents dans le diagramme de Venn et sont mutuellement exclusifs (leur intersection est vide). p (A ou B) $= p$ (A ou C) $= p$ (A) $+ p$ (C).

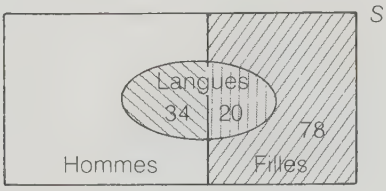


Exemple 3

p (R ou $\heartsuit$) $= p$ (R ou [$\heartsuit$ qui n'est pas un R])
$= p$ (R) $+ p$ ($\heartsuit$ qui n'est pas un R)
$= 4/52 + 12/52 = 16/52$

Exemple 4

p (langues ou fille) $= p$ [Langues ou (filles sans concentration langues)]
$= p$ (langues) $+ p$ (filles sans concentration langues)
$= 54/200 + (98 - 20)/200 = 0{,}27 + 0{,}39$
$= 0{,}66$



5.7 Événements indépendants p (A et B)

Si une pièce de monnaie est lancée deux fois, le résultat du deuxième jet ne sera pas affecté par le résultat du premier. Si deux dés sont lancés, le résultat du premier dé n'est pas influencé par le résultat du second. Supposons qu'un sac contient 6 boules blanches et 4 boules rouges et que l'on pige au hasard une boule, on note sa couleur puis on la replace dans le sac. Une seconde boule est pigée. Quelle est la probabilité que celle-ci soit rouge? La réponse (0,4) ne dépend pas du premier choix puisque la boule fut replacée dans le sac; ainsi les deux événements sont indépendants puisque la réalisation de l'un n'affecte en rien la probabilité de réalisation de l'autre.

Si toutefois, la première boule **n'est pas** remise dans le sac avant le second tirage, quelle est la probabilité que la seconde boule soit rouge? Si la première boule pigée est rouge, il ne reste plus que 9 boules dans le sac dont 3 sont rouges et p (la seconde est rouge) = 3/9 mais si la première est blanche alors 4 des 9 boules du sac sont rouges et p (la seconde est rouge) = 4/9. Le résultat du second tirage dépend du résultat du premier; on dit alors que les deux événements sont **dépendants.**

> A et B sont **indépendants** si la réalisation de A n'affecte pas la probabilité de réalisation de B et vice-versa.
>
> A et B sont **dépendants** si la réalisation de l'un affecte la probabilité de réalisation de l'autre.

Exemple 1

Deux candidats A et B ont chacun une liste de 10 hommes et 8 femmes. Une personne doit être choisie au hasard de chacune des listes pour représenter les candidats lors d'un banquet. Les événements C = {le représentant de A est une femme} et D = { le représentant de B est un homme} sont indépendants. En effet le choix d'une femme comme représentant de A n'affecte pas le choix du représentant de B.

Exemple 2

Deux cartes sont pigées au hasard d'un jeu de 52 cartes et ce sans remise. Les événements A = {la première est un coeur} et B = {la seconde est un coeur} sont dépendants. Si la première carte est un coeur, il reste 51 cartes dont 12 coeurs et la probabilité que la seconde carte soit un coeur est de 12/51; si la première carte n'est pas un coeur il y a 13 coeurs dans les 51 cartes qui restent pour le second choix et $p\ (B) = 13/51$.

Notation: le symbole $p\ (B/A)$ veut dire «la probabilité de l'événement B étant donné que l'événement A s'est déjà produit».

Dans l'exemple ci-dessus, on aurait pu écrire p (un homme représente le candidat B) $= p\ (D/C) = p$ (un homme représente le candidat B étant donné qu'une femme représente le candidat A) = 10/18. Il est à noter que dans ce cas $p\ (D/C) = p\ (D)$.

Dans l'exemple 2 ci-dessus p (B/A) $= p$ (la deuxième carte soit un coeur étant donné que la première est un coeur) = 12/51.

Quand on a deux événements A et B, avant de calculer les probabilités, on doit se poser la question «A et B sont-ils indépendants» et y répondre.

Si A et B sont indépendants $p\ (A$ et $B) = p\ (A) * p\ (B)$

Si A et B sont dépendants $p\ (A$ et $B) = p\ (A) * p\ (B/A)$

Exemple 3

Se référant à l'exemple 1, chaque candidat A et B possède une liste contenant 10 hommes et 8 femmes. Quelle est la probabilité que le représentant de A soit une femme et celui de B soit un homme?

C = {représentant de A est une femme}, D = {représentant de B est un homme}, $p\ (C) = 8/18$, $p\ (D) = 10/18$, C et D sont indépendants donc $p\ (C$ et $D) = 8/18 * 10/18 = 20/81$.

Exemple 4

Deux cartes sont pigées sans remise d'un jeu de 52 cartes. Quelle est la probabilité que les deux cartes soient des coeurs?

A = {première carte est un coeur}, B = {deuxième carte est un coeur}, A et B sont dépendants $p\ (A$ et $B) = p\ (A)\ p\ (B/A) = 13/52 * 12/51 = 1/17$.

Exemple 5

Une carte est pigée d'un jeu de 52 cartes, puis remise dans le jeu avant qu'une seconde pige soit effectuée. Quelle sont les probabilités que les deux cartes soient des coeurs?

Les deux événements sont indépendants, la probabilité de chacun est 13/52 = 1/4 et p (2 coeurs) = 1/4 * 1/4 = 1/16. Une autre façon de poser la même

$\frac{13}{52} * \frac{13}{52}$

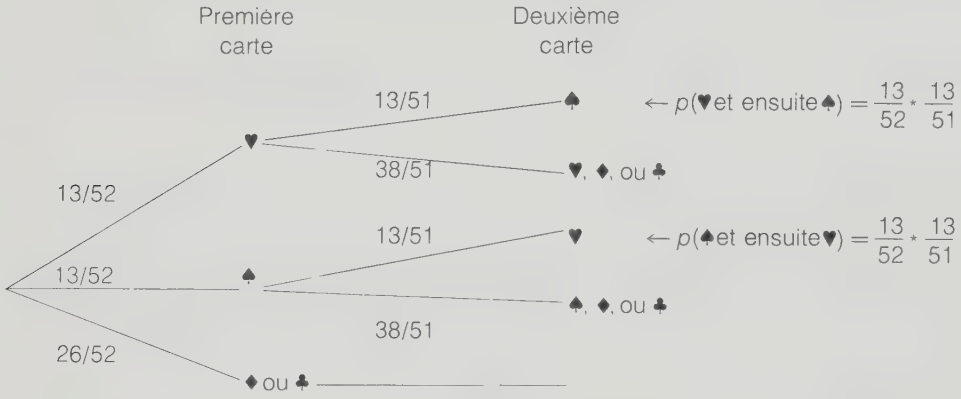
même question est: «Si deux cartes sont pigées avec remise d'un jeu de 52 cartes, quelle est la probabilité qu'elles soient toutes les deux des coeurs?»

Exemple 6 • •

Si deux cartes sont pigées d'un jeu de 52 cartes, quelle est la probabilité d'obtenir un coeur et un pique?

Piger deux cartes simultanément revient à en choisir une première sans remise puis en piger une seconde. Piger un coeur et un pique ne tient pas compte de l'ordre dans lequel elles sont pigées; on peut tout aussi bien avoir un coeur en premier puis un pique ou l'inverse.

Un diagramme en arbre illustre bien la situation.



$$p\ (1\ \text{coeur et 1 pique}) = \left(\frac{13}{52}\right)\left(\frac{13}{51}\right) + \left(\frac{13}{52}\right)\left(\frac{13}{51}\right) = 2\left(\frac{13}{204}\right) = \frac{13}{102}$$

Il est souvent plus facile de trouver une solution sans l'utilisation d'un arbre ou d'un diagramme de Venn. Essayons de comprendre le raisonnement suivant:

p (1 coeur et 1 pique) = p [(la première carte est un coeur et la seconde est un pique) ou (la première est un pique et la seconde un coeur)].
= p (la première est un coeur et la seconde est un pique) + p (la première est un pique et la seconde est un coeur)
= p (la première est un coeur) * p (la seconde est un pique étant donné que la première est un coeur) + p (la première est un pique) * p (la seconde est un coeur étant donné que la première est un pique).
= (13/52)(13/51) + (13/52)(13/51)
= 2(13/204)
= 13/102

Des événements mutuellement exclusifs permettent d'utiliser p (A ou B) = p (A) + p (B). Note: Il serait un peu long d'écrire le raisonnement de chaque

problème de la façon précédente. On pourra simplifier l'écriture de la façon suivante:

$p(1\heartsuit \text{ et } 1\spadesuit) = p(\heartsuit_1\spadesuit_2 \text{ ou } \spadesuit_1\heartsuit_2) = p(\heartsuit_1\spadesuit_2) + p(\spadesuit_1\heartsuit_2)$
$= p(\heartsuit_1) * p(\spadesuit_2/\heartsuit_1) + p(\spadesuit_1) * p(\heartsuit_2/\spadesuit_1) = 13/52 * 13/51 + 13/52 * 13/51 = 13/102.$

Exemple 7

Quelle est la probabilité d'obtenir exactement trois faces si on lance une pièce de monnaie 4 fois?

On utilisera la notation p (FFPF) pour indiquer la probabilité d'obtenir face au premier, second et quatrième lancers et pile au troisième. On peut donc écrire simplement:

$$\begin{aligned} p \text{ (3 F en 4 jets)} &= p \text{ (FFFP ou FFPF ou FPFF ou PFFF)} \\ &= p \text{ (FFFP)} + p \text{ (FFPF)} + p \text{ (FPFF)} + p \text{ (PFFF)} \\ &= p \text{ (F)} * p \text{ (F)} * p \text{ (F)} * p \text{ (P)} + p \text{ (F)} * p \text{ (F)} * \\ &\quad p \text{ (P)} * p \text{ (F)} + p \text{ (F)} * p \text{ (P)} p \text{ (F)} * p \text{ (F)} + \\ &\quad p \text{ (P)} * p \text{ (F)} * p \text{ (F)} * p \text{ (F)}. \\ &= (1/2)^4 + (1/2)^4 + (1/2)^4 + (1/2)^4 \\ &= 4(1/2)^4 \\ &= 4/16 = 1/4 \end{aligned}$$

Explications

(1) Ce sont 4 événements différents (FFPF) (PFFF) (FPFF) (FFFP).

(2) Donc mutuellement exclusifs, donc p (A ou B ou C ou D) $= p$ (A) $+ p$ (B) $+ p$ (C) $+ p$ (D).

(3) p (FFPF) est formé de 4 événements indépendants: le résultat du second jet est indépendant du premier, donc p (FFPF) $= p$ (F) $* p$ (F) $* p$ (P) $* p$ (F).

(4) p (F) $= 1/2$, p (P) $= 1/2$; la pièce est régulière.

Ne pas confondre des événements «mutuellement exclusifs» et «indépendants». Supposons que Georges et Marie jouent au tennis le samedi à 15 h. Soit G l'événement «Georges gagne son match» et M l'événement «Marie gagne son match». Si Georges et Marie ne jouent pas l'un contre l'autre, G et M sont des événements indépendants mais pas mutuellement exclusifs. S'ils jouent l'un contre l'autre, G et M sont des événements mutuellement exclusifs mais ne sont pas indépendants.

Si on veut calculer p (A ou B), il faut vérifier si A et B sont mutuellement exclusifs. Si oui, alors p (A ou B) $= p$ (A) $+ p$ (B).

Si on veut calculer p (A et B), il faut vérifier si A et B sont indépendants. Si oui, p (A et B) $= p$ (A) $* p$ (B).

5.8 VOCABULAIRE ET SYMBOLES

probabilité	choisir au hasard
événements également probables	événements mutuellement exclusifs
fréquence relative	p (A ou B)
espace échantillonnal	$\overline{A}$
sous-ensemble	p (A ou B)
diagramme de Venn	p (B/A)

Utiliser un diagramme de Venn ou un arbre si nécessaire.

1. Dans un groupe de 400 étudiants, 50 choisissent la philosophie, l'anglais et la zoologie, 160 la philosophie et l'anglais, 100 l'anglais et la zoologie, 60 la philosophie et la zoologie, 300 l'anglais, 140 la zoologie, 210 la philosophie.

(a) Combien d'étudiants n'ont pas choisi l'anglais ni la philosophie ni la zoologie?

(b) Combien ont choisi l'anglais ou la philosophie mais pas la zoologie?

(c) Combien ont choisi l'anglais mais n'ont pas choisi la philosophie ou la zoologie?

2. À un poste de vérification, 1% des autos ont des freins en mauvais état et des phares défectueux et causent trop de pollution; 19% causent trop de pollution; 14% ont des freins en mauvais état; 5% ont des phares défectueux et des freins en mauvais état; 3% ont des freins en mauvais état et causent trop de pollution; 3% ont des phares défectueux et causent trop de pollution; 16% ont des phares défectueux.

(a) Quel pourcentage des autos ont, soit de mauvais freins ou un mauvais éclairage (ou les deux)?

(b) Quel pourcentage ont des mauvais freins ou un mauvais éclairage (mais pas les deux)?

3. Un analyste en valeurs mobilières prévoit avec une probabilité de 0,45 que les actions de Canada Rubber augmenteront de 5 points; qu'elles resteront stables avec une probabilité de 0,24 et qu'elles diminueront de 5 points avec une probabilité de 0,31. Un autre analyste prévoit que les probabilités seront respectivement 0,47; 0,27; et 0,29; et un troisième analyste 0,40; 0,27; 0,35 respectivement. Commenter ces prévisions.

4. Assumant que les probabilités que les résultats d'un étudiant à un test de personnalité soient plus de 600, entre 500 et 599, entre 400 et 499, moins de 400 sont respectivement 0,16; 0,34; 0,34; 0,16. Quelle est la probabilité que son résultat soit (a) 400 ou plus (b) moins de 400 ou entre 500 et 599?

5. Sachant que 70% des actes criminels se produisent le soir et que 40% de ceux-ci sont violents, de plus 20% des actes criminels se produisant le jour sont violents. Si un acte criminel est choisi au hasard parmi les dossiers de la police d'où proviennent ces statistiques, quelle est la probabilité que celui-ci soit violent?

6. Si deux dés sont lancés, quelle est la probabilité que la somme des résultats soit (a) 10; (b) 9; (c) 9 ou plus; (d) moins que 9?

7. Une pièce de monnaie est lancée 1 000 fois et on obtient 440 faces. Si la pièce est lancée une autre fois, quelle est la probabilité d'obtenir (a) une face (b) un pile?

8. Si une carte est pigée d'un jeu de 52 cartes, quelle est la probabilité que l'on obtienne (a) un as (b) un 4 de carreau, un 7 de coeur (c) un 2 ou un 5 (d) un valet de pique (e) pas un 3 (f) un 10 ou un 9 ou un valet?

9. Une bille est choisie au hasard dans un sac contenant 5 billes rouges, 10 noires et 10 blanches. Quelle est la probabilité qu'elle soit (a) rouge (b) rouge ou blanche (c) pas rouge (d) ni rouge ni blanche?

10. Trouver la probabilité d'obtenir un 3 au moins une fois quand un dé est lancé deux fois.

11. Si une pièce de monnaie est lancée 3 fois, trouver la probabilité d'obtenir (a) aucune face (b) exactement une face (c) au moins une face.

12. Une famille est composée de 3 enfants. Trouver la probabilité que (a) ce soient 3 filles (b) un des enfants est un garçon (c) il y ait au moins un garçon. Assumer que la probabilité d'avoir un garçon est 1/2.

13. Un de vos amis vient d'obtenir un poste de vendeur. Vous lui offrez un pari de \$100 contre \$10 qu'il gagnera plus de \$10 000 en commission pendant sa première année. Que peut-on dire de sa probabilité suggestive de gagner \$10 000 cette année (a) s'il accepte le pari (b) s'il refuse le pari?

14. Lancer 2 pièces de monnaie 48 fois et enregistrer le nombre de fois où l'on n'a pas une face, une face, deux faces. Comparer ces résultats avec ceux obtenus en calculant les probabilités.

15. Un homme a 6 clefs dont une seule peut déverrouiller la porte de sa maison. S'il essaie les clefs une à la fois en choisissant au hasard et ce avec remise, quelle est la probabilité qu'il débarre sa porte (a) au premier essai (b) au troisième (c) au cinquième?

16. Pendant une campagne de charité, le Club Optimiste distribue des tracts à la moitié des résidents d'un quartier en les choisissant au hasard et le Club Lions lui, les distribue à tous les trois résidents selon leur adresse. Quelle est la probabilité qu'un résident de ce quartier choisi au hasard ait reçu (a) deux tracts (b) aucun tract (c) au moins un tract?

17. La compagnie Acme fabrique 10 000 réfrigérateurs chaque semaine. Chaque unité est soumise à trois vérifications *A, B, C* avant d'être livrée. Il y a 2% de rejet dès la première inspection *A*. 5% de ceux qui restent sont rejetés à la deuxième inspection *B* et finalement 1% du reste est rejeté lors de l'inspection *C*. Quelle est la probabilité qu'un réfrigérateur, choisi au hasard lors de la production, ne soit rejeté à aucune des inspections (le réfrigérateur est en bon état)?

18. 52 hommes et 48 femmes escaladent le Mont Washington. 14 personnes furent incommodées dont 6 femmes. Si un membre du groupe est choisi au hasard, quelle est la probabilité qu'il soit incommodé ou qu'il soit une femme?

●19. Un médecin trouve qu'un patient ayant le symptôme *A* a une probabilité de 0,6 d'avoir la tuberculose alors que la probabilité est de 0,5, s'il a le symptôme *B*. Quelle est la probabilité qu'un patient ayant les symptômes *A* ou *B* ait la tuberculose?

20. Une pièce de monnaie est lancée 3 fois. Quelle est la probabilité d'obtenir (a) aucune face (b) exactement 1 face (c) exactement 2 faces (d) exactement 3 faces (e) au moins une face?

21. (a) Un des nombres 2; 3; 4 est choisi au hasard. Quelle est la probabilité qu'il soit pair?

(b) On pige successivement 3 nombres avec remise. Quelle est la probabilité (I) aucun des nombres n'est pair (II) un seul est pair (III) deux sont pairs (IV) les 3 sont pairs?

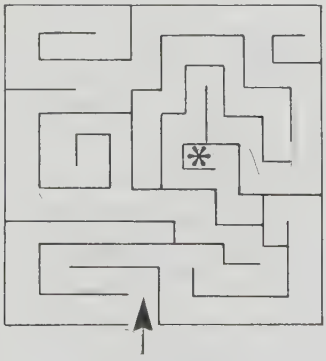
22. Une boîte de cigarettes en contient 10 avec filtre et 5 sans filtre. Si deux cigarettes sont prises au hasard, quelle est la probabilité que (a) les deux aient un filtre (b) au moins une ait un filtre?

23. -Quatre joueurs de tennis doivent jouer chacun un set contre chacun des trois autres. Un expert donne des probabilités qu'un joueur gagne au moins deux sets comme suit:

Joueur	A	B	C	D
Probabilité	0,7	0,8	0,2	0,3

Est-ce que la somme des probabilités peut être plus grande que 1? Expliquer.

24. Un enfant essaie de trouver son chemin dans le labyrinthe ci-dessous. Il choisit au hasard à chaque fois qu'une situation de choix se présente et ne refait jamais le même trajet, sauf quand il arrive dans un cul-de-sac. S'il ne fait pas d'erreur, il peut atteindre le centre en 40 secondes; chaque erreur cause un retard de 8 secondes. Quelle est la probabilité qu'il atteigne le centre en 70 secondes ou moins?



25. Un dé est lancé 4 fois. Quelle est la probabilité d'obtenir un 5 seulement au quatrième jet?

26. Dans un examen à choix multiples, il y a 10 questions avec 5 choix chacune. Si un étudiant connait la réponse à 5 questions et choisit au hasard la réponse des 5 autres, quelle est la probabilité qu'il obtienne un résultat de 70% ou plus?

27. M. et Mme Couture habitent à Montréal. La probabilité que M. Couture se rende à Québec le 31 décembre est de 0,2; la probabilité que Mme Couture s'y rende est de 0,1. La probabilité que M. Couture s'y rende sachant que Mme Couture s'y rend aussi est de 0,3. Quelle est la probabilité que (a) M. et Mme Couture se rendent à Québec le 31 décembre? (b) Au moins un des deux se rende à Québec? (c) Ni l'un ni l'autre ne se rendent à Québec le 31 décembre?

28. Si 3 fruits sont choisis au hasard d'un panier contenant 4 pommes, 5 poires et 1 pêche. Quelle est la probabilité que (a) la pêche soit prise? (b) Au moins deux pommes soient choisies? (c) Une pêche et deux pommes soient choisies?

29. Un sac contient 4 billes rouges et 6 blanches. Compléter le tableau suivant sachant que R: avec Remise et SR: Sans Remise.

	Piger 2 billes		Piger 3 billes	
	R	SR	R	SR
p (aucune blanche)				
p (exactement une est blanche)				
p (exactement 2 sont blanches)				
p (exactement 3 sont blanches)				

30. Un dé est lancé deux fois. Quelle est la probabilité que la somme des résultats soit (a) 11? (b) 11 ou plus? (c) 11 si le premier résultat est 5? (d) 11 si au moins un des deux résultats est 5?

31. Si une des bougies d'un moteur 6 cylindres est défectueuse, quelle est la probabilité que la bougie défectueuse soit découverte (a) si une bougie choisie au hasard est remplacée par une neuve? (b) si 2 bougies choisies au hasard sont remplacées par des neuves?

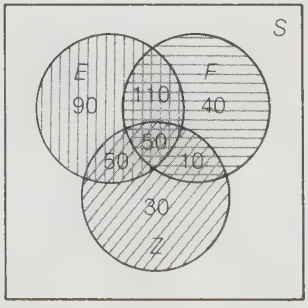
32. Si un projet de loi doit franchir trois lectures en chambre avant de devenir une loi et que les probabilités que ce projet de loi soit accepté à chacune des ces lectures sont respectivement 0,40; 0,30; 0,20. Quelle est la probabilité que ce projet de loi devienne une loi?

●33. Un jeu de Bingo simplifié est composé de boules numérotées de 0 à 9. Les boules sont choisies au hasard et les joueurs couvrent les numéros correspondants sur une carte. Le gagnant est le premier qui réussit à couvrir 5 numéros sur une même rangée ou en diagonale. S'il vous manque les numéros 3, 7, 9 pour compléter une rangée, quelle est la probabilité que vous y parveniez (a) dans les 3 prochains appels? (b) 4 appels exactement?

●34. Dans une boîte il y a deux pièces de monnaie, une est régulière «p (P) = p (F) = 1/2», mais l'autre possède deux faces «p (F) = 1 p (P) = 0». Une pièce est choisie au hasard dans la boîte et est lancée 6 fois. Les 5 premiers lancers donnent un résultat «F». Quelle est la probabilité d'obtenir un F au sixième lancer?

RÉPONSES

1. (a) 20. (b) 90 + 110 + 40 = 240. (c) 90.



3. Les prévisions du second et du troisième analyste sont impossibles, car la somme des probabilités ne donne pas 1.

5. Pour 1 000 actes criminels, 700 se produisent le soir et 280 sont violents, 300 se produisent le jour et 60 sont violents; 280 + 60 = 340 actes criminels violents

p (actes violents) $= \dfrac{340}{1000} = 0{,}34$

7. Si on utilise la définition de fréquence relative on obtient (a) $\dfrac{440}{1000} = 0{,}44$ (b) $1 - 0{,}44 = 0{,}56$. Si la pièce est régulière on obtient (a) 1/2 (b) 1/2. Au chapitre 7, nous verrons que si la pièce est régulière, la probabilité d'obtenir 440 faces en 1000 essais est très petite, mais plus grande que 0,00.

9. (a) 5/25 = 1/5. (b) 15/25 = 3/5. (c) 1 − 1/5 = 4/5 ou 10/25 + 10/25 = 20/25 = 4/5. (d) 10/25 = 2/5.

11. FFF; FFP; FPF; PFF; FPP; PFP; PPF; PPP. (a) 1/8. (b) 3/8. (c) 1 −1/8 = 7/8.

13. (a) Sa probabilité objective est au moins 10/11 (b) est moins de 10/11.

15. (a) 1/6 (b) 1/6 (c) 1/6.

17. Sur 10 000 unités, 9 800 passent la première inspection; de celles-ci, 9 310 passent l'inspection *B* et 9 217 l'inspection *C*. Donc p (bon état) est 0,92. Autrement p (passe *A*) p (passe *B* étant donné qu'il a passé *A*) p (passe *C* étant donné que *A* et *B* ont été passées) = (1 − 0,02) (1 − 0,05) (1 − 0,01) = 0,98 * 0,95 * 0,99 = 0,92. On n'a pas assez d'information pour donner une réponse. On cherche p (*T*/*A* ou *B*) et il nous manque p(*A*), p(*B*), p(*A* et *B*) et p(*T*/*A* et *B*).

21. (i) (a) p = 2/3; (b) $(1/3)^3 = 1/27$; (ii) $3(2/3)(1/3)^2 = 2/9$; (iii) $3(2/3)^2(1/3)^1 = 4/9$; (iv) $(2/3)^3 = 8/27$.

23. Oui; lire p (*A g B*) comme la probabilité «que le joueur *A* gagne son set contre *B*» alors p (*X g Y*) + p (*Y g X*) doit égaler 1 pour chaque couple de joueurs. La réponse est plus facile à comprendre si les joueurs *A* et *B* sont des professionnels et les joueurs *C* et *D* sont des débutants. Ainsi, les probabilités de gagner au moins 2 sets sont 1, 1, 0, 0, respectivement. Les probabilités suivantes attribuées à chacun des sets semblent peu réalistes, mais résultent des prévisions de l'expert.

	A	*B*	*C*	*D*
p (*A g* ------)	–	0,5	0,4	1
p (*B g* ------)	0,5	–	1	0,6
p (*C g* ------	0,6	0	–	0,3
p (*D g* ------)	0	0,4	0,7	–

25. p (pas 5 dans les 3 premiers jets et 5 au 4e) $= (5/6)^3 (1/6) = 0{,}1$.

27. (a) p (M et Mme) = p (M) p (Mme) = 0,1 * 0,3 = 0,03.
(b) p (M ou Mme) = p (M) + p (Mme) − p (M et Mme) = 0,2 + 0,1 − 0,03 = 0,27.
(c) p (pas M, pas Mme) = 1 − p (M ou Mme)
= 1 − 0,27 = 0,73.

29.

Nombre de blanches	Piger 2 billes		Piger 3 billes	
	R	*SR*	*R*	*SR*
0	$(0{,}40)^2 = 0{,}16$	$\frac{4}{10} * \frac{3}{9} = 0{,}13$	$(0{,}4)^3 = 0{,}06$	$\frac{4}{10} * \frac{3}{9} * \frac{2}{8} = 0{,}03$
1	$2 * 0{,}40 * 0{,}60 = 0{,}48$	$2 * \frac{4}{10} * \frac{6}{9} = 0{,}53$	$3(0{,}4)^2\,(0{,}6) = 0{,}29$	$3 * \frac{4}{10} * \frac{3}{9} * \frac{6}{8} = 0{,}30$
2	$(0{,}60)^2 = 0{,}36$	$\frac{6}{10} * \frac{5}{9} = 0{,}33$	$3(0{,}4)\,(0{,}6)^2 = 0{,}43$	$3 * \frac{4}{10} * \frac{6}{9} * \frac{5}{8} = 0{,}50$
3	0	0	$(0{,}6)^3 = 0{,}22$	$\frac{6}{10} * \frac{5}{9} * \frac{4}{8} = 0{,}17$

31. (a) $1/6 = 0{,}17$ (b) $2\,(1/6 * 5/5) = 2\,(5/6 * 1/5) = 1/3$

33. $0{,}3 * 0{,}2 * 0{,}1 = 0{,}006$. (b) Pour gagner en 4 appels, le dernier doit vous donner Bingo. Il y a ici une complication, si on obtient un 3 au premier appel, on peut obtenir un autre 3 au second ou au troisième appel et gagner en 4, mais si on obtient pas un 3, 7 ou 9 au premier appel, il ne peut y avoir de répétition dans les autres appels.

p (3; 7; 9 [peu importe l'ordre] aux appels 2; 3; 4) $= 0{,}7 * 0{,}3 * 0{,}2 * 0{,}1 = 0{,}0042$
p (3; 7; 9 [peu importe l'ordre] aux appels 1; 3; 4) $= 0{,}3 * 0{,}8 * 0{,}2 * 0{,}1 = 0{,}0048$
p (3; 7; 9 [peu importe l'ordre] aux appels 1; 2; 4) $= 0{,}3 * 0{,}2 * 0{,}9 * 0{,}1 = 0{,}0054$
$0{,}0144$

p (de gagner en 4 appels) $= 0{,}0144$

Chapitre 6

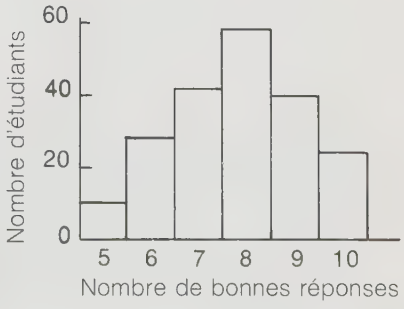
Distributions de probabilités (cas discret)

Ce chapitre est une introduction aux différentes distributions de probabilités. Il y a plusieurs types de distributions. Nous étudierons en particulier la distribution binômiale dans ce chapitre et la distribution normale dans le suivant. La distribution binômiale est très importante et est essentielle à la compréhension de plusieurs notions des chapitres suivants.

6.1 DISTRIBUTIONS DE PROBABILITÉS

Nous sommes familiers (depuis le chapitre 2) avec les distributions de fréquences et leur représentation sous forme d'histogramme, un exemple:

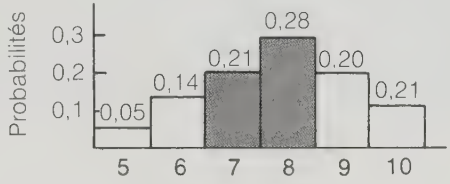
Nombre de bonnes réponses dans un test Vrai ou Faux	Nombre d'étudiants
5	10
6	28
7	42
8	56
9	40
10	24
	200



Si un étudiant est choisi au hasard, la probabilité qu'il réponde correctement à 5 questions est de 10/200 = 0,05. De la même manière, on peut trouver la probabilité que le nombre de bonnes réponses soit: 6; 7; 8; 9; ou 10. Les résultats sont compilés dans le tableau suivant:

Nombre de bonnes réponses	Probabilités
5	0,05
6	0,14
7	0,21
8	0,28
9	0,20
10	0,12
	1,00

Ce tableau représente une **distribution de probabilités.** Il est à noter que les événements dans la colonne des résultats sont mutuellement exclusifs. Un étudiant ne peut avoir en même temps 5 et 8 bonnes réponses dans le même test. De plus, la somme des probabilités est égale à 1. Cette distribution de probabilités est une **distribution de fréquences relatives:** p (5 bonnes réponses) = 0,05 puisque 10 étudiants sur 200 ont répondu correctement à 5 questions. Cette distribution particulière peut être représentée par un histogramme. L'aire totale sous l'histogramme est 1.



De plus, la probabilité d'obtenir 7 à 8 réponses correctes aux questions est: 0,21 + 0,28 = 0,49 qui peut être obtenu en additionnant les aires des rectangles correspondant à 7 et 8 bonnes réponses (aires ombragées de l'histogramme).

Une distribution de probabilités est un cas particulier d'une distribution de fréquences relatives et possède également une moyenne et un écart-type. Généralement, on considère une distribution de probabilités pour une population et non pour un échantillon, c'est pourquoi deux nouveaux symboles sont introduits: μ (mu) m minuscule grec et σ (sigma) s minuscule grec.

μ = moyenne de la population

σ = écart-type de la population

Les formules permettant de trouver μ et σ d'une distribution de probabilités sont:

$$\mu = \Sigma X p(X)$$

$$\sigma = \sqrt{\Sigma (X - \mu)^2 p(X)}$$

N.B.: On reconnaît dans ces formules la parenté certaine avec $\overline{X}$ et S d'une distribution de fréquences.

Exemple 1

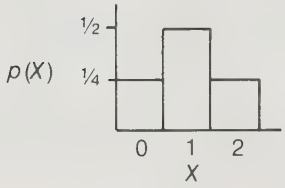
Une pièce de monnaie est lancée deux fois. Quelle est la probabilité d'obtenir (a) 0 face, (b) 1 face, (c) 2 faces? Trouver une distribution de probabilités, calculer sa moyenne et son écart-type puis, tracer son histogramme.

Soit X le nombre de faces.

X	$p(X)$	$Xp(X)$	$X-\mu$	$(X-\mu)^2$	$(X-\mu)^2p(X)$
0	1/4	0	−1	1	1/4
1	1/2	1/2	0	0	0
2	1/4	1/2	1	1	1/4
		1			1/2

$\mu = \Sigma Xp(X) = 1.$

$\sigma = \sqrt{\Sigma(X-\mu)^2p(X)} = \sqrt{1/2} = 0{,}7$



Exemple 2

Après un grand nombre d'expériences, la distribution ci-dessous donne la probabilité qu'un rat franchisse le labyrinthe en moins de 60 secondes après X tentatives. (a) Trouver la moyenne et l'écart-type de cette distribution. (b) Quelle est la probabilité qu'un rat prenne 3 essais ou moins pour franchir le labyrinthe dans le temps requis? (c) Tracer un histogramme.

X	$p(X)$	$Xp(X)$	$X-\mu$	$(X-\mu)^2$	$(X-\mu)^2p(X)$
1	0,20	0,20	− 1,5	2,25	0,450
2	0,40	0,80	− 0,5	0,25	0,100
3	0,20	0,60	0,5	0,25	0,050
4	0,10	0,40	1,5	2,25	0,225
5	0,10	0,50	2,5	6,25	0,625
		2,50			1,450

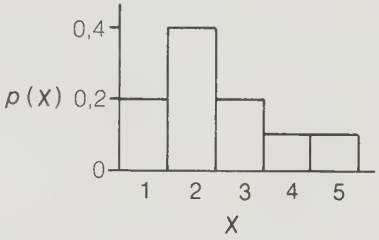
(a) $\mu = \Sigma Xp(X) = 2{,}5$

$\sigma^2 = \Sigma(X-\mu)^2p(X) = 1{,}450$

$\sigma = \sqrt{1{,}450} = 1{,}20$

(b) $p(X = 1, 2, \text{ou } 3)$

$= 0{,}20 + 0{,}40 + 0{,}20 = 0{,}80$



Si l'on a une distribution de fréquences, on peut obtenir une distribution de probabilités en divisant chaque nombre de la colonne f par le nombre total de résultats (effectifs). Si l'on a une distribution de probabilités, on peut immédiatement trouver le pourcentage des résultats dans chacune des classes (en multipliant par 100) et trouver le nombre de résultats dans chaque classe (la fréquence) si le nombre de résultats est connu.

classes	f	$p(X)$
10–19	6	6/50 = 0,12
20–29	26	26/50 = 0,52
30–39	18	18/50 = 0,36
	50	1,00

classes	$p(Y)$	%	si $n = 20$: f
1–3	0,3	30	6
4–6	0,6	60	12
7–9	0,1	10	2
	1,00	100	20

Il faut se rappeler qu'une distribution de probabilités est un cas particulier d'une distribution de fréquences, que la somme des probabilités est 1 et que la probabilité qu'un résultat soit compris entre a et b est égale à l'aire sous l'histogramme entre a et b.

6.2 PRÉ-TEST SUR LA PERMUTATION ET LES COMBINAISONS.

Avant de poursuivre l'étude des distributions de probabilités, nous ferons une étude rapide des permutations et combinaisons. Si vous pouvez répondre correctement aux questions suivantes, passez immédiatement à la section 6.6.

${}_nP_x$ = nombre de permutations de n objets pris X à la fois.

${}_nC_x$ = nombre de combinaisons de n objets pris X à la fois.

1. De combien de manières peut-on former un comité de 2 membres pris dans une association de 10 membres?
2. De combien de manières peut-on choisir un président et un vice-président dans une association de 10 membres?
3. $5! = ?$
4. Combien de mots de 3 lettres peut-on former avec les lettres a, e, i, o, u, si une lettre ne peut être répétée dans un même mot?

Réponses: 1) ${}_{10}C_2 = 45$; 2) ${}_{10}P_2 = 90$; 3) 120; 4) ${}_5P_3 = 60$.

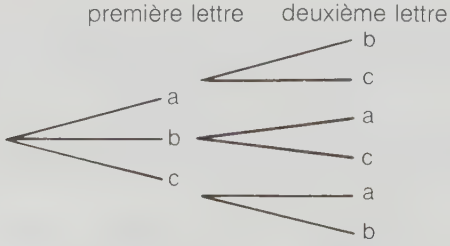
6.3 PERMUTATIONS

Une **permutation** est un arrangement ordonné sans répétition des objets d'un ensemble. Ainsi en se demandant: «Combien de permutations de 2 lettres peut-on obtenir avec les lettres $\{a, b, c\}$?» implique que ab est différent de ba; les deux mêmes lettres sont présentes, mais l'ordre d'apparition est différent.

Avant d'obtenir une formule pour calculer les permutations, examinons la situation en construisant un diagramme en arbre.

Exemple 1

Combien de permutations de 2 lettres peut-on former avec les lettres $\{a, b, c\}$?



À partir de cet arbre, on peut écrire les six permutations différentes: $ab; ac; ba; bc; ca; cb$. Il est possible de trouver ce résultat sans construire un arbre, mais celui-ci nous aidera à mieux comprendre. En regardant notre arbre de gauche à droite, on constate qu'il y a 3 branches, représentant les 3 choix possibles pour la 1re lettre (a; b ou c) puis, chacune de

ces branches donne naissance à deux autres branches, les deux choix qui restent pour la 2e lettre (si la 1re lettre est *b*, la 2e peut être *a* ou *c*). Il y a donc 3 * 2 permutations (3 choix pour la 1re lettre * 2 choix pour la 2e). La notation pour le nombre de permutations de 3 objets pris 2 à la fois est ${}_3P_2$ et l'on sait que ${}_3P_2 = 3 * 2 = 6$. Différentes situations peuvent faire appel au même principe. Il faut se rappeler que les mots clefs sont: l'ordre et l'absence de répétition.

Les situations suivantes ont toutes pour résultat ${}_3P_2 = 3 * 2 = 6$.

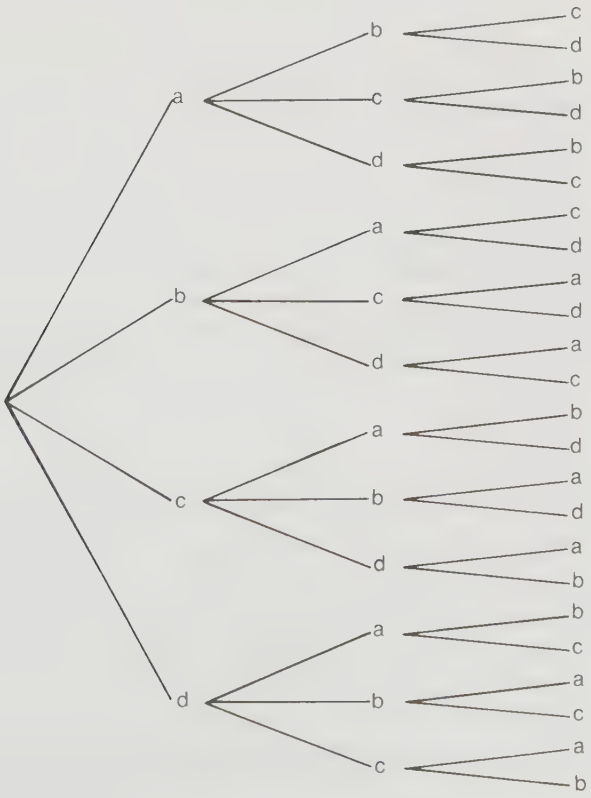
1. Combien de nombres de 2 chiffres peut-on former à partir des chiffres 1; 3; 4 si aucun chiffre n'est répété?
2. De combien de manières différentes peut-on choisir un président et un secrétaire dans un comité de 3 membres?
3. Un étudiant doit choisir un cours optionnel en 1re année et de même en 2e année et doit choisir parmi les cours suivants {chimie, physique, zoologie}. Combien de possibilités a-t-il?

Exemple 2

Combien de permutations peut-on obtenir de 4 objets si on en prend 3 à la fois? Supposons que les objets sont *a, b, c, d.* On peut écrire la liste suivante:

abc	*acb*	*bac*	*bca*	*cab*	*cba*	*dab*	*dba*	
abd	*adb*	*bad*	*bda*	*cad*	*cda*	*dac*	*dca*	
acd	*adc*	*bcd*	*bdc*	*cbd*	*cdb*	*dbc*	*dcb*	réponse: 24

Ou l'on peut construire l'arbre suivant et compter le nombre de branches.



Le résultat est 24. À gauche, il y a 4 branches (4 choix possibles pour la 1re lettre); chacune d'elles donne naissance à 3 autres branches (3 choix possibles pour la 2e lettre); puis finalement, (2 choix pour la 3e lettre); ainsi ${}_4P_3 = 4 * 3 * 2 = 24$. On peut également représenter le raisonnement de la façon suivante:

	1re lettre		2e lettre		3e lettre
Nombre de choix	4	*	3	*	2

Exemple 3

${}_{10}P_4 = ?$ Donner la liste des permutations. Construire un arbre peut s'avérer dans ce cas assez laborieux. On peut choisir 4 objets parmi 10 sans répétition. Représentons notre raisonnement comme précédemment:

	1re lettre		2e lettre		3e lettre		4e lettre
Nombre de choix	10	*	9	*	8	*	7

On a 10 choix pour emplir la 1re case. Il en reste 9 pour la 2e et ainsi de suite.

Nous venons de voir que: ${}_3P_2 = 3 * 2$
${}_4P_3 = 4 * 3 * 2$
alors ${}_{10}P_4 = 10 * 9 * 8 * 7$.

Peut-on trouver ${}_{12}P_5 = ?$ Si n et X sont des entiers positifs et que $X \leqslant n$, ${}_nP_X = ?$ On constate que ${}_nP_X$ aura X facteurs; le premier étant n, le second $n - 1$, le troisième $n - 2$ et le X^e sera $[n - (X - 1)]$. Ainsi ${}_{10}P_8 = 10 * 9 * 8 * 7 * 6 * 5 * 4 * 3$; ${}_3P_1 = 3$; ${}_5P_2 = 5 * 4$; ${}_{12}P_5 = 12 * 11 * 10 * 9 * 8$.

6.4 LA NOTATION FACTORIELLE

Il est souvent utile d'abréger des produits du type $3 * 2 * 1$ ou $5 * 4 * 3 * 2 * 1$. De tels produits sont représentés par un point d'exclamation placé après le plus grand chiffre de la séquence. Donc $3 ! = 3 * 2 * 1$; $5 ! = 5 * 4 * 3 * 2 * 1$.

Définition: Si n est un entier positif, $n ! = n\,(n - 1)\,(n - 2) \ldots 3 * 2 * 1$

3 ! se lit 3 factoriel, n ! se lit n factoriel.

Exemple 1

$$ {}_5P_2 = 5 * 4 = \frac{5 * 4 * 3 * 2 * 1}{3 * 2 * 1} = \frac{5!}{3!} = \frac{5!}{(5-2)!} $$

Exemple 2

$$_{10}P_4 = 10 * 9 * 8 * 7 = \frac{10 * 9 * 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1}{6 * 5 * 4 * 3 * 2 * 1} = \frac{10!}{6!} = \frac{10!}{(10-4)!}$$

Exemple 3

$$_{27}P_{12} = \frac{27!}{(27-12)!} = \frac{27!}{15!}$$

Exemple 4

$$_nP_k = ?$$

Par analogie avec les trois premiers exemples, on peut écrire directement $_nP_k = \frac{n!}{(n-k)!}$. Examinons maintenant $_nP_n$ (les permutations de n objets pris n à la fois). $_nP_n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = ?$

Nous savons que $_3P_3 = 3 * 2 * 1 = 3!$ or, $_3P_3 = \frac{3!}{(3-3)!} = \frac{3!}{0!}$; donc $3! = \frac{3!}{0!}$, et ceci a du sens seulement si $0! = 1$. On vient donc de définir $0! = 1$ et $_nP_n = \frac{n!}{0!} = n!$

6.5 COMBINAISONS

Si A est un ensemble de n éléments, alors un sous-ensemble de A contenant X éléments est appelé une combinaison de n objets pris X à la fois. La notation représentant le nombre total de combinaisons (le nombre des sous-ensembles de A contenant X éléments chacun) est $_nC_X$. Tout comme dans les permutations, les répétitions ne sont pas permises; l'ensemble contenant $\{a, b, a\}$ est identique à l'ensemble $\{a, b\}$. La différence essentielle entre les combinaisons et les permutations est que dans une combinaison on ne tient pas compte de l'ordre

Exemple 1

Combien y a-t-il de combinaisons des 3 lettres a, b, c prises 2 à la fois?

1. On donne la liste: $\{a, b\}$, $\{a, c\}$, $\{b, c\}$, donc $_3C_2 = 3$.
2. On sait que les lettres a, b donnent 2 permutations (ab et ba), mais 1 combinaison a, b. Ainsi 2 lettres dans A donnent $_2P_2 = 2$ permutations, mais seulement 1 combinaison. Ici, il y a au total $_3P_2 = 6$ permutations. Donc pour trouver le nombre de combinaisons, nous devons diviser $_3P_2$ par $_2P_2$:

$$_3C_2 = \frac{_3P_2}{_2P_2} = \frac{6}{2} = 3$$

Exemple 2

De combien de manières peut-on former un comité de 3 membres pris dans un groupe de 10 personnes?

Un comité formé d'André, Bernard et Claude est identique au comité formé d'André, Claude et Bernard. L'ordre n'importe pas. On a donc à calculer ${}_{10}C_3 = ?$ Donner la liste serait difficile; essayons donc la 2e méthode du premier exemple. Il y a ${}_{10}P_3$ permutations de 10 membres pris 3 à la fois, mais les 3 membres du comité peuvent être permutés de 3! façons et donnent seulement une combinaison; (*ABC; ACB; BAC; BCA; CAB; CBA*) sont toutes les permutations d'André, Bernard et Claude, mais $\{A, B, C\}$ est la seule combinaison de ces trois personnes donc:

$$ {}_{10}C_3 = \frac{{}_{10}P_3}{{}_3P_3} = \frac{10!/7!}{3!/0!} = \frac{10!}{3!7!} = 120 $$

Généralement, le nombre de combinaisons de n objets pris X à la fois est donné par:

$$ {}_nC_X = \frac{{}_nP_X}{{}_XP_X} = \frac{n!}{X!(n - X)!} $$

Un exemple:

$$ {}_5C_3 = \frac{5!}{3!2!} = \frac{5 * 4 * 3 * 2 * 1}{(3 * 2 * 1)(2 * 1)} = 5 * 2 = 10 $$

Il est préférable de simplifier les facteurs communs au numérateur et au dénominateur avant d'effectuer le calcul. On peut également évaluer ${}_nC_X$ en écrivant au numérateur X facteurs décroissants de 1 en débutant par n, puis au dénominateur tout simplement $X!$

$$ {}_5C_3 = \frac{5 * 4 * 3}{3 * 2 * 1} = 10 $$

La plus grande difficulté à résoudre un problème sera de décider si l'ordre est important. Voici quatre questions dont la réponse est dans chaque cas ${}_4C_3 = 4$.

1. De combien de façons différentes peut-on choisir 3 des 4 nombres suivants 1; 3; 5; 7?
2. Si une pièce de monnaie est lancée 4 fois, combien de séquences peuvent contenir 3 faces?
3. Un sac contient 4 boules numérotées de 1 à 4. Si on pige 3 boules, combien de résultats différents sont possibles?
4. Un département de 4 personnes doit nommer 3 représentants pour aller à un congrès. De combien de façons peut-on le faire?

6.6 HYPOTHÈSES DE LA DISTRIBUTION BINÔMIALE

Nous continuons l'étude des distributions de probabilités et leur histogramme, mais nous limiterons notre étude aux problèmes qui répondent aux hypothèses suivantes:

1. La même expérience est répétée n fois (n essais sont faits).
2. Chaque expérience a deux résultats possibles. Appelons ces 2 résultats «succès» et «échec». (Un succès n'implique pas un bon résultat et réciproquement. Si on lance un dé, un succès peut être d'obtenir un 2 et un échec d'obtenir 1; 3; 4; 5; 6.) Soit p la probabilité de succès lors d'un essai et q la probabilité d'un échec. Comme on doit obtenir un succès ou un échec lors d'un essai, on aura $p + q = 1$ et $q = 1 - p$.
3. Le résultat d'un essai est indépendant du résultat lors d'un autre essai. Ce sont les hypothèses de la distribution binômiale: nous justifions le mot **binômiale** à la section 6.8.

Si une pièce de monnaie est lancée trois fois, les hypothèses sont satisfaites: $n = 3$; un succès est interprété, par exemple, par «obtenir une face» et $p = 1/2$, $q = 1/2$; de plus le résultat obtenu lors d'un jet n'est pas affecté par les autres jets.

Considérons un médecin qui essaie un nouveau médicament sur 35 patients. Il veut comparer ce nouveau médicament avec l'ancien et voit que dans 70% des cas le nouveau est plus efficace, donc $p = 0{,}70$; $q = 0{,}30$ et $n = 35$.

Supposons que l'on pige sans remise 4 cartes d'un jeu de 52 et que l'on a un succès si on pige un coeur. Ici les hypothèses ne sont **pas** vérifiées, car la probabilité de piger un coeur au premier essai est de 13/52 et de 12/51 au deuxième essai si on pige un coeur en premier, ou de 13/51 si on n'a pas pigé un coeur en premier. Le résultat de la seconde pige n'est pas indépendant du résultat obtenu lors de la première.

6.7 p (X SUCCÈS EN n TENTATIVES)

Ici nous retrouvons un résumé de deux problèmes que nous avons déjà effectués et que nous pouvons résoudre facilement.

Exemple 1

Une pièce de monnaie régulière est lancée n fois. Quelle est la probabilité d'obtenir face, 0; 1; . . . . n fois si (a) $n = 1$; (b) $n = 2$; (c) $n = 3$? $p = 1/2$, $q = 1/2$. Posons X = nombre de faces obtenues.

(a) $n = 1$

X	$p(X)$
0	1/2
1	1/2
	1

(b) $n = 2$

X	$p(X)$
0	1/4
1	1/2
2	1/4
	1

(c) $n = 3$

X	$p(X)$
0	1/8
1	3/8
2	3/8
3	1/8
	1

Exemple 2

Un sac contient 4 boules rouges et 6 boules vertes. Si n boules sont pigées au hasard avec remise, quelle est la probabilité d'obtenir exactement 0, 1, ... n boules rouges si (a) $n = 1$; (b) $n = 2$; (c) $n = 3$? $p = 0{,}4$; $q = 0{,}6$. Soit X le nombre de boules rouges.

(a)

$n = 1$

X	$p(X)$
0	0,6
1	0,4
	1

(b)

$n = 2$

X	$p(X)$
0	$(0{,}6)^2 = 0{,}36$
1	$2 * 0{,}4 * 0{,}6 = 0{,}48$
2	$(0{,}4)^2 = 0{,}16$
	1

(c)

$n = 3$

X	$p(X)$
0	$(0{,}6)^3 = 0{,}216$
1	$3 * 0{,}4 * (0{,}6)^2 = 0{,}432$
2	$3 * (0{,}4)^2 (0{,}6)^1 = 0{,}288$
3	$(0{,}4)^3 = 0{,}064$
	1

Revoyons rapidement comment ces probabilités furent obtenues:
p (1 face en 3 essais) $= p$ (FPP ou PFP ou PPF) $= p$ (FPP) $+ p$ (PFP) $+ p$ (PPF) $= 1/2 * 1/2 * 1/2 + 1/2 * 1/2 * 1/2 + 1/2 * 1/2 * 1/2 = 3 * (1/2)^3 = 3 * 1/8$.
p (2 boules rouges si on pige 3 boules du sac) $= p$ (RRV ou RVR ou VRR) $= p$ (RRV)$+ p$ (RVR) $+ p$ (VRR) $= (0{,}4)^2 * 0{,}6 + 0{,}4 * 0{,}6 * 0{,}4 + 0{,}6 * (0{,}4)^2 = 3 * (0{,}4)^2 * (0{,}6) = 0{,}288$.

Nous voulons résoudre des problèmes semblables pour de plus grandes valeurs de n. Un exemple: la probabilité d'obtenir 17 faces si une pièce est lancée 20 fois? Revenons au cas où on a 1 face en 3 tentatives. Jusqu'à maintenant, on a donné la liste des différentes possibilités qui contiennent 1 face: p (FPP ou PFP ou PPF). Or, on sait que l'on peut obtenir 1 face en 3 tentatives de ${}_3C_1$ façons et que la probabilité que chaque combinaison contenant 1 face et 2 piles est de $1/2 * (1/2)^2$; donc p (1 face en 3 tentatives) $= {}_3C_1 * 1/2 * (1/2)^2$

Un autre exemple: p (2 boules rouges quand on pige 3 boules d'un sac contenant 4 boules rouges et 6 boules vertes). On peut obtenir 2 R en 3 tentatives de ${}_3C_2$ façons; en effet ${}_3C_2 = \dfrac{3 * 2}{1 * 2} = 3$ façons qui sont RRV, RVR, VRR et chacune de celles-ci ont la même probabilité qui est $0{,}4^2 * 0{,}6$, car p (R) $= 0{,}4$ et p (V) $= 0{,}6$. Finalement, p (2 R en 3 piges) $= {}_3C_2 * 0{,}4^2 * 0{,}6$.

On peut maintenant généraliser:

> Si une expérience est répétée n fois et que les hypothèses d'une distribution binômiale sont satisfaites, alors la probabilité de X succès sera donnée par p (X succès) $= {}_nC_X p^X q^{n-X}$.

En effet, chaque combinaison de X succès et $(n - X)$ échecs a une probabilité de $p^X q^{n-X}$ et on a ${}_nC_X$ combinaisons de X succès en n tentatives. X doit être plus petit ou égal à

n; de plus, il faut se rappeler que $p^0 = q^0 = 1$. On peut vérifier, en utilisant cette formule, les résultats des exemples 1 et 2.

La table 2 appendice C donne les valeurs de ${}_nC_X p^X q^{n-X}$ pour n variant de 2 à 20 avec différentes valeurs de p.

Exemple 3

Trouver ${}_{15}C_4\,(0,6)^4\,(0,4)^{11}$. Ici $n = 15$, $X = 4$. Dans la table 2 appendice C, on trouve 15 dans la colonne n puis 4 dans la colonne X, puis on se déplace horizontalement jusqu'à la colonne $p = 0,60$ et on trouve ${}_{15}C_4\,(0,6)^4\,(0,4)^{11} = 0,007$.

Exemple 4

Une pièce de monnaie est lancée 3 fois. (a) Quelle est la probabilité d'obtenir 0 face, 1 face, 2 faces, 3 faces. (b) Quelle est la probabilité d'obtenir au plus 1 face en 3 essais? (c) Quelle est la probabilité d'obtenir au moins 1 face?

(a) $n = 3$; $p = 0,5$; soit X = le nombre de faces.

X	${}_nC_X p^X q^{n-X} = p(X)$
0	0,125
1	0,375
2	0,375
3	0,125

(b) p (0 ou 1) $= {}_3C_0 p^0 q^3 + {}_3C_1 p^1 q^2 = 0,125 + 0,375 = 0,50$.

(c) p (1;2 ou 3) $= 1 - p(X = 0) = 1 - {}_3C_0 p^0 q^3 = 1 - 0,125 = 0,875$.

Exemple 5

Une famille a 5 enfants. Quelle est la probabilité que (a) au moins un enfant soit un garçon? (b) au moins 2 soient des garçons? En supposant que la probabilité d'avoir un garçon est de 0,5.

Donc $p = 0,5$; $q = 0,5$ et X = nombre de garçons en utilisant la table 2 appendice C pour $n = 5$.

(a) $p\,(X = 0) = 0,031$ et $p\,(X = 1$ ou plus$) = 1 - p\,(X = 0) = 1 - 0,03 = 0,97$.

(b) $p\,(X = 1) = 0,156$; $p\,(X = 2$ ou plus$) = 1 - p\,(X = 0$ ou $1) = 1 - p\,(X = 0) - p\,(X = 1) = 1 - 0,031 - 0,156 = 0,81$.

On peut écrire aussi $p(X = 2$ ou plus$) = p\,(X = 2) + p\,(X = 3) + p\,(X = 4) + p\,(X = 5) = 0,312 + 0,312 + 0,156 + 0,031 = 0,81$.

6.8 POURQUOI L'EXPRESSION «DISTRIBUTION BINÔMIALE»?

Cette expression vient du développement d'un binôme de la forme $(x + y)^n$, développement qui fut fait par Newton vers 1670. Par exemple:

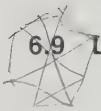
$$(x + y)^n = x^n + {}_nC_{n-1}x^{n-1}y + {}_nC_{n-2}x^{n-2}y^2 + \ldots + {}_nC_1xy^{n-1} + y^n.$$

$$\begin{aligned}(x + y)^3 &= x^3 + {}_3C_2x^2y + {}_3C_1xy^2 + y^3 \\ &= x^3 + 3x^2y + 3xy^2 + y^3.\end{aligned}$$

On constate que dans une distribution binômiale, on a

$$(p + q)^n = p^n + {}_nC_{n-1}p^{n-1}q + {}_nC_{n-2}p^{n-2}q^2 + \ldots + {}_nC_1pq^{n-1} + q^n$$

et cette dernière expression prouve que la somme des probabilités de la distribution binômiale est 1, car $q = 1 - p$ et $p + q = 1$; donc $(p + q)^n = (1)^n = 1$.



6.9 LA MOYENNE ET L'ÉCART-TYPE D'UNE DISTRIBUTION BINÔMIALE

Une distribution binômiale est un cas particulier d'une distribution de probabilités et aussi un cas particulier d'une distribution de fréquences. Ainsi on peut s'attendre à trouver sa moyenne et son écart-type, comme nous l'avons fait à l'article 6.1. On obtient les formules suivantes:

$$\mu = np,\ \sigma = \sqrt{npq}$$

La formule de la moyenne peut être trouvée intuitivement. Si une pièce de monnaie est lancée 100 fois, on peut s'attendre à obtenir 50 faces et $\mu = np = 100 * 1/2 = 50$. Si un dé est lancé 60 fois, on peut s'attendre à obtenir 10 fois la face 3; en effet $n = 60$, $p = 1/6$ et $\mu = np = 60 * 1/6 = 10$.

Exemple 1

Trouver la moyenne et l'écart-type de la distribution binômiale si $n = 3$ et $p = 0{,}5$.

$$\mu = np = 3 * 0{,}5 = 1{,}5$$
$$\sigma = \sqrt{npq} = \sqrt{3 * 0{,}5 * 0{,}5} = 0{,}9$$

Exemple 2

Deux billes sont pigées successivement, avec remise, dans un sac contenant 4 billes rouges et 6 billes vertes. Trouver la moyenne et l'écart-type pour la distribution de probabilité du nombre de billes rouges pigées:

$$p = 0{,}4;\ q = 0{,}6;\ n = 2$$
$$\mu = np = 2 * 0{,}4 = 0{,}8$$
$$\sigma = \sqrt{npq} = \sqrt{2 * 0{,}4 * 0{,}6} = \sqrt{0{,}48} = 0{,}7$$

6.10 APPLICATIONS DE LA DISTRIBUTION BINÔMIALE

La connaissance de la distribution binômiale nous aidera à comprendre la distribution normale qui sera étudiée au prochain chapitre. Voici quelques applications ayant un certain intérêt.

Exemple 1

Un médecin étudiant un nouveau traitement pour guérir la leucémie sait que 30% des patients sont toujours vivants 8 ans après le début du traitement. Il essaie son nouveau médicament sur 10 patients atteints de leucémie, choisis au hasard, et découvre que 7 sont encore vivants 8 ans après le début du traitement. Peut-il affirmer que son nouveau traitement est préférable à l'ancien? Reprenons la situation et écrivons-la en termes de probabilité: quelle est la probabilité qu'au moins 7 patients sur 10 soient vivants 8 ans après le début de l'ancien traitement? $p = 0{,}3$; $q = 0{,}7$; p (au moins 7 vivants) $= p$ (7 ou 8, ou 9, ou 10 vivants) $= 0{,}009 + 0{,}001 + 0{,}00... + 0{,}00... \cong 0{,}01$.

La probabilité qu'au moins 7 patients sur 10 seront vivants 8 ans après le début de l'ancien traitement est approximativement 1%. Donc son nouveau médicament semble très prometteur, car il a obtenu la certitude que 7 patients sur 10 seront vivants 8 ans après le début du traitement. Il peut donc continuer à se servir de celui-ci et l'essayer sur un plus grand nombre de patients.

Exemple 2

Un contremaître a découvert que 1% des boulons produits par une machine sont défectueux. Il prend un échantillon de 600 boulons et en trouve 8 défectueux. Quelle est la probabilité qu'il y ait 8 boulons ou plus de défectueux dans l'échantillon de 600, si la machine en produit 1% de défectueux?
$p = 0{,}01$; $q = 0{,}99$; p (8 défectueux ou plus sur 600) $= 1 - p$ (7 défectueux ou moins). La table 2 appendice C ne permet pas de résoudre pour $n = 600$. Le calcul serait passablement long. Nous verrons au prochain chapitre comment trouver une bonne approximation de cette probabilité en utilisant la loi normale.

6.11 VOCABULAIRE ET SYMBOLES

distribution de probabilités
permutation
factoriel $n!$
distribution binômiale
distribution de fréquences relatives
arrangement ${}_nP_x$
combinaison ${}_nC_x$

6.12 EXERCICES

1. (a) Les résultats d'une population sont distribués selon la table ci-dessous. Si un résultat est choisi au hasard, quelle est la probabilité d'obtenir 2? 5? 8? 5 ou 8? que le résultat soit plus petit que 11?

X	f
2	6
5	12
8	8
11	14

(b) Donner la distribution de probabilités et tracer son histogramme.
(c) Trouver la moyenne et l'écart-type de cette distribution de probabilités.

2. (a) Trois pièces de monnaie doivent être lancées 800 fois. Utiliser la table ci-dessous pour déterminer le nombre de fois où l'on obtiendra i) aucune pièce ne donne une face; ii) 1 pièce donne face; iii) 2 pièces donnent face; iv) les 3 donnent face.

Nombre de faces	$p\ (X)$
0	1/8
1	3/8
2	3/8
3	1/8

(b) Trouver la moyenne et l'écart-type de la distribution de probabilités de la table ci-dessus.
(c) Quel est le pourcentage des lancers qui présenteront deux pièces ayant un résultat face?

3. (a) Un sac contient 4 boules rouges et 6 boules vertes. Si 3 boules sont pigées successivement avec remise, la distribution de probabilités de piger X boules rouges est donnée dans l'exemple 2 section 6.7. Tracer un histogramme de cette distribution.
(b) Trouver la moyenne et l'écart-type de la distribution de probabilités.

4. La distribution de probabilités suivante donne le résultat moyen de 2 000 étudiants.

RM	Probabilité
3,00 – 4,00	0,24
2,00 – 2,99	0,62
1,00 – 1,99	0,12
0,00 – 0,99	0,02

(a) Trouver la moyenne et l'écart-type de cette distribution.
(b) Écrire la distribution de fréquences correspondante.

5. (a) ${}_9P_3 = ?$ (b) ${}_9C_3 = ?$ (c) ${}_9C_6 = ?$

6. Si une plaque d'immatriculation est formée de 2 lettres différentes, suivies de trois chiffres différents, combien de plaques différentes peut-on obtenir? (N.B.: les lettres O et I ne sont pas utilisées.)

7. De combien de manières peut-on choisir un président et un vice-président dans une association de 60 membres?

8. Un enfant possède des pièces de 1c; 5c; 10c; 25c; 50c; $1.00 (dollar en

argent). S'il décide de déposer 3 pièces dans sa tirelire, quels sont les montants d'argent qu'il peut déposer?

9. De combien de manières peut-on former un comité de 4 membres choisis dans un groupe de 20 personnes?

10. De combien de façons différentes peut-on asseoir 5 personnes sur une rangée de 5 chaises?

11. Une association de 100 professeurs doit nommer 5 membres pour former un comité de sélection. De combien de façons est-ce possible?

12. Combien de mots de 3 lettres peut-on former avec les lettres du mot «périodique» si une lettre ne peut être répétée?

13. (a) Un groupe de 4 garçons et 3 filles décide de s'élire un comité de 3 personnes pour assister à un congrès. Combien y a-t-il de comités possibles?
(b) Combien de comités contiennent exactement une fille?
(c) Combien de comités contiennent au moins une fille?

14. Une urne contient 8 boules numérotées de 1 à 8. De combien de façons peut-on piger 3 boules successivement et sans remise?

15. Un ensemble U contient 4 éléments. Combien de sous-ensembles peut-on former?

●16. De combien de façons peut-on asseoir 4 hommes et 3 femmes dans une rangée si on doit alterner les hommes et les femmes?

●17. De 5 économistes et 6 sociologues, on doit former un comité de 3 économistes et 3 sociologues. Combien de comités peut-on former si
(a) tout le monde peut y participer?
(b) un économiste ne peut y participer?
(c) 2 sociologues doivent absolument faire partie de ce comité?

●18. Combien de nombres de 3 chiffres peut-on former (a) si les répétitions ne sont pas permises, 0 ne peut être le premier chiffre et le nombre doit être pair? (b) si chaque chiffre a 2 de moins que le précédent (864 est admis, mais pas 863 ou 368)? (c) si une répétition est admise mais le premier chiffre n'est pas 0?

19. Quelle est la probabilité d'obtenir le chiffre un
(a) exactement 4 fois si un dé est lancé 5 fois?
(b) au moins 4 fois?
(c) au plus 4 fois?

20. Si 4 cartes sont pigées d'un jeu régulier de 52 cartes, quelle est la probabilité de piger 4 coeurs?

21. Un questionnaire à choix multiples comprend 10 questions ayant 5 choix chacune. Quelle est la probabilité de répondre correctement à 8 questions si on procède au hasard?

22. Un sac contient 10 boules rouges, 50 boules vertes, 40 boules noires. Quelle est la probabilité de piger 3 boules rouges en 4 tentatives si chaque boule est remise dans le sac après chaque pige?

23. Quel est le nombre moyen de succès espéré en 200 tentatives si l'expérience est binômiale et que la probabilité d'un succès est de 0,4? Quel est l'écart-type de cette distribution de probabilités?

24. Posons que la probabilité d'avoir une fille est de 1/2 et que le sexe d'une naissance n'est pas affecté par les naissances précédentes.

(a) Trouver la distribution de probabilités du nombre de filles dans une famille de 4 enfants.

(b) Trouver la moyenne et l'écart-type de cette distribution.

● 25. Les années antérieures nous montrent que 10% des élèves sont absents pour cause de maladie au début de mars. Quelle est la probabilité que, cette année, il n'y ait pas plus de 2 absents dans une classe de 30 au début de mars?

26. Après plusieurs tests, un fabricant de graines démontre que 90% de celles-ci germeront. Il les vend en paquets de 100 et donne une garantie de 85% de germination. Quelle est la probabilité qu'un paquet particulier ne rencontre pas les clauses de la garantie?

27. Une paire de dés est lancée 200 fois.

(a) Quel est le nombre moyen de fois où la somme donnera 7?

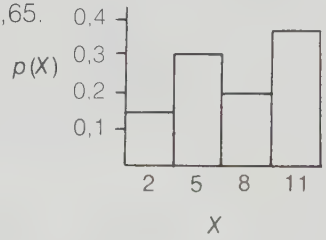
(b) Quel est le nombre de fois le plus probable où on aura une somme de 7?

RÉPONSES

1. (a) $6/40 = 0{,}15;\ 0{,}30;\ 0{,}20;\ 0{,}50;\ 1 - 0{,}35 = 0{,}65.$

(b)

X	$p\ (X)$
2	0,15
5	0,30
8	0,20
11	0,35
	1,00

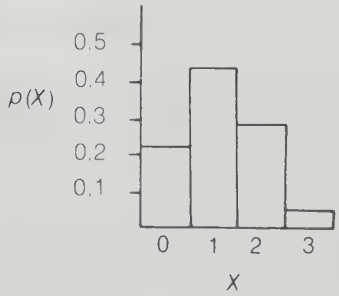


(c) $\mu = 2 * 0{,}15 + 5 * 0{,}30 + 8 * 0{,}20 + 11 * 0{,}35 = 7{,}25$

$\sigma^2 = (-5{,}25)^2 * 0{,}15 + (-2{,}25)^2 * 0{,}30 + (0{,}75)^2 * 0{,}20 + (3{,}75)^2 * 0{,}35 = 10{,}69$

$\sigma = \sqrt{10{,}69} = 3{,}3$

3. (a)



(b) $\mu = 1{,}2;\ \sigma = 0{,}8$

5. (a) $9 * 8 * 7 = 504$. (b) $\frac{9 * 8 * 7}{1 * 2 * 3} = 84$. (c) $\frac{9!}{6!3!} = {}_9C_3 = 84$.

7. C'est une permutation, car A président et B vice-président est différent de B président et A vice-président. ${}_{60}P_2 = 60 * 59 = 3\,540$.

9. ${}_{20}C_4 = \frac{20 * 19 * 18 * 17}{4 * 3 * 2 * 1} = 4845$.

11. ${}_{100}C_5 = 75\ 287\ 520$

13. (a) ${}_7C_3 = 35$; (b) Choisir 1 fille parmi 3 et 2 garçons parmi 4. ${}_3C_1 * {}_4C_2 = 3 * 6 = 18$; (c) ${}_3C_1 * {}_4C_2 + {}_3C_2 * {}_4C_1 + {}_3C_3 * {}_4C_0 = 18 + 12 + 1 = 31$ **ou** 35 − (nombre de comités sans fille) $= 35 - 4 = 31$.

15. Première solution: un sous-ensemble peut avoir 0; 1; 2; 3; ou 4 éléments, donc ${}_4C_0 + {}_4C_1 + {}_4C_2 + {}_4C_3 + {}_4C_4 = 1 + 4 + 6 + 4 + 1 = 16$; deuxième solution: pour chaque élément du U, on a deux possibilités; l'élément est ou n'est pas dans le sous-ensemble. On aura donc $2 * 2 * 2 * 2 = 16$ sous-ensembles.

17. (a) ${}_5C_3 * {}_6C_3 = 10 * 20 = 200$.

(b) ${}_4C_3 * {}_6C_3 = 4 * 20 = 80$ (choisir 3 des 4 économistes qui peuvent faire partie du comité, puis 3 des 6 sociologues).

(c) ${}_5C_3 * {}_4C_1 = 10 * 4 = 40$.

19. (a) ${}_5C_4\,(1/6)^4(5/6)^1 = 25/6^5 \approx 0{,}003$ (b) $25/6^5 + {}_5C_5\left(\frac{1}{6}\right)^5 = \frac{26}{6^5} \approx 0{,}003$

(c) $1 - 0{,}0001 = 0{,}9999$

21. ${}_{10}C_8\,(0{,}2)^8\,(0{,}8)^2 = 0{,}00007$

23. $\mu = 200 * 0{,}4 = 80$

$\sigma = \sqrt{200 * 0{,}4 * 0{,}6} = \sqrt{48} = 6{,}9$.

25. On ne peut répondre, car les hypothèses de la binômiale ne sont pas vérifiées, un enfant malade peut transmettre sa maladie (grippe, rhume...).

27. (a) $200 * \frac{1}{6} = 33{,}3$ (b) 33.

Chapitre 7

Distributions normales et résultats standardisés

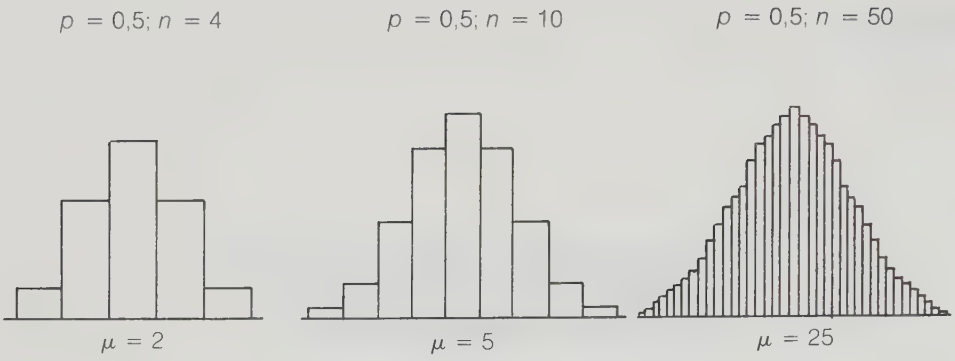
À la fin du chapitre 2, nous avons parlé brièvement de la distribution continue. Dans la plupart des distributions de fréquences que nous avons utilisées, X pouvait prendre un nombre fini de valeurs. Nous allons maintenant nous préoccuper des distributions continues d'une population dans laquelle est possible un nombre infini de résultats. Le polygone de fréquences qui illustre une distribution discrète de X deviendra une courbe régulière.

La distribution continue la plus importante est sans contredit la distribution normale. Il existe plusieurs sortes de distributions normales; nous étudierons en premier la distribution normale centrée et réduite dont la moyenne est 0 et l'écart-type est 1. Ensuite, nous introduirons les résultats standardisés où l'unité de mesure est l'écart-type. Les résultats standardisés possèdent plusieurs applications, ils permettent en particulier de comparer entre eux des résultats provenant de distributions normales centrées, réduites différentes.

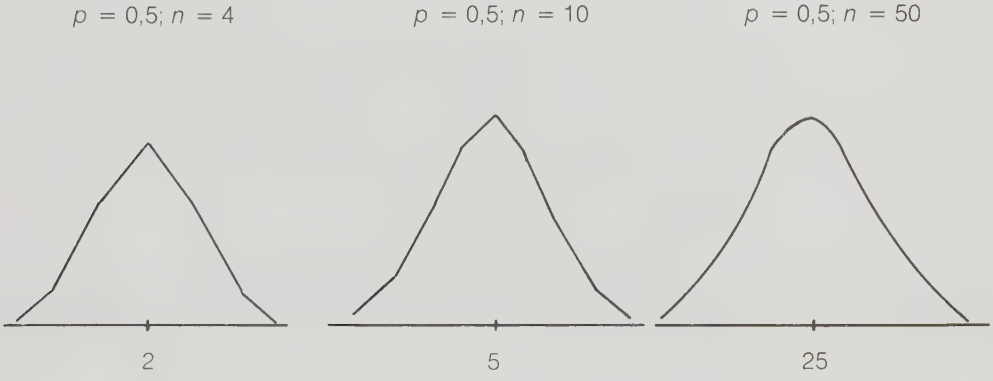
Nous verrons également comment trouver la cote z pour différentes distributions. On montrera que la distribution normale est une bonne approximation d'une distribution binômiale en se rappelant que la moyenne d'une distribution binômiale est $\mu = np$ *et que son écart-type est* $\sigma = \sqrt{npq}$.

7.1 HISTOGRAMME D'UNE DISTRIBUTION BINÔMIALE

Considérons quelques distributions binômiales ayant $p = 1/2$ et différentes valeurs de n. Voici les histogrammes pour $n = 4$; 10; 50 (se rappeler que ce sont des fréquences relatives ou des distributions de probabilités; les échelles verticales ont été omises pour qu'on puisse examiner convenablement la forme des histogrammes).



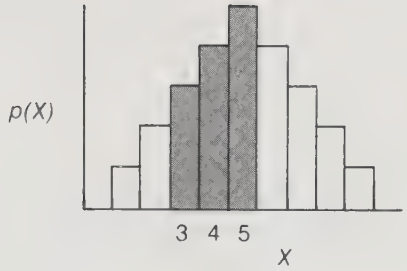
Il est à noter que dans chaque cas l'histogramme est symétrique par rapport à sa moyenne, que l'aire de chaque histogramme est 1 et que si un polygone de fréquences est tracé, celui-ci se rapproche d'une courbe en forme de cloche quand n augmente.



Si la hauteur d'un rectangle dans l'histogramme est connue, son aire est connue (sa largeur est 1); la probabilité d'obtenir un résultat compris entre deux valeurs est obtenue en trouvant l'aire comprise entre ces deux valeurs.

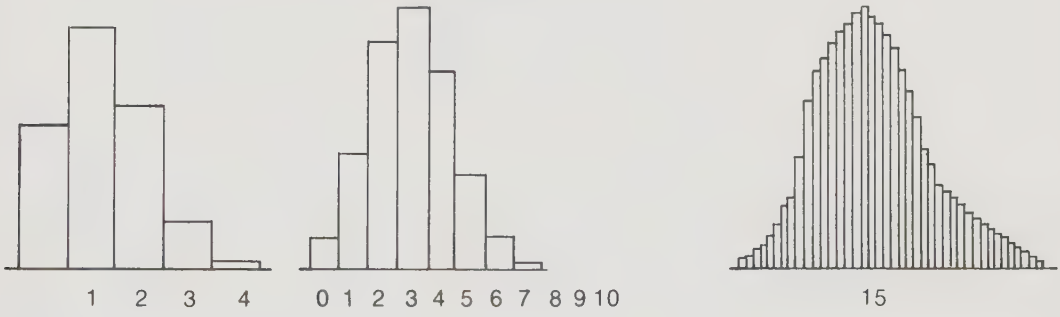
Exemple 1

Si une pièce de monnaie est lancée 10 fois, quelle est la probabilité d'obtenir une face 3; 4 ou 5 fois?



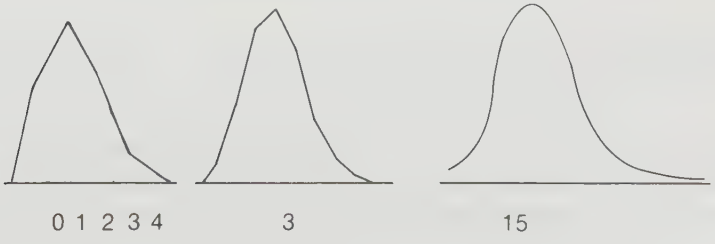
Si 3 est le milieu de classe dont les bornes sont 2,5 et 3,5 et de même pour 4 et 5, alors la probabilité d'avoir un résultat compris entre 2,5 et 5,5 quand $n = 10$ et $p = 0{,}5$ peut être trouvée de la même façon. Examinons maintenant les histogrammes pour $p = 0{,}3$; $n = 4$; 10 et 50.

$p = 0{,}3$; $n = 4$ $p = 0{,}3$; $n = 10$ $p = 0{,}3$; $n = 50$



Ici les aires sous les histogrammes demeurent égales à 1, mais les polygones de fréquences ne sont pas symétriques par rapport à leur moyenne pour des petites valeurs de n.

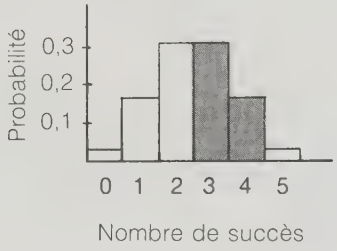
$p = 0{,}3$; $n = 4$ $p = 0{,}3$; $n = 10$ $p = 0{,}3$; $n = 50$



Au fur et à mesure que n augmente, le polygone devient de plus en plus symétrique et ressemble de plus en plus à une courbe en forme de cloche.

7.2 COURBES NORMALES

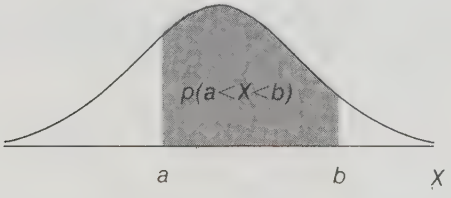
L'aire sous un histogramme d'une distribution de probabilités est 1 et la probabilité qu'un résultat soit compris entre a et b est représentée par l'aire de l'histogramme entre a et b. Si $p = 0{,}5$ et $n = 5$, la probabilité de 3 ou 4 succès (entre 2,5 et 4,5) en 5 essais est égale à l'aire ombragée ci-dessous.



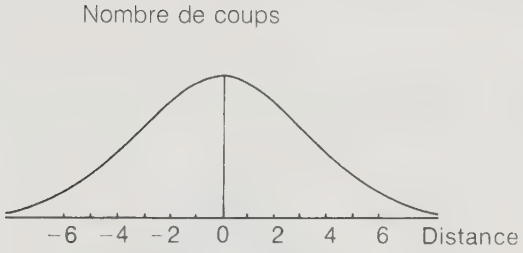
De la même manière, une courbe continue représente une distribution de probabilités si:

(a) L'aire totale sous la courbe est 1.

(b) La probabilité qu'un résultat soit entre a et b est égale à l'aire sous la courbe entre a et b.



Il y a un type de distribution de probabilités appelé **distribution normale** qui est fondamental en statistique moderne. C'est une distribution continue; la courbe qui la représente est appelée **courbe normale.** Historiquement, elle fut découverte au 18e siècle en faisant l'étude des erreurs (si un homme vise une cible avec un fusil et que les distances entre les points d'impact sur la cible et le centre de celle-ci sont mesurées, ces distances sont normalement distribuées). Cette façon de voir est plus facile à comprendre que la courbe en cloche obtenue à partir de la distribution binômiale pour une grande valeur de n.

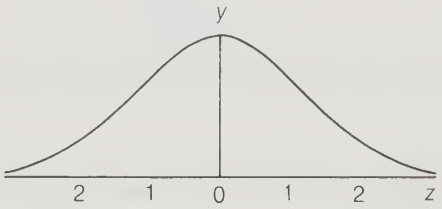


Il y a un grand nombre de courbes normales qui diffèrent par leur moyenne et leur écart-type. En premier lieu, nous étudierons la courbe normale centrée, réduite où $\mu = 0$ et $\sigma = 1$.

7.3 LA COURBE NORMALE CENTRÉE, RÉDUITE

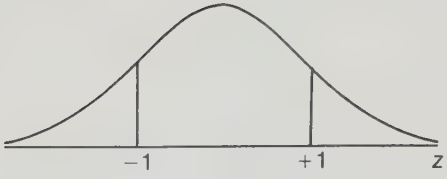
L'équation d'une courbe normale centrée, réduite est:

$$y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

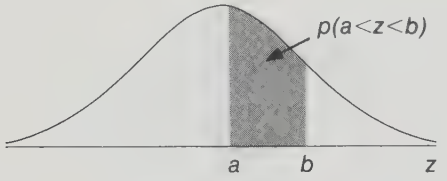


Il n'est pas nécessaire de mémoriser cette équation. Elle est donnée seulement pour montrer que son graphique est la courbe normale centrée, réduite. Il est à remarquer qu'on utilise z comme variable indépendante à la place de X et ce, dans le but de différencier la courbe normale centrée réduite des autres courbes normales. e est le nombre irrationnel (2,718...) rencontré en calcul différentiel. Bien qu'il ne soit pas nécessaire de connaître l'équation, nous devrons connaître certaines caractéristiques de la courbe normale centrée, réduite.

1. Elle s'étend indéfiniment dans les deux directions tout en s'approchant de l'axe horizontal sans le toucher au fur et à mesure que z devient grand ou petit. Pratiquement, la plus grande partie de l'aire sous cette courbe se trouve entre -3 et 3 et il ne sera pas nécessaire de considérer des valeurs de z en dehors de cet intervalle.
2. Le sommet de la courbe se trouve à $z = 0$ et sa moyenne est 0.
3. La courbe est symétrique par rapport à l'axe vertical.
4. Une courbe est concave vers le haut si: ∪ ou ⌣ ou ⌣ ; elle est concave vers le bas si: ∩ ou ⌢ ou ∩. La courbe normale centrée, réduite change de concavité à + 1 et − 1 (points d'inflexion).



5. L'écart-type d'une distribution normale centrée, réduite est 1.
6. Une distribution normale centrée, réduite est une distribution de probabilités; l'aire totale sous la courbe est 1 et la probabilité qu'un résultat soit entre *a* et *b* est égale à l'aire sous la courbe dans cet intervalle.

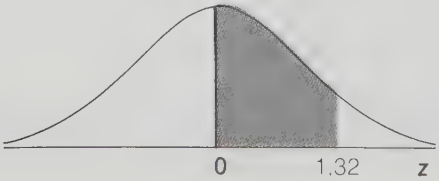


7.4 L'AIRE SOUS LA COURBE NORMALE CENTRÉE, RÉDUITE.

L'aire sous la courbe normale est donnée à la table 4 appendice C. Les entrées de cette table donnent l'aire sous la courbe normale entre la moyenne ($z = 0$) et la valeur de z. Dans cette table, la colonne de gauche donne la valeur de z à une décimale, la seconde décimale est donnée par la ligne du haut en se déplaçant horizontalement. Dans tous les problèmes, il est **bon d'esquisser l'aire cherchée.**

Exemple 1

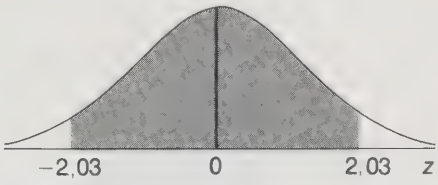
Quelle est la probabilité que z soit compris entre 0 et 1,32? Trouver dans la table 4 la valeur de 1,3 dans la colonne de gauche puis se déplacer horizontalement jusqu'à la colonne 0,02. On lit 0,4066 puis en arrondissant à deux décimales, $p\ (0 < z < 1{,}32) = 0{,}41$.



Exemple 2

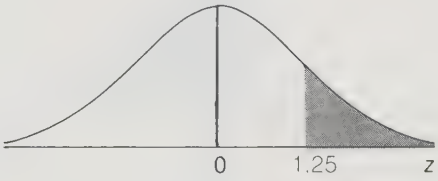
$p\ (-2{,}03 < z < 0)$

Les valeurs négatives de z ne sont pas données dans la table. Mais la courbe normale étant symétrique par rapport à $z = 0$, alors $p\ (-2{,}03 < z < 0) = p\ (0 < z < 2{,}03) = 0{,}48$.



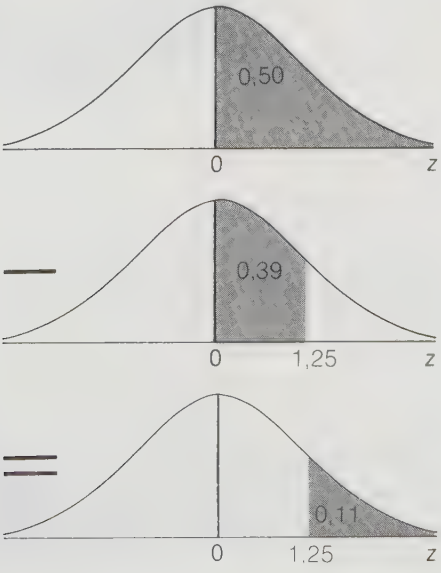
Exemple 3

Si un résultat z est choisi au hasard, quelle est la probabilité qu'il soit supérieur à 1,25? Il faut trouver $p\ (z > 1{,}25) = ?$



Sachant que l'aire totale sous la courbe est 1, on peut affirmer que l'aire de $z = 0$ à ∞ sera de 0,50 (la moitié de la courbe). Donc $p\ (z > 1{,}25) = 0{,}50 - p\ (0 < z < 1{,}25) = 0{,}50 - 0{,}39 = 0{,}11$.

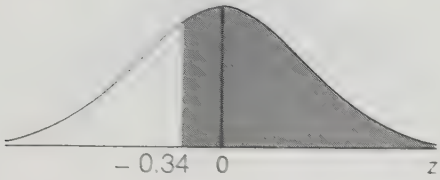
Illustrons cette situation:



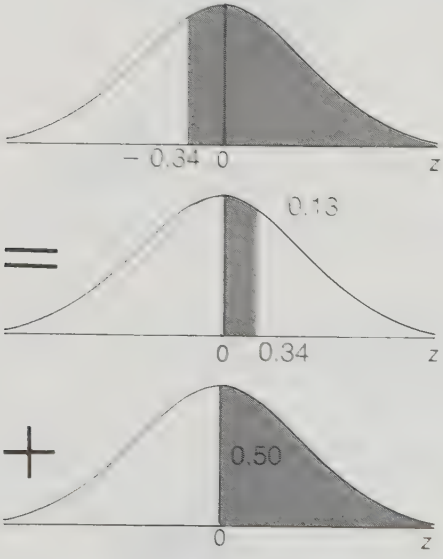
Exemple 4

$p\ (z > -0{,}34) = ?$

Par symétrie, l'aire entre $z = -0{,}34$ et $z = 0$ est la même que celle entre $z = 0$ et $z = 0{,}34$. Donc on devra additionner celle-ci à l'aire de la demi-courbe de droite.



Graphiquement:

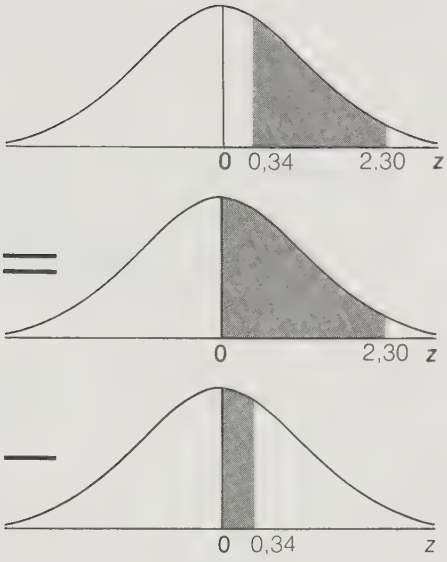


$$p\ (z > -0{,}34) = p\ (0 < z < 0{,}34) + p\ (z \geq 0)$$
$$= 0{,}1331 + 0{,}5$$
$$= 0{,}633$$

Exemple 5

Si un résultat est choisi au hasard dans une population distribuée normalement avec une moyenne de 0 et un écart-type de 1, quelle est la probabilité que celui-ci soit compris entre 0,34 et 2,30?

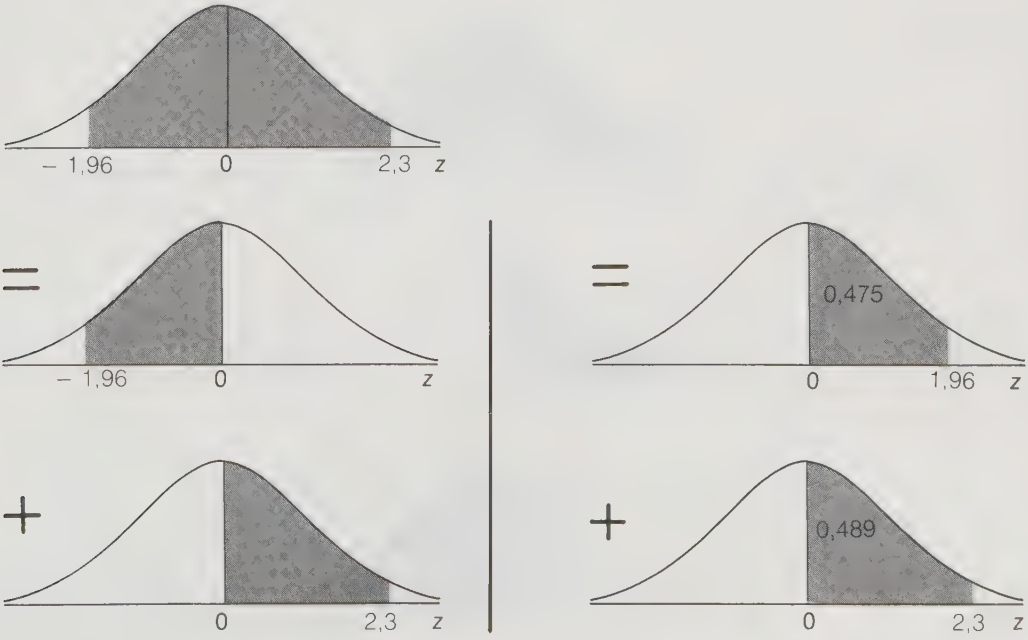
$p\ (0{,}34 < z < 2{,}30) = ?$



$$p\,(0{,}34 < z < 2{,}30) = p\,(0 < z < 2{,}30) - p\,(0 < z < 0{,}34) = 0{,}489 - 0{,}133 = 0{,}36.$$

Exemple 6

Trouver $p\,(-1{,}96 < z < 2{,}30) = ?$



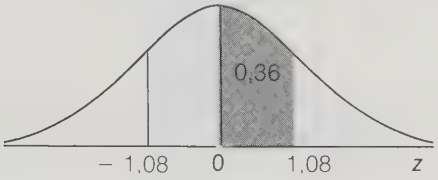
$$p\,(-1{,}96 < z < 2{,}3) = p\,(0 < z < 1{,}96) + p\,(0 < z < 2{,}3) = 0{,}475 + 0{,}489 = 0{,}96$$

Voir ici l'importance d'illustrer la situation.

Il peut arriver que l'on ait à résoudre le problème inverse: l'aire sous la courbe normale est connue (la probabilité) et on veut trouver la valeur de z. Il faudra utiliser la table 4 en sens inverse.

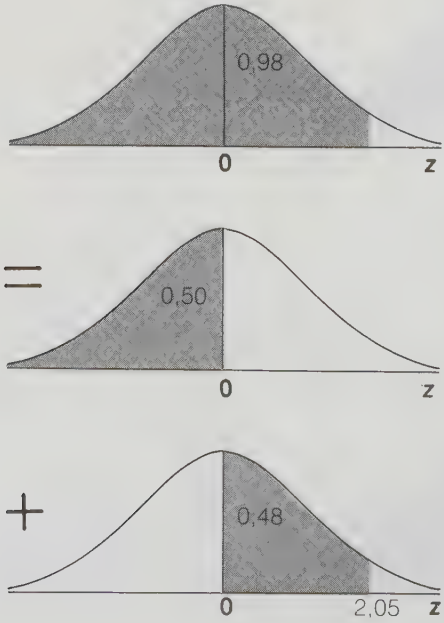
Exemple 7

Trouver z si l'aire sous la courbe normale entre 0 et z est 0,36. Dans la table 4, la valeur la plus rapprochée de 0,36 est 0,3599, elle est située sur la ligne $z = 1{,}0$ et dans la colonne 0,08, donc $z = 1{,}08$. Si $z = -1{,}08$ l'aire entre 0 et z est également 0,36 à cause de la symétrie.



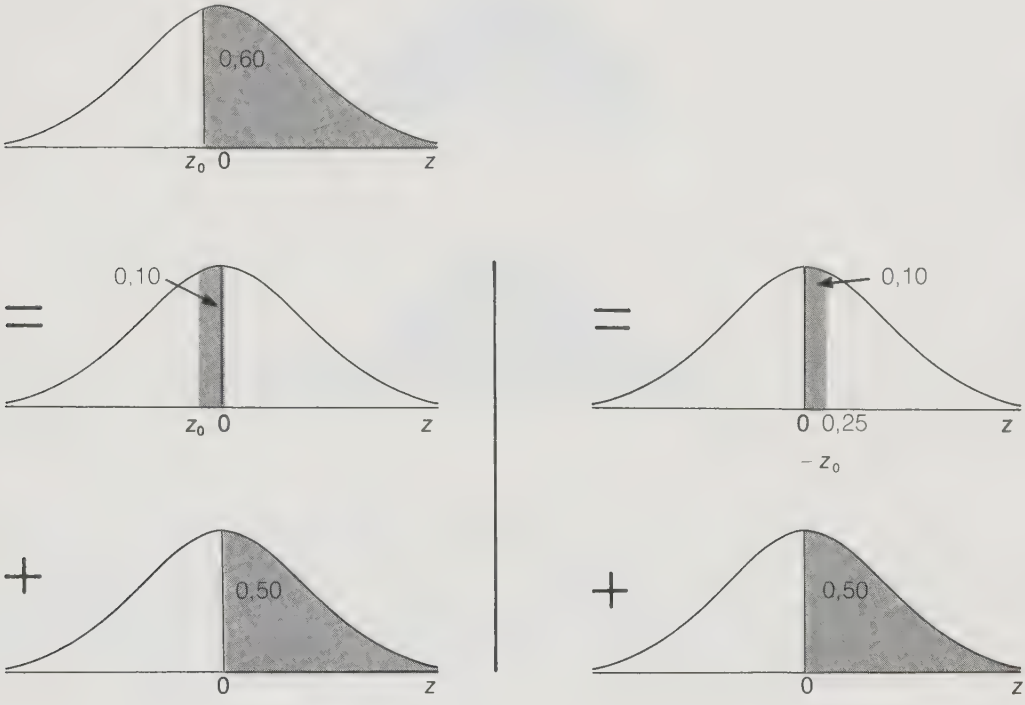
Exemple 8

Trouver le 98e centile des résultats z. Il faut trouver z sachant que l'aire à gauche de z est 0,98. L'aire à gauche de $z = 0$ est 0,50 donc l'aire entre 0 et z sera $0{,}98 - 0{,}50 = 0{,}48$. Par la table 4 on trouve $z = 2{,}05$.



Exemple 9

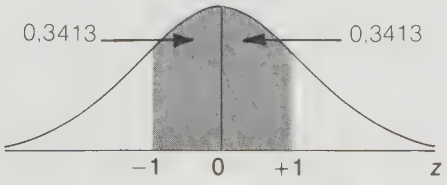
Quelle est la valeur de z_0 si 0,60 est la probabilité qu'un nombre z choisi au hasard soit plus grand que z_0?



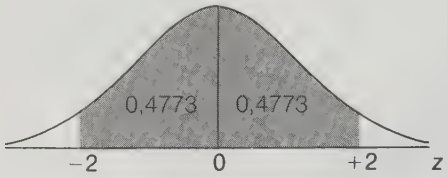
z_0 doit être négatif, car l'aire est supérieure à 0,50. L'aire correspondante à 0,10 donne $z = 0{,}25$ et comme z est à gauche de $z = 0$ on aura $z = -0{,}25$.

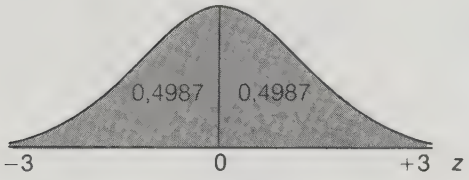
7.5 POURCENTAGES DES RÉSULTATS COMPRIS ENTRE LA MOYENNE ET $\pm$ 1; $\pm$ 2; $\pm$ 3 FOIS L'ÉCART-TYPE

Dans une courbe normale centrée réduite, $\sigma = 1$. Pour cette raison, les résultats sont souvent appelés résultats en unité d'écart-type. Par exemple, un résultat $z = 2{,}56$ est à 2,56 unités d'écart-type de la moyenne ($\mu = 0$).



$p\ (0 < z < 1) = 0{,}3413$ donc $2 * 0{,}3413 = 0{,}6826$ et 68% des résultats se trouvent compris entre $\pm$ 1 écart-type de la moyenne.





De la même manière, on trouve que dans une distribution normale, 95,5% des résultats sont compris entre ± 2 écart-types et 99,7% sont compris entre $\pm$ 3 écart-types. Il ne faut pas oublier que ces résultats sont valides uniquement pour une distribution normale. Beaucoup de débutants croient que ceux-ci sont applicables à toutes les distributions; c'est une grossière erreur comme nous le verrons dans l'exemple suivant.

Exemple 1

X: 1; 1; 1; 1; 1; 1; 1; 1; 1; 9; moyenne = 1,8, l'écart-type = 2,4. (Vérifier ces résultats en utilisant les formules de la section 7.6.) Quel pourcentage des résultats sont compris entre (a) 1 (b) 2 (c) 3 écart-types. La réponse à chacune de ces questions est 90%.

7.6 MOYENNE ET ÉCART-TYPE D'UNE POPULATION

Pour une population discrète on obtient les formules suivantes:

$$\text{Moyenne: } \mu = \frac{\Sigma X}{N}$$

$$\text{Écart-type: } \sigma = \sqrt{\frac{\Sigma (X - \mu)^2}{N}}$$

Dans le cas de données groupées on obtient $\mu = \dfrac{\Sigma fX}{N}$ et $\sigma = \sqrt{\dfrac{\Sigma f(X - \mu)^2}{N}}$

Nous ferons peu usage de ces formules sous cette forme.

Exemple 1

Soit les résultats X: 10; 30; 40; 40; trouver μ et σ.

X	$X-\mu$	$(X-\mu)^2$
10	−20	400
30	0	0
40	10	100
40	10	100
120		600

$\mu = 120/4 = 30.$
$\sigma = \sqrt{600/4} = \sqrt{150} = 12{,}2$

On peut utiliser la méthode courte pour calculer la moyenne: soustraire une constante A à chaque résultat, trouver la moyenne des nouveaux résultats puis ajouter la constante A pour obtenir la moyenne. La méthode courte pour calculer σ est donnée par la formule:

$$\sigma = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/N}{N}}.$$

Exemple 2

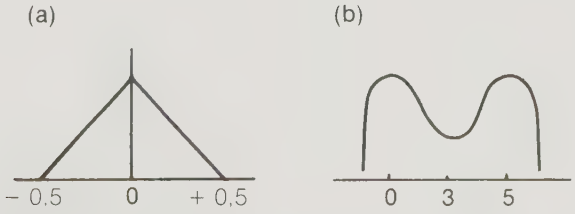
Soit les résultats Y: 1; 2; 4; 4; utiliser la méthode courte pour calculer σ.

$\Sigma Y = 11$, $N = 4$; Y^2: 1; 4; 16; 16; $\Sigma Y^2 = 37$

$$\sigma = \sqrt{\frac{37 - 11^2/4}{4}} = \sqrt{1{,}69} = 1{,}3$$

Que se passe-t-il dans une population continue? Les résultats individuels ne peuvent être donnés, car il y en a une infinité. En général, on doit connaître l'équation de la distribution puis, utiliser le calcul intégral pour calculer la moyenne. Il y a exception dans le cas de distributions symétriques, la moyenne étant située à l'axe de symétrie. Par exemple, dans les distributions suivantes on aura:

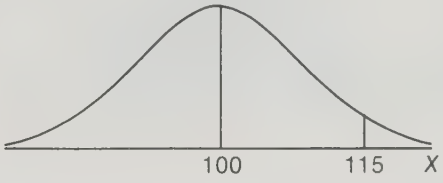
Exemple 3



(a) $\mu = 0$. (b) $\mu = 3$.

7.7 VARIABLE CENTRÉE ET RÉDUITE

En général, une distribution normale n'aura pas une moyenne de 0 et un écart-type de 1, mais elle aura une courbe en forme de cloche tout comme la distribution normale centrée, réduite.



Si on veut répondre à la question «Quel est le pourcentage des résultats compris entre 100 et 115, dans une distribution normale dont $\mu = 100$ et $\sigma = 10$», il semble que nous aurions besoin d'une autre table donnant une distribution normale ayant le μ et σ donnés. Comme il peut y avoir une infinité de valeurs pour μ et σ, il est impensable d'avoir une table pour chaque nouveau cas. Nous allons plutôt trouver un moyen de comparer une distribution normale centrée, réduite.

La cote z permettra de comparer 2 résultats provenant de 2 distributions différentes; un exemple: Jean a obtenu 78% en mathématiques. La moyenne de sa classe est 75% avec un écart-type de 2%. Pierre, lui, a obtenu 81% et la moyenne de sa classe est 80% avec un écart-type de 5%. Lequel des deux est le meilleur en mathématiques? La cote z servira à étudier une distribution normale où la moyenne n'est pas 0 et l'écart-type n'est pas 1.

La question importante à laquelle il faut répondre est la suivante: si on a un ensemble de résultats X dont on connaît la moyenne μ et l'écart-type σ, comment trouver un nouvel ensemble de résultats dont la moyenne sera 0 et l'écart-type 1?

$$z = \frac{X - \mu}{\sigma}$$

Exemple 1

Soient les résultats X suivants, trouver les cotes z.

X	$X - \mu$	$(X - \mu)^2$	$z = \frac{X - 8}{3}$
5	−3	9	−1
7	−1	1	−1/3
7	−1	1	−1/3
13	5	25	5/3
32		36	

$\mu = 32/4 = 8.$
$\sigma = \sqrt{36/4} = 3.$

Les résultats z sont -1; $-1/3$; $-1/3$; $5/3$. Vérifier que leur moyenne est 0 et leur écart-type est 1.

Exemple 2

(a) Trouver μ_x et σ_x des résultats X ci-dessous.
(b) Soit $Y = X - \mu_X$, trouver μ_y et σ_y.
(c) Soit $z = \dfrac{X - \mu_X}{\sigma_X} = \dfrac{Y}{\sigma_Y}$; trouver μ_z et σ_z.

X	$Y = X - \mu_X$	$(X - \mu_X)^2 = (Y - \mu_Y)^2$	$z = \dfrac{X - \mu_X}{\sigma_X} = \dfrac{Y}{\sigma_Y}$	$(z - \mu_z)^2$
5	−3	9	−1	1
7	−1	1	−1/3	1/9
7	−1	1	−1/3	1/9
13	5	25	5/3	25/9
32	0	36	0	4

(a) $\mu_X = 32/4 = 8$; $\sigma_X = \sqrt{36/4} = 3$.
(b) $\mu_Y = 0/4 = 0$; $\sigma_Y = \sqrt{36/4} = 3$.

Soustrayant de chaque X leur moyenne 8, on obtient un nouvel ensemble Y dont la moyenne est 0, mais dont l'écart-type reste inchangé. (Il faut se rappeler que le fait de soustraire une constante à chaque valeur X n'affecte pas leur dispersion et l'écart-type n'est pas affecté.)

(c) $\mu_z = 0/4 = 0$; $\sigma_z = \sqrt{4/4} = 1$. Divisant chaque résultat Y par leur écart-type 3, on obtient un nouvel ensemble z dont la moyenne demeure 0, mais dont l'écart-type est 1/3 du précédent.

$$\text{Si } \mu_Y = \frac{\Sigma Y}{N} = 0, \quad \text{alors } \frac{\Sigma(Y/\sigma)}{N} = \frac{1}{\sigma} * \frac{\Sigma Y}{N} = \frac{1}{\sigma} * 0 = 0;$$

$$\text{Si } \frac{\Sigma(Y-0)^2}{N} = \sigma^2, \quad \text{alors } \frac{\Sigma(Y/\sigma - 0)^2}{N} = \frac{1}{\sigma^2}\frac{\Sigma Y^2}{N} = \frac{1}{\sigma^2}\sigma^2 = 1.$$

Exemple 3

(a) Trouver μ et σ des résultats X ci-dessous.
(b) Trouver les résultats correspondants z.
(c) Trouver la moyenne et l'écart-type des résultats z.

X	$X-3$	$(X-3)^2$	$z = \dfrac{X-3}{\sqrt{6}}$	z^2
1	−2	4	$-2/\sqrt{6}$	4/6
1	−2	4	$-2/\sqrt{6}$	4/6
3	0	0	0	0
7	4	16	$4/\sqrt{6}$	16/6
12		24	0	4

(a) $\mu = 12/4 = 3$; $\sigma = \sqrt{24/4} = \sqrt{6}$.
(b) $z = \dfrac{X - \mu}{\sigma} = \dfrac{X - 3}{\sqrt{6}}$;

Les résultats z sont $-2/\sqrt{6}$; $-2/\sqrt{6}$; 0 et $4/\sqrt{6}$.

(c) La moyenne des $z = 0/4 = 0$, l'écart-type est:

$$= \sqrt{\frac{\Sigma(z-0)^2}{N}} = \sqrt{\frac{4}{4}} = 1.$$

Dans certains problèmes, z est connu et il faut retrouver le résultat X.

Exemple 4

Dans une population, la moyenne est 100 et l'écart-type est 10. Quel est le résultat si la cote z est $-1{,}3$?

$z = -1{,}3$; $\mu = 100$ et $\sigma = 10$.

$$z = (X - \mu)/\sigma,$$

$$-1{,}3 = \frac{X - 100}{10}$$

$$-13 = X - 100$$

$$X = 87.$$

Exemple 5

Dans une distribution dont $\mu = 40$ et $\sigma = 3{,}1$, quel est le résultat correspondant à $z = 2{,}1$?

$$z = (X - \mu)/\sigma$$
$$2{,}1 = (X - 40)/3{,}1$$
$$2{,}1 * 3{,}1 = 6{,}5 = X - 40$$
$$X = 46{,}5$$

Nous utiliserons la cote z très souvent dans les chapitres suivants. Il y a différentes formes de la variable z, mais dans tous les cas, il faut soustraire la moyenne et diviser par l'écart-type. Si Y est un ensemble de résultats dont la moyenne est $\overline{Y}$ et l'écart-type s, alors

$$z = (Y - \overline{Y})/s.$$

Un peu plus loin, nous étudierons des distributions dont les éléments sont des moyennes $\overline{X}$ (chaque résultat est la moyenne d'un certain échantillon) si on appelle $\mu_{\overline{X}}$ (lire mu de X barre) la moyenne de cette distribution $\sigma_{\overline{X}}$ l'écart-type, on aura alors $z = (\overline{X} - \mu_{\overline{X}})/\sigma_{\overline{X}}$.

Voici quelques autres exemples:

Résultats	Moyenne	Écart-type	$z =$
p	P	σ_p	$(p - P)/\sigma_p$
$\overline{X} - \overline{Y}$	$\mu_X - \mu_Y$	$\sigma_{(\overline{X}-\overline{Y})}$	$[\overline{X} - \overline{Y} - (\mu_X - \mu_Y)]/\sigma_{(\overline{X}-\overline{Y})}$
$\overline{X}$	μ	$\sigma_{\overline{X}}$	$(\overline{X} - \mu)/\sigma_{\overline{X}}$
X	$\overline{X}$	s	$(X - \overline{X})/s$
$p_X - p_Y$	$P_X - P_Y$	$\sigma_{(p_X - p_Y)}$	$\dfrac{(p_X - p_Y) - (P_X - P_Y)}{\sigma_{(p_X - p_Y)}}$

Ne vous en faites pas, ces formules reviendront plus tard selon les besoins. Voici quelques exemples.

Exemple 6

La proportion de la population fumant plus d'un paquet de cigarettes par jour est de 0,40. Une distribution p dont la moyenne est P et l'écart-type σ_p = 0,06. Quelle est la cote z correspondante à (a) 0,46 (b) 0,28?

Réponses

(a) $z = (p - P)/\sigma_p = (0{,}46 - 0{,}40)/0{,}06 = 1$
(b) $z = (0{,}28 - 0{,}40)/0{,}06 = -2.$

Il est à noter que le résultat $p = 0{,}46$ est à 1 écart-type (0,06)au-dessus de la moyenne (0,40) et sa cote z est 1. De même, le résultat $p = 0{,}28$ est à 2 écart-types $(2 * 0{,}06)$ en-dessous de la moyenne (0,40) et sa cote z est -2. La cote z nous indique donc à combien d'écart-types se situe un résultat d'une distribution normale de part et d'autre de la moyenne.

7.8 UTILISATION DES RÉSULTATS CENTRÉS ET RÉDUITS

Exemple 1

Un étudiant a obtenu 82% à son examen final de mathématiques; la moyenne de la classe est 75% avec un écart-type de 10. En sciences économiques, il a obtenu 86% et la moyenne est 80% avec un écart-type de 14. Est-il meilleur en mathématiques ou en sciences économiques?

En mathématiques $z = \dfrac{82 - 75}{10} = 0{,}70$

En sciences économiques $z = \dfrac{86 - 80}{14} = 0{,}43$

Son résultat de mathématiques est 0,70 unités au-dessus de la moyenne alors qu'en économique il est 0,43 unités au-dessus de la moyenne. Donc, il a une meilleure note en mathématiques.

Exemple 2

Dans un test d'aptitudes mathématiques, Pierre a obtenu un résultat brut de 88,3. Pour ce test la moyenne est 75 avec un écart-type de 8. Chaque année, les résultats sont ramenés à une moyenne de 500 avec un écart-type de 100 (pour pouvoir comparer les résultats d'une année à l'autre). Quel résultat net donnera-t-on à Pierre?

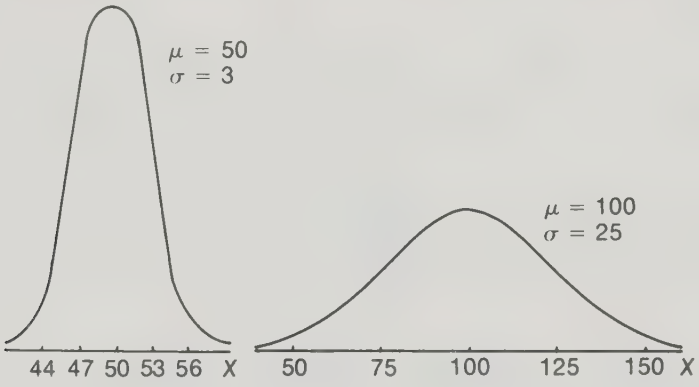
Le résultat de Pierre est $\dfrac{88{,}3 - 75}{8} = 1{,}66.$ Dans une distribution ou μ = 500 et $\sigma = 100$, un résultat de 666 correspond à une cote z de 1,66; en effet, il suffit de résoudre $1{,}66 = \dfrac{X - 500}{100}$ et on donnera à Pierre la note 666.

Nous allons montrer maintenant la relation qui existe entre une distribution normale quelconque et une distribution normale centrée, réduite.

7.9 DISTRIBUTIONS NORMALES EN GÉNÉRAL

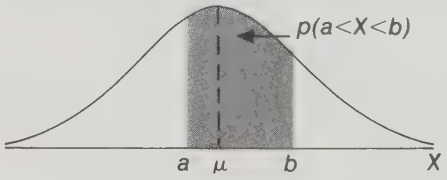
Il y a une famille de distributions normales. Des valeurs particulières de μ et σ donnent un membre de cette famille.

Voici quelques exemples:

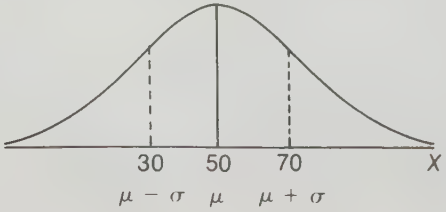


La courbe normale centrée réduite est aussi un membre de cette famille ($\mu = 0$ et $\sigma = 1$). Tous les membres de la famille ont certaines particularités en commun:

1. Ils sont tous des distributions de probabilités. L'aire totale sous la courbe est 1. L'aire sous la courbe entre $X = a$ et $X = b$ donne la probabilité que X soit compris entre a et b.



2. Toutes les courbes normales s'étendent de $-\infty$ à $+\infty$ et 95,5% de l'aire sous la courbe est comprise entre $\mu \pm 2\sigma$.
3. Chaque courbe normale est symétrique par rapport à la moyenne et a une forme de cloche.
4. Chaque courbe normale change de concavité à une distance de $\pm$ un écart-type (σ) de la moyenne μ. Ainsi si $\mu = 50$ et $\sigma = 20$, les points d'inflexion sont 30 et 70 comme ci-dessous.



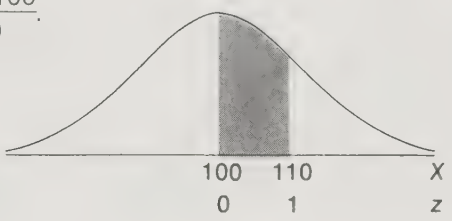
7.10 AIRE SOUS UNE COURBE NORMALE

Pour trouver l'aire sous une courbe normale de moyenne μ et d'écart-type σ entre $X = a$ et $X = b$, il suffit de trouver la cote z pour a et b (posons z_1 et z_2 ces valeurs) puis trouver l'aire sous la courbe normale centrée, réduite entre z_1 et z_2 en utilisant la table 4 appendice C. Examinons quelques cas.

Exemple 1

Des résultats X sont normalement distribués avec $\mu = 100$ et $\sigma = 10$. Quel pourcentage de résultats sont compris entre (a) 100 et 110 (b) 80 et 105?

(a) $z = \dfrac{X - 100}{10}$.

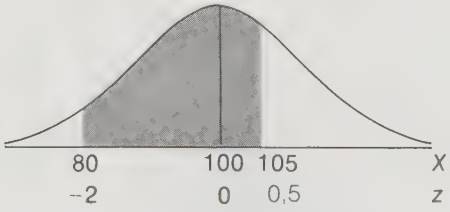


Quand $X = 100$, $z = \dfrac{100 - 100}{10} = 0$; $X = 110$, $z = 1$. $p(100 < X < 110)$

$= p\,(0 < z < 1) = 0{,}341$ donc 34,1% des résultats sont compris entre 100 et 110.

Quand $X = 80$, $z = \dfrac{80 - 100}{10} = -2$. Quand $X = 105$, $z = \dfrac{105 - 100}{10} = 0{,}5$

alors $p\,(80 < X < 105) = p\,(-2 < z < 0{,}5) = 0{,}4773 + 0{,}1915 = 0{,}669$



Donc 66,9% des résultats sont compris entre 80 et 105. Il est souvent très utile de montrer sur le même axe horizontal les valeurs de X et de z comme nous venons de le faire dans les figures précédentes.

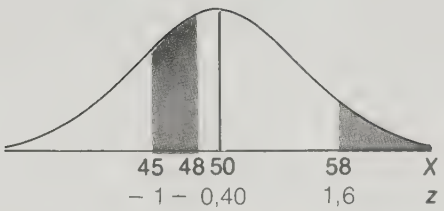
Exemple 2

Des résultats X sont normalement distribués avec $\mu = 50$ et $\sigma = 5$. Quelle est la probabilité qu'un résultat soit (a) entre 45 et 48 (b) plus grand que 58?

Ici, $z = \dfrac{X - \mu}{\sigma}$ alors $z = \dfrac{X - 50}{5}$;

disposons les résultats X et z dans un tableau.

X	z
45	−1
48	− 0,4
58	1,6



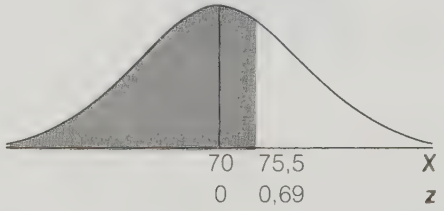
(a) $p(45 < X < 48) = p(-1 < z < -0,4) = 0,341 - 0,155 = 0,186$
(b) $p(X > 58) = p(z > 1,6) = 0,5 - 0,445 = 0,055$

Dans les exemples 1 et 2 nous avons pris pour acquis que les résultats X présentaient une distribution continue. Examinons maintenant des cas où nous aurions une distribution discrète.

Exemple 3

Le poids de 1 000 étudiants est distribué normalement avec une moyenne de 70 kg et un écart-type de 8 kg. Combien ont un poids de 75 kg ou moins? (Chaque poids est mesuré à 1 kg près.)

$z = \dfrac{X - \mu}{\sigma}$ devient $\dfrac{X - 70}{8}$. Comme le poids est enregistré à 1 kg près, un homme ayant un poids de 75 kg a un poids inférieur à 75,5 kg et c'est ce dernier résultat qui est utilisé dans un calcul de z.



$p\ (X < 75,5) = p\ (z < 0,69) = 0,5 + 0,2549 = 0,7549.$

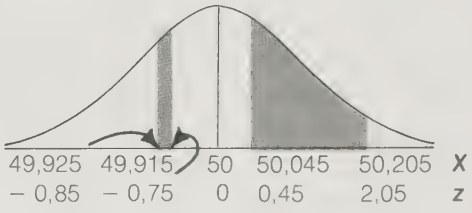
Donc 755 étudiants ont un poids inférieur à 75 kg. 1 000 résultats donnent une distribution discrète, mais à cause du grand nombre, on obtient un histogramme qui ressemble à celui de la section 7.1 au début du chapitre. On ne commet donc pas une grosse erreur en utilisant une distribution continue normale. De plus, on prend comme valeur 75,5 kg au lieu de 75, car dans l'histogramme, 75 se situe au milieu du rectangle et comme la probabilité est l'aire du rectangle, il est préférable de prendre la borne supérieure de celui-ci pour avoir l'aire de ce rectangle au complet.

Exemple 4

Les diamètres des billes de roulement sont distribués normalement avec une moyenne de 50 mm et un écart-type de 0,1 mm. D'une quantité de 10 000 billes de roulement mesurées au centième de mm près, combien auront un diamètre (a) compris entre 50,05 et 50,20 mm (b) égal à 49,92 mm?

$$z = \frac{X - 50{,}000}{0{,}1}$$

	X	z
(a)	50,045	0,45
	50,205	2,05
(b)	49,915	− 0,85
	49,925	− 0,75



(a) $p\ (50{,}045 < X < 50{,}205) = p\ (0{,}45 < z < 2{,}05)$
$= 0{,}480 - 0{,}174 = 0{,}306$
3060 billes de roulement mesureront entre 50,05 et 50,20 mm.

(b) $p\ (X = 49{,}92\text{ mm}) = p\ (49{,}915 < X < 49{,}925\text{ mm})$
$= p\ (-0{,}85 < z < -0{,}75)$
$= 0{,}302 - 0{,}273 = 0{,}029$

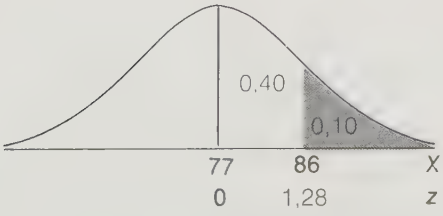
Donc, 290 billes de roulement auront un diamètre égal à 49,92 mm.

Exemple 5

Les résultats d'un examen en statistiques sont distribués normalement avec une moyenne de 77 et un écart-type de 7. Si 10% des résultats les plus hauts doivent avoir un A, quelle est la note la plus basse à laquelle on pourra attribuer un A?

$$0{,}50 - 0{,}10 = 0{,}40 = p\ (0 < z < 1{,}28)$$

$$z = \frac{X - 77}{7}.$$



Si $z = 1{,}28$, $X = 8{,}96 + 77 = 85{,}96$.
La note la plus basse sera alors 86%.

La notation ***A (z)*** sera utilisée pour désigner l'aire sous la courbe normale centrée, réduite entre 0 et z. Dans l'exemple 5 on a $A\ (z) = 0{,}40$ donc $z = 1{,}28$.

7.11 APPROXIMATION D'UNE DISTRIBUTION BINÔMIALE PAR UNE LOI NORMALE

Comme nous l'avons déjà mentionné, au fur et à mesure que le nombre d'essais n devient grand, la distribution binômiale est approximée par une distribution normale. Une distribution normale avec une moyenne $\mu = np$ et un écart-type $\sigma = \sqrt{npq}$ peut être utilisée comme approximation d'une distribution binômiale **si *np* et *nq* sont tous les deux plus grand que 5.**

Exemple 1

Trouver la probabilité d'obtenir au plus 3 faces en 12 lancers d'une pièce de monnaie en utilisant (a) une distribution binômiale et la table 2 appendice C (b) l'approximation par une distribution normale.

(a) $n = 12; p = 0{,}5$

X	$p(X)$
0	0,000
1	0,003
2	0,016
3	0,054
	0,073

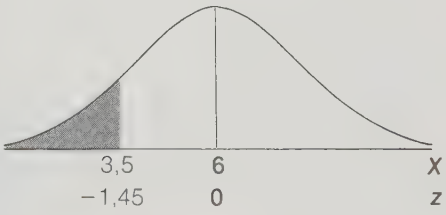
p (au plus 3 faces) = 0,073.

(b) Pour utiliser une distribution normale dans le cas de données discrètes, il est nécessaire d'agir comme si on avait une distribution continue. «On doit donc interpréter au plus 3 faces comme moins de 3,5 faces» (borne supérieure du rectangle dont le centre est 3).

$\mu = np = 12 * 1/2 = 6$

$\sigma = \sqrt{npq} = \sqrt{12 * 1/2 * 1/2} = 1{,}732$

$z = \dfrac{X - 6}{1{,}732}$



Si $X = 3{,}5; z = -2{,}50/1{,}73 = -1{,}45$

$p\ (X < 3{,}5) = p\ (z < -1{,}45) = 0{,}5 - 0{,}426 = 0{,}074$

Exemple 2

Une pièce est lancée 900 fois. Trouver la probabilité que le nombre de faces soit compris entre 400 et 500 fois (400 et 500 inclus).

$n = 900,\ p = 1/2,\ q = 1/2,\ \mu = np = 900 * 1/2 = 450,\ \sigma = \sqrt{npq}$
$= \sqrt{900 * 1/2 * 1/2} = 15,\ z = \dfrac{X - 450}{15}.$

400 à 500 faces devient 399,5 à 500,5 en considérant le cas continu, car $np > 5$ et $nq > 5$.

X	z
399,5	− 3,37
500,5	+ 3,37

$p\ (399{,}5 < X < 500{,}5) = p\ (-3{,}37 < z < 3{,}37) = 2 * 0{,}4996 = 0{,}9992$. Nous sommes à peu près certains d'obtenir face de 400 à 500 fois (inclus) en 900 tentatives.

7.12 VOCABULAIRE ET SYMBOLES

distribution normale
cote z
μ
distribution normale centrée, réduite
$A\ (z)$
σ

7.13 EXERCICES

Dans chaque problème, faire une figure illustrant la situation.

1. Trouver l'aire entre
   (a) $z = 0$ et $z = 1{,}41$
   (b) $z = -0{,}6$ et $z = 0$
   (c) $z = -1{,}23$ et $z = 0{,}53$
   (d) $z = -1{,}23$ et $z = -0{,}53$
   (e) $z = 0{,}46$ et $z = 2{,}31$
   (f) $z = -\infty$ et $z = 1{,}28$

2. Quelle est la probabilité qu'un résultat z choisi au hasard d'une distribution normale soit
   (a) plus grand que 1,65?
   (b) entre 0,46 et 2,59?
   (c) entre $-0{,}46$ et 2,59?
   (d) moins que 0,46?
   (e) plus grand que $-1{,}3$?
   (f) plus petit que $-1{,}17$?
   (g) entre $\pm 1{,}96$?

3. Trouver z si l'aire sous la courbe entre 0 et z est 0,4236.

4. Trouver z si la probabilité qu'un résultat soit compris entre z et 3,42 est 0,2058.

5. Dans une distribution normale, quel pourcentage des résultats z sont compris entre $-0{,}16$ et 1,97?

6. Dans un ensemble de mesures normalement distribuées avec $\mu = 0$ et $\sigma = 1$, quel pourcentage des résultats est
   (a) dans la région $\mu \pm 1{,}2\sigma$?
   (b) à l'extérieur de $\mu \pm 2\ \sigma$?

7. Trouvez z si (a) 95% (b) 98% des résultats d'une distribution normale sont compris entre $-z$ et z.

8. Trouver (a) le 90e centile (b) le 75e centile d'une distribution normale de z.

9. Trouver l'intervalle semi interquartile Q d'une distribution normale z.

10. 20 000 résultats sont normalement distribués avec $\mu = 0$ et $\sigma = 1$. Combien sont (a) plus grands que 1 (b) entre 0,5 et 1,53?

11. Un psychologue fait passer un test de personnalité à 6 étudiants. Voici les résultats obtenus: 28; 34; 32; 31; 30; 31. Il veut écrire un rapport dans lequel il veut normaliser les résultats à une moyenne de 50 et un écart-type de 10. Quels seront les résultats de ce rapport?

12. Si les résultats de l'exercice 11 proviennent d'une population quels seraient alors leur moyenne μ et l'écart-type σ?

13. Des résultats X sont normalement distribués avec $\mu = 30$ et $\sigma = 20$. Si un résultat est choisi au hasard, quelle est la probabilité qu'il soit (a) compris entre 60 et 80 (b) moins que 60 (c) plus grand que 40?

14. 300 étudiants ont une taille moyenne de 1,65 m avec un écart-type de 5 cm. Les tailles sont normalement distribuées et mesurées au cm près.

(a) Combien d'étudiants mesurent entre 1,67 et 1,71 m?
(b) Combien mesurent moins de 1,64 m?
(c) 30% des étudiants mesurent moins de combien?
(d) Combien d'étudiants ont une grandeur qui diffère de la moyenne par plus d'un écart-type?

15. Si une pièce de monnaie est lancée 100 fois, quelle est la probabilité que le nombre de faces soit
(a) Compris entre 40 et 60? (40 et 60 inclus)
(b) Compris entre 41 et 59? (41 et 59 inclus)
(c) Compris entre 43 et 57? (43 et 57 inclus)

16. Un examen «Vrai ou Faux» a 20 questions. Quelle est la probabilité qu'un étudiant ne connaissant pas le sujet et répondant au hasard obtienne 15 bonnes réponses ou plus?

17. Jean a obtenu les résultats suivants en physique au dernier semestre.

Examen	Résultat	Moyenne de la classe	Écart-type de la classe
1	86	85	5
2	90	85	10
3	60	55	5
4	74	72	5

(a) Quel est le résultat moyen de Jean?
(b) Quel est son résultat moyen si le professeur utilise la cote z pour chaque examen puis, redistribue les notes avec une moyenne de 78 et un écart-type de 6?
(c) Lequel de (a) ou (b) donne une meilleure idée du résultat de Jean pour ce cours?

18. 500 étudiants qui suivirent le cours de psychologie I ont obtenu une moyenne de 77 avec un écart-type de 10. Les notes finales doivent être redistribuées de la façon suivante: 10% des meilleures notes ont un A, les 30% suivantes un B, 50% un C, et le dernier 10% un E (échec). Trouver les notes qui permettront de faire cette distribution.

19. Une machine fabrique des boulons dont le diamètre moyen est 2,40 cm avec un écart-type de 0,05 cm. (a) Si les diamètres de 6 000 boulons sont normalement distribués, combien auront un diamètre supérieur à 2,39 cm? (b) Si la machine est

réajustée de façon à ce que le diamètre moyen augmente de 0,06 cm tout en gardant le même écart-type. Quel sera le pourcentage des boulons qui auront un diamètre supérieur à 2,55 cm?

20. Un super-marché est éclairé par des fluorescents dont la vie moyenne est 3 500 heures avec un écart-type de 600 heures. (a) Si les tubes sont allumés 10 heures par jour, 6 jours par semaine et 52 semaines par an, quelle proportion des lumières devra-t-on remplacer après un an? (b) Après combien de semaines devra-t-on remplacer 10% des lumières?

21. Richard prétend qu'il peut reconnaître le vin *A* de trois autres marques du même type. Pendant un mois, on lui fait subir 30 essais et il reconnaît 14 fois le vin *A*. Quelle est la probabilité qu'il identifie le vin *A* 14 fois ou plus s'il ne fait que deviner? Sa prétention est-elle justifiée?

22. Se basant sur l'expérience des années précédentes un manufacturier de machine à laver affirme que la vie moyenne de ses machines est de 5,75 ans avec un écart-type de 2 ans.

(a) Quelle garantie doit-il offrir (période de temps) s'il est disposé à réparer 1% des machines durant cette période. (b) S'il donne une garantie de 2 ans, quel pourcentage des machines devra-t-il réparer avant l'expiration de celles-ci?

23. Mendel a découvert que si l'on sème des pois produits par des grands plants de pois, 75% des plants résultant sont grands et 25% sont nains. Si on sème 100 plants quelle est la probabilité qu'au moins 28 plants soient nains?

24. Une paire de dés est lancée 180 fois. Quelle est la probabilité que les deux dés présentent le même résultat

(a) au moins 40 fois?
(b) entre 30 et 40 fois?
(c) exactement 35 fois?

25. Une distributrice de café est ajustée pour emplir des tasses de 200 cm^3 avec un écart-type de 12 cm^3. Si des tasses de 225 cm^3 sont utilisées quel est le pourcentage des tasses qui déborderont?

26. La moyenne des quotients intellectuels des étudiants admis au collège Beloeil est 118 avec un écart-type de 5. Si 40 étudiants admis ont entre 120 et 125, combien d'étudiants furent admis assumant que les Q.I. sont normalement distribués?

Réponses

1. (a) 0,4207
   (b) 0,2257
   (c) 0,3907 + 0,2019 = 0,5926
   (d) 0,3907 − 0,2019 = 0,1888
   (e) 0,4896 − 0,1772 = 0,3124
   (f) 0,5 + 0,3997 = 0,8997

3. 1,43 ou − 1,43

5. 0,0636 + 0,4756 = 0,5392; 53,9%

7. (a) $0,95/2 = 0,4750$; $z = 1,96$
(b) $0,98/2 = 0,49$; $z = 2,33$

9. $A(z) = 0,25$ donc $z = 0,67 = Q_3$, $Q_1 = -0,67$

$$Q = \frac{0,67 - (-0,67)}{2} = 0,67$$

11. $\overline{X} = 31,0$; $s = 2,0$. Résultats z: $-1,5$; $1,5$; $0,5$; 0; $-0,5$; 0.

Résultats normalisés: 35; 65; 55; 50; 45; 50.

13. $z = \dfrac{X - 30}{20}$. (a) $p(60 < X < 80) = p(1,5 < z < 2,5) = 0,4938 - 0,4332 = 0,06$

X	z
60	1,5
80	2,5
40	0,5

(b) $p(X < 60) = p(z < 1,5) = 0,5 + p(0 < z < 1,5) = 0,5 + 0,4332 = 0,93$

(c) $p(X > 40) = p(z > 0,5) = 0,5 - p(0 < z < 0,5) = 0,5 - 0,1915 = 0,31$

15. $n = 100$, $p = 0,5$, $\mu = np = 50$, $\sigma = \sqrt{100 * 0,5 * 0,5} = 5$;

X	Y
39,5	$-2,10$
60,5	2,10
40,5	$-1,90$
59,5	1,90
42,5	$-1,50$
47,5	1,50

(a) $p(39,5 < X < 60,5) = p(-2,10 < z < 2,10) = 2 * 0,482 = 0,96$

(b) $p(40,5 < X < 59,5) = p(-1,90 < z < 1,90) = 2 * 0,471 = 0,94$

(c) $p(42,5 < X < 47,5) = p(-1,50 < z < 1,50) = 2 * 0,433 = 0,87$.

17. (a) 77,5 (b) Moyenne des résultats $z = (0,2 + 0,5 + 1 + 0,4)/4 = 0,52$ donc $0,52 = \dfrac{X - 78}{6}$, donc $X = 81$

(c) Le résultat (b) donne une bonne idée de sa capacité relative dépendant du choix par le professeur de la moyenne et de l'écart-type (s'il utilise toujours la même moyenne et le même écart-type, la note de Jean peut être comparée aux étudiants des sessions précédentes).

19. (a) $z = \dfrac{2,395 - 2,40}{0,05} = -0,10$; $p(z > -0,10)$
$= 0,5 + 0,0398 = 0,5398$; $0,5398 * 6\,000 = 3239$ boulons

(b) $z = \dfrac{2,555 - 2,46}{0,05} = 1,90$; $p(z > 1,90)$
$= 0,5 - 0,4713 = 0,029$; 2,9% ont plus de 2,55 cm de diamètre.

21. $n = 30$, $p = 1/4$, $\mu = np = 7,5$, $\sigma = \sqrt{npq} = 2,37$, $z = \dfrac{13,5 - 7,5}{2,37} = 2,53$

p (identifier 14 fois ou plus) $= 0,5 - 0,4943 = 0,0057$. La probabilité qu'il devine est moins que 0,6% donc sa prétention est probablement justifiée.

23. p (nain) $= p = 1/4$, $n = 100$, $\mu = 25$, $\sigma = \sqrt{100 * 1/4 * 3/4} =$ 4,33. Si $X = 27{,}5, z = \dfrac{27{,}5 - 25{,}0}{4{,}33} = 0{,}58; p\,(X > 27{,}5) = p\,(z > 0{,}58) = 0{,}5 - 0{,}219 = 0{,}28$

25. $z = \dfrac{X - 200}{12}$; $p\,(X > 225) = p\,(z > 2{,}08)$

$= 0{,}50 - 0{,}48124 = 0{,}01876$; 1,9% des tasses déborderont.

Chapitre 8

Comment choisir un échantillon?

Nous avons maintenant les connaissances nécessaires en probabilité et en distributions de probabilités pour aborder l'inférence statistique. Dans ce chapitre, nous donnerons une vue d'ensemble sur la manière de mener une expérience utilisant des statistiques. Dans la plupart de ces expériences, il faut mener une enquête auprès d'une population en recueillant des données d'une portion de celle-ci, c'est-à-dire, d'un échantillon. Ce chapitre nous donne les grandes lignes permettant de choisir adéquatement un échantillon représentatif de la population.

8.1 PRÉPARATION DE L'EXPÉRIENCE STATISTIQUE

Il y a cinq étapes:

1. On doit bien identifier le sujet de l'étude et la population à laquelle cette étude est destinée.

Un exemple: on se propose d'étudier la relation qui existe entre les fumeurs et le cancer du poumon; l'une des questions qu'on doit se poser est «Qui fume?». Doit-on considérer seulement les fumeurs de cigarettes: les fumeurs de cigares et de pipe doivent-ils être classés avec les non-fumeurs ou ne pas être considérés du tout? On doit également classer les fumeurs de cigarettes selon le nombre de cigarettes fumées par jour et également selon le nombre d'années. La personne fume-t-elle plus de deux paquets par jour? Plus d'un paquet par jour? Veut-on comparer l'incidence du cancer du poumon entre les fumeurs et les non-fumeurs ou seulement entre ceux qui ont une certaine occupation, ou qui habitent une certaine région ou un genre d'environnement? Le nombre de questions est illimité; tout dépend du temps, du budget, et de l'importance de l'enquête que l'on veut mener.

Généralement, on doit restreindre au minimum le nombre de questions à cause de la difficulté de recueillir les données et de les analyser. Il est préférable d'avoir peu de questions, mais de pouvoir en tirer des conclusions valables que d'avoir beaucoup de questions, mais sans signification. Par exemple, compter le nombre d'oiseaux qui survolent le mont St-Hilaire ne nous renseigne pas sur le genre d'oiseaux qui vivent sur le mont St-Anne. Il faut bien planifier son travail avant de commencer la cueillette des données.

2. On doit décider si l'expérience s'adresse à toute la population ou à un échantillon.

Si on veut faire un recensement, les techniques de statistiques descriptives étudiées dans les chapitres précédents sont suffisantes; on peut présenter un tableau des données, tracer un histogramme ou un diagramme en bâtonnets, trouver la moyenne, l'écart-type.

Un exemple: si on veut trouver la grandeur moyenne des élèves de 3e année de l'école primaire Christ-Roy, et que le nombre d'élèves n'est pas trop grand, on peut mesurer la taille de chaque élève (recensement). Pour une grande population, le recensement est ordinairement trop coûteux et prend trop de temps. Un gouvernement peut à l'occasion faire un recensement par le biais des rapports d'impôt pour déterminer le salaire moyen, l'âge moyen des travailleurs, les catégories de travailleurs, les professionnels, etc. Le gouvernement compile annuellement ces statistiques dans le livre Statistiques Canada. Pour la plupart des expériences, on utilisera un échantillon. Des résultats obtenus pour l'échantillon, on généralisera pour la population par le biais de l'inférence statistique.

Comment doit-on choisir l'échantillon? Ce chapitre répond à cette question. Quelle grandeur doit avoir l'échantillon? Cette question est étudiée au chapitre 10. Pour l'instant, on prendra pour acquis que la taille de l'échantillon est connue. Un point est à retenir: l'expérience ainsi que l'échantillon doivent être connus avant la cueillette des données.

3. On doit déterminer les résultats à recueillir, comment on va les recueillir et quels critères serviront à l'interprétation.

Un exemple: on veut déterminer le Q.I. de 40 paires de jumeaux identiques séparés peu après la naissance et élevés dans des milieux différents. On doit décider en premier quel type de test on utilisera. De plus, est-ce la même personne qui fera passer le test à chaque paire de jumeaux? (Le choix d'une même personne assure-t-il l'uniformité des résultats ou celle-ci sera-t-elle influencée par les résultats obtenus par les premières paires de jumeaux?) On doit déterminer les critères qui permettront de comparer les différents milieux et comment obtenir l'information.

Il est essentiel de déterminer avant l'expérience les critères permettant d'interpréter les résultats. Pour vérifier la régularité d'une pièce de monnaie, on décide d'effectuer 100 lancers. Si l'on obtient 50 faces, on affirme que la pièce est régulière, si l'on obtient 3 faces, on est assuré que la pièce n'est pas régulière. Mais si l'on obtient 20, 40, 48 faces, comment peut-on être assuré que la pièce est régulière? Si la pièce est régulière, la probabilité d'obtenir de 41 à 59 faces (inclus) est de 0,95 (exercice 15, section 7.13). Il y a donc une probabilité de 0,05 que le nombre de faces soit inférieur à 41 ou supérieur à 59. Avant de faire l'expérience, on peut affirmer qu'il y a 95% des chances que le nombre de faces soit compris entre 41 et 59 et qu'il y a 5% des chances de se tromper, car par hasard, il peut arriver que même si la pièce est régulière, on obtienne un résultat en dehors de l'intervalle 41-59. Nous reviendrons sur ce sujet plus tard.

Il est à noter que les critères permettant d'analyser les résultats sont connus **avant** la cueillette des données.

4. On doit choisir un échantillon si nécessaire, puis recueillir les données.

C'est la partie la plus longue et la plus laborieuse, d'où la nécessité de bien planifier son travail.

5. On doit analyser les résultats, tirer des conclusions et estimer la précision des résultats.

Il est important de donner la précision des résultats, car sans celle-ci, le rapport est sans signification et n'a aucune valeur scientifique. «Il serait trop facile d'affirmer n'importe quoi, n'importe quand.»

8.2 ÉCHANTILLONS ALÉATOIRES

Supposons que l'on désire connaître la grandeur des habitants d'une petite ville; on décide de choisir un échantillon de 30 personnes. Il est évident qu'on ne choisira pas les 30 premiers enfants qui arrivent à l'école primaire, pas plus que les 30 joueurs de ballon-panier de l'équipe locale. On désire que l'échantillon représente le plus possible la population de la ville, ce qui permettra de généraliser les résultats.

Comment choisir un échantillon pour qu'il soit représentatif? Si la population est homogène, on peut choisir au hasard tout échantillon que l'on désire. Dans le cas de notre petite ville, va-t-on choisir 2 grands hommes, 3 enfants en bas de 2 ans et ainsi de suite? Ceci suppose que l'on connaît très bien les caractéristiques de la population, alors pourquoi choisir un échantillon pour étudier ce que l'on connaît? L'échantillon ne doit pas être formé uniquement de ce qui est considéré comme typique de la population (l'échantillon ne doit pas être subjectif) car alors les résultats peuvent être faussés par les préjugés de celui qui choisit l'échantillon.

Un échantillon ne peut être choisi par commodité. Un exemple: en interviewant les personnes qui circulent dans un centre-d'achat, car dépendant de l'heure et de la journée on risque de ne pas avoir une bonne représentation de la population.

Un échantillon doit être choisi objectivement. Une des façons d'y parvenir est de choisir un échantillon aléatoire; c'est un échantillon choisi de sorte que chaque échantillon de même grandeur pris dans la population a la même probabilité d'être choisi. La théorie de l'inférence étudiée plus loin assume que l'échantillon est aléatoire.

Un échantillon peut théoriquement être choisi avec ou sans remise. En pratique, dans la plupart des cas, on le choisit sans remise. Si on interviewe des étudiants, il serait pour le moins étrange de choisir deux fois le même étudiant! Si on fait des tests d'endurance, on ne choisira pas deux fois le même athlète.

Si un échantillon est choisi sans remise, il n'est pas nécessaire d'écrire la liste de tous les échantillons possibles et d'en choisir un au hasard parmi celle-ci. Un échantillon est aléatoire si chaque membre qu'on lui ajoute a la même probabilité d'être choisi que tous les autres qui restent dans la population. Un exemple: les échantillons de grandeur 2 sans remise de la population A; B; C sont $\{A;B\}$ $\{A;C\}$ $\{B;C\}$. La probabilité de choisir l'un ou l'autre de ces échantillons est 1/3. Si on choisit le premier membre de l'échantillon au hasard et le deuxième membre parmi ceux qui restent après le premier choix, la probabilité de choisir l'échantillon $\{A;B\} = p\,(AB \text{ ou } BA) = p\,(AB) + p\,(BA) = 1/3 * 1/2 + 1/3 * 1/2 = 2 * 1/6 = 1/3$ comme précédemment.

Comment choisir un échantillon aléatoire de 30 étudiants pris dans une population de 400 étudiants? Une méthode consisterait à écrire les noms de chacun des 400 étudiants sur des bouts de papier, de déposer le tout dans un récipient, puis de choisir au hasard 30 noms. Une autre méthode consisterait à donner à chacun un numéro, puis de choisir dans une table de nombres aléatoires 30 nombres. On peut obtenir une table de nombres aléatoires, en procédant de la façon suivante: on écrit les nombres 0 à 9 sur des cartes séparées puis après les avoir bien brassées, on pige une carte, on enregistre le résultat, puis on la remet et on brasse à nouveau. Après 5 piges, on laisse un espace et on continue pour 5 autres piges et ainsi de suite. La table 3 appendice C représente une table de nombres aléatoires.

Pour choisir 30 étudiants sur 400, on donne un numéro de 1 à 400 à chaque étudiant, on choisit un point de départ au hasard dans la table 3, puis dans la direction que l'on veut, on lit les 3 premiers chiffres de chaque groupe de 5, ou encore, on ignore les espaces et on lit successivement par groupe de 3. Si un groupe forme un nombre supérieur à 400, on l'ignore, dans les autres cas, l'étudiant dont le numéro est choisi appartient à l'échantillon.

Exemple 1

Choisir un échantillon de grandeur 12 dans une population de grandeur 80. Par exemple, en démarrant à la 18e ligne et 5e colonne de la table 3, et en lisant horizontalement vers la droite, puis en revenant à la 19e ligne, on obtient:

18 | . . . 62542 30536 14777 72360
19 | 70791 39030 . . .

On donne à chaque membre de la population un nombre compris entre 1 et 80. En débutant avec 62 dans la table, les nombres identifiés par A sont retenus pour l'échantillon.

A	A	A	A	A	A	A	B	C	A
62	54	23	05	36	14	77	77	23	60

A	A	A	D	A
70	79	13	90	30

Le deuxième 77 (marqué B) est rejeté.
Le deuxième 23 (C) est également rejeté.
Le nombre 90 (D) est rejeté, car la population ne dépasse pas 80.

Choisir un échantillon en écrivant les noms ou en attribuant des nombres peut s'avérer dans bien des cas trop long ou tout simplement impossible. Ces méthodes ne peuvent être utilisées si on doit choisir un échantillon de 8,000 votants dans une région métropolitaine de quelques millions d'habitants.

8.3 AUTRES MÉTHODES D'ÉCHANTILLONNAGE

L'échantillonnage stratifié consiste d'abord à subdiviser la population en sous-groupes (ou strates) et à choisir ensuite l'échantillon au sein de chacun d'eux. Généralement, chaque sous-groupe est représenté dans l'échantillon proportionnellement à son importance dans la population. Un exemple: si on veut étudier l'opinion des travailleurs d'une ville et que le dernier recensement de la ville montre que 5% des travailleurs sont des professionnels, 15% des techniciens, 25% des ouvriers spécialisés etc, on pourra choisir un échantillon respectant les mêmes proportions, on peut envisager plus d'un facteur dans ce procédé, par exemple, les groupes d'âge, le sexe, la race, la langue etc... C'est un modèle réduit de la population à étudier.

Échantillonnage par amas. Pour choisir un échantillon de 30 étudiants de 400 étudiants habitant 4 résidences, on peut choisir au hasard un étage de chaque résidence puis, choisir au hasard les 30 étudiants parmi ceux habitant les étages sélectionnés. L'échantillonnage par amas est peu dispendieux, les personnes étant déjà regroupées. Par contre, on perd de la précision si les individus d'un groupement sont trop semblables. Un exemple: si on veut prédire le comportement de la population lors des prochaines élections, il serait non souhaitable de choisir l'échantillonnage parmi un seul groupe en interrogeant les membres du club Kiwanis.

Échantillonnage systématique. Pour choisir un échantillon de 30 étudiants d'un groupe de 400, on peut, par exemple, choisir au hasard un premier nom dans le bottin des étudiants et ensuite choisir chaque 13e étudiant que l'on rencontre dans ce bottin jusqu'à l'obtention des 30 noms requis. On utilise ici 13, car $400/30 \simeq 13$. Cette méthode est simple et fréquemment utilisée.

8.4 VOCABULAIRE

recensement
échantillon aléatoire
échantillonnage systématique
échantillon représentatif
échantillonnage stratifié
échantillonnage par amas

8.5 EXERCICES

1. Parmi les échantillons suivants, lesquels sont aléatoires? S'il ne le sont pas, expliquer pourquoi.

(a) Une association de chimistes veut savoir combien, parmi ceux qui ont obtenu un Ph. D. en chimie depuis 5 ans, ont écrit des articles scientifiques. Elle fait parvenir un questionnaire à tous ceux qui ont obtenu un Ph. D. en chimie depuis 5 ans, et base son étude sur les réponses reçues.

(b) Un hebdomadaire veut faire un sondage concernant l'avortement, en choisissant un abonné de chaque région postale de la liste.

(c) Un fabricant de générateurs d'automobile fait passer un test de qualité à chaque 100^{e} générateur produit.

(d) Un psychologue choisit 30 étudiants en donnant un numéro à chacun des 485 élèves d'une école. Il utilise les trois nombres du milieu de chaque groupement de 5 d'une table de nombres aléatoires. Il rejette chaque nombre déjà choisi et ceux supérieurs à 485.

(e) On choisit un échantillon de 50 personnes prises dans le bottin téléphonique de Montréal. Chaque appel est fait le soir et choisi au hasard jusqu'à ce qu'on obtienne 50 réponses.

(f) Un médecin fait une recherche sur le cancer du poumon; il découvre un médicament pouvant le guérir. Il prescrit ce nouveau médicament à chaque troisième patient qui présente un cancer du poumon gardant les autres pour fin de comparaison.

2. Dans une table de nombres aléatoires, dans quelle proportion aurons-nous le chiffre 4?

3. Quelle est la probabilité que, dans une table de nombres aléatoires, deux chiffres qui se suivent soient identiques.

4. Choisir un échantillon de taille 10, sans remise, d'une population numérotée de 1 à 8432 en partant de la 3^{e} ligne et de la 2^{e} colonne en lisant de gauche à droite dans la table 3 puis, en remontant les lignes précédentes.

5. Choisir sans remise un échantillon de grandeur 4 d'une population de taille 10 en partant de la 10^{e} ligne et de la 5^{e} colonne de la table 3, puis en lisant verticalement vers le bas, puis en conservant le premier chiffre de chaque groupe de 5.

6. Commenter les situations suivantes:

(a) L'office de la consommation demande aux gens qui entrent dans un super-marché s'ils utilisent des «Kleenex».

(b) Un sociologue demande à un échantillon aléatoire de 75 étudiants pris dans l'ensemble des classes du cours d'introduction à la sociologie: «Combien de fois avez-vous volé des objets valant plus de $10?»

7. Choisir 10 nombres (entre 0 et 9) au hasard en utilisant la table 3 appendice C. Que doit-on espérer obtenir comme moyenne et écart-type de ces 10 nombres? Quels sont la moyenne et l'écart-type des 10 nombres effectivement choisis?

8. Un échantillon aléatoire est-il un échantillon représentatif?

RÉPONSES

1. (a) Non-aléatoire. Ceux qui n'ont pas publié seront enclin à ne pas répondre.

(b) Non-aléatoire. Les régions rurales sont surreprésentées par rapport aux villes. Bien que le choix soit aléatoire, les réponses sont volontaires et on favorise ainsi probablement ceux qui ont plus d'instruction.

(c) C'est un échantillon systématique, non-aléatoire. On aurait peut-être de la difficulté à découvrir une défectuosité périodique qui se produirait par exemple à tous les 7 générateurs.

(d) C'est un échantillon aléatoire. Il aura besoin d'une table plus longue que celle donnée.

(e) Non-aléatoire. On obtiendra certainement quelques téléphones d'affaires et on élimine les individus qui ne sont pas habituellement à la maison le soir.

(f) C'est un mauvais échantillon. Ce n'est pas un échantillon systématique. La population n'est pas assez grande. De plus, les patients choisis ne sont pas tous rendus à la même évolution. En outre, ce genre de recherche pose une question d'éthique liée à la technique d'échantillonnage.

3. 1/10 (il y a 10 paires 00, 11, 22, ... 99 et 100 nombres de 2 chiffres).

5. En numérotant la population de 0 à 9 au lieu de 1 à 10, cela permet d'utiliser des nombres de 1 chiffre. Les nombres retenus sont 8, 9, 0, 3.

7. C'est une distribution de probabilité uniforme $p(0) = p(1) = \ldots p(9) = 0{,}1; \mu = 4{,}5; = \sigma \sqrt{8{,}25} = 2{,}87$.

Chapitre 9

Intervalle de confiance pour les moyennes

Nous allons maintenant nous attaquer à un problème important: celui de choisir un échantillon, de l'étudier, de tirer des conclusions sur le comportement de la population d'où l'échantillon est prélevé. C'est l'inférence statistique.

Une partie de ce chapitre est très théorique, mais la distribution de la moyenne échantillonnale sera facilement comprise si on comprend bien comment l'inférence de la moyenne d'une population est faite. Finalement, on pourra estimer la moyenne d'une population connaissant la moyenne d'un échantillon choisi au hasard.

Deux cas peuvent se présenter: (1) σ est connu. (2) σ est inconnu mais n *est grand (au moins 30). Si σ est inconnu et* n *plus petit que 30, nous devrons attendre au chapitre 12 avant d'estimer la moyenne d'une population.*

9.1 ÉCHANTILLONNAGE AVEC ET SANS REMISE

Supposons que des boules numérotées sont placées dans un sac. Si une boule est pigée du sac, on a le choix de la remettre dans celui-ci ou non avant d'en piger une seconde. Si la boule est remise, et le sac est bien agité avant d'en piger une seconde, alors chaque boule peut être repigée. On a un échantillon avec remise. Si la boule n'est pas remise, chaque boule peut être pigée seulement une fois.

Exemple 1

Quatre étudiants (A, B, C, D) passent un examen de 20 questions. Le nombre de bonnes réponses de chacun est donné par le tableau suivant:

Étudiant:	A	B	C	D
Nombre de bonnes réponses:	4	8	12	20

Donner la liste de tous les échantillons de taille 2 qu'on peut choisir avec remise dans cette population.

Il y a $4 * 4 = 16$ échantillons, puisque chaque étudiant choisi comme premier élément peut être associé avec chacun des 4 éléments comme second choix. L'ensemble des 16 échantillons est:

4;4	8;4	12;4	20;4
4;8	8;8	12;8	20;8
4;12	8;12	12;12	20;12
4;20	8;20	12;20	20;20

Exemple 2

Cinq vendeurs (M, N, O, P, Q) ont vendu le 15 juin le nombre de voitures suivant:

Vendeur:	M	N	O	P	Q
Nombre de voitures:	1	2	3	4	4

Donner la liste de tous les échantillons de taille 2 **sans remise** qu'on peut obtenir de cette population.

Il y a ${}_5C_2 = 10$ échantillons

	Échantillons	Résultats
1	M;N	1;2
2	M;O	1;3
3	M;P	1;4
4	M;Q	1;4
5	N;O	2;3
6	N;P	2;4
7	N;Q	2;4
8	O;P	3;4
9	O;Q	3;4
10	P;Q	4;4

9.2 DISTRIBUTION DE LA MOYENNE ÉCHANTILLONNALE

Dans l'exemple 1 de la section précédente considérons les 4 étudiants comme la population ($N = 4$). Tous les échantillons de taille 2 ($n = 2$ pour chaque échantillon) furent pris, avec remise, de cette population. Chacun de ces échantillons a une moyenne. Voici la liste des moyennes de chaque échantillon:

$\bar{X}$:				
	4	6	8	12
	6	8	10	14
	8	10	12	16
	12	14	16	20

Ces 16 moyennes doivent être considérées comme les résultats d'une nouvelle distribution. Nous pouvons les compiler comme on le ferait avec toute autre distribution:

$\bar{X}$	f
4	1
6	2
8	3
10	2
12	3
14	2
16	2
20	1

Cette nouvelle distribution faite à partir des moyennes de tous les échantillons possibles de taille 2 est appelée la **distribution de la moyenne échantillonnale.** C'est un des concepts fondamentaux de l'inférence statistique et nous devrons découvrir ses principales propriétés.

Comme la distribution de la moyenne échantillonnale est une distribution de fréquences, elle a sa propre moyenne et son propre écart-type. Étant donné que les $\bar{X}$ représentent les éléments de la distribution de la moyenne échantillonnale et sont eux-mêmes des moyennes (de chaque échantillon), nous utiliserons la notation $\mu_{\bar{X}}$ (lire mu de $\bar{X}$ barre) pour la moyenne de cette distribution: μ car c'est une moyenne et $\bar{X}$ un indice pour la distinguer de la moyenne de la population. De la même façon, la notation $\sigma_{\bar{X}}$ (lire sigma de $\bar{X}$ barre) est utilisée pour représenter l'écart-type de la distribution de la moyenne échantillonnale. Cet écart-type est aussi appelé **l'erreur-type de la moyenne.**

Calculons la moyenne $\mu_{\bar{X}}$ et l'écart-type $\sigma_{\bar{X}}$ de l'exemple précédent. Pour simplifier les calculs, soustrayons 11 de chaque valeur de $\bar{X}$ avant de multiplier par f. Pour $\sigma_{\bar{X}}$ on utilisera la formule longue où on divisera par N et non $N-1$.

$\bar{X}$	f	$\bar{X}-11$	$f(\bar{X}-11)$	$f(\bar{X}-11)^2$
4	1	−7	− 7	49
6	2	−5	−10	50
8	3	−3	− 9	27
10	2	−1	− 2	2
12	3	1	3	3
14	2	3	6	18
16	2	5	10	50
20	1	9	9	81
	16		0	280

$$\mu_{\bar{X}} = 11 + \frac{0}{16} = 11 \qquad \sigma_{\bar{X}} = \sqrt{280/16} = \sqrt{35/2}$$

On gardera $\sqrt{35/2}$ au lieu de 4,2 puisque la relation avec l'écart-type de la population sera plus facile à établir.

Calculons maintenant μ et σ de la population originale.

X	$X-\mu$	$(X-\mu)^2$
4	−7	49
8	−3	9
12	1	1
20	9	81
44		140

$\mu = 44/4 = 11 \qquad \sigma = \sqrt{140/4} = \sqrt{35}$

Comparons ces résultats avec ceux de la distribution d'échantillonnage des moyennes, on obtient les relations suivantes:

1. $\mu_{\bar{X}} = \mu$. $(11 = 11)$.
2. $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. $(\sqrt{35}/2 = \sqrt{35}/\sqrt{2}.)$

Exemple 1

Considérons les 5 vendeurs de l'exemple 2 de la section précédente, comme la population.

(a) Trouver la moyenne de chacun des 10 échantillons de taille 2 obtenu sans remise puis, former la distribution de l'échantillonnage des moyennes.

(b) Trouver $\mu_{\bar{X}}$ et $\sigma_{\bar{X}}$

(c) Calculer μ et σ

(d) Comparer les résultats obtenus en (b) et (c).

(a) $\bar{X}$: 1,5; 2; 2,5; 2,5; 2,5; 3; 3; 3,5; 3,5; 4. voir le tableau

$\bar{X}$	f
1,5	1
2	1
2,5	3
3	2
3,5	2
4	1
	10

(b) $\bar{X}-2,5$	$f(\bar{X}-2,5)$	$f(\bar{X}-2,5)^2$
− 1,0	− 1,0	1,0
− 0,5	− 0,5	0,25
0	0	0
0,5	1,0	0,5
1,0	2,0	2,0
1,5	1,5	2,25
	3,0	6,0

$$A = 2,5;\ \mu_{\bar{X}} = 2,5 + \frac{3}{10} = 2,8$$

$$\sigma_{\bar{X}} = \sqrt{\frac{6 - 3^2/10}{10}} = \sqrt{\frac{51}{100}}.$$

(c) X: 1; 2; 3; 4; 4; $\mu = 14/5 = 2,8$

$$\Sigma X^2 = 1 + 4 + 9 + 16 + 16 = 46,\ \sigma = \sqrt{\frac{46 - 14^2/5}{5}} = \sqrt{\frac{34}{25}}.$$

(d) $\mu_{\bar{X}} = \mu$, comme auparavant.

Cette fois, il n'est pas vrai que $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. La relation est $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$. Le facteur de correction $\sqrt{\frac{N-n}{N-1}}$ est nécessaire quand on utilise un échantillonnage sans remise. En pratique, ce facteur de correction est rarement utilisé. Si N est plus grand que $20n$, ce facteur est presque 1. Par exemple, si $N = 10\,000$ et $n = 100$ alors $\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{9900}{9999}} = 0{,}995 \simeq 1$. Ce facteur n'est jamais utilisé pour une population infinie. Dans cet exemple, $\sigma_{\bar{x}} = \sqrt{51/100}$, $\sigma = \sqrt{34/25}$, $N = 5$, $n = 2$, et $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{34/25}}{\sqrt{2}}\sqrt{\frac{5-2}{5-1}} = \sqrt{\frac{34 * 3}{25 * 2 * 4}} = \sqrt{\frac{51}{100}} = \sigma_{\bar{x}}$.

Il est maintenant nécessaire de généraliser ce qui vient d'être fait dans ces deux cas et de décrire une procédure. Elle comporte 3 étapes:

1. Tous les échantillons de taille n sont pris d'une certaine population dont la moyenne est μ et l'écart-type est σ, soient S_1, S_2, S_3, ... ces échantillons.
2. La moyenne de chaque échantillon est trouvée. Ces moyennes $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ ..... sont les éléments de la **distribution de la moyenne échantillonnale.**
3. La moyenne $\mu_{\bar{x}}$ et l'écart-type $\sigma_{\bar{x}}$ de la distribution de la moyenne échantillonnale sont calculés. On peut alors établir les relations suivantes:
   (a) $\mu_{\bar{x}} = \mu$.
   (b) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ si la population est infinie ou si les échantillons sont pris avec remise ou si $N > 20n$.

   $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ si la population est finie et les échantillons choisis sans remise **et** $N \leqslant 20n$.

Nous ne démontrerons pas ici les conclusions précédentes. Deux exemples supplémentaires vous aideront à admettre leur validité. Posons $n = 1$, chaque échantillon possède un seul élément de la population et la moyenne est le résultat lui-même. La distribution de la moyenne échantillonnale est donc formée des mêmes résultats que la population. Ainsi $\mu_{\bar{x}} = \mu$ et $\sigma_{\bar{x}} = \sigma$. De plus, si $n = 1$, on aura $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{1}}\sqrt{\frac{N-1}{N-1}}$, et cette relation reste valide pour $n = 1$.

Posons maintenant $n = N$; alors il y a seulement un seul échantillon qui est la population elle-même. La distribution de la moyenne échantillonnale ne possède qu'un seul résultat qui est μ et l'écart-type est 0 (un seul résultat n'a pas de dispersion). On aura alors $\mu_{\bar{x}} = \mu$ et $\sigma_{\bar{x}} = 0$. En posant $n = N$ dans $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N-N}{N-1}} = 0$.

Normalement on s'intéressera aux échantillons de taille plus grande que 1 donc $\frac{\sigma}{\sqrt{n}} < \sigma$ et le facteur de correction $\sqrt{\frac{N-n}{N-1}}$ est plus petit que 1, donc $\sigma_{\bar{x}} < \sigma$.

9.3 THÉORÈME DE LA LIMITE CENTRALE

Jusqu'à maintenant, nous avons discuté de la moyenne et de l'écart-type de la distribution de la moyenne échantillonnale, mais nous n'avons pas parlé de la forme de celle-ci. La forme est précisée par le théorème suivant:

Théorème de la limite centrale

1. La distribution de la moyenne échantillonnale est normalement distribuée si la population est normalement distribuée.
2. Même si la population n'est pas normalement distribuée, la distribution de la moyenne échantillonnale est approchée par une distribution normale si n est grand. Plus la distribution de la population diffère d'une distribution normale plus n devra être grand pour que la distribution de la moyenne échantillonnale s'approche d'une distribution normale. En pratique, si $n \geq 30$, on considérera la distribution d'échantillonnage des moyennes comme une distribution normale.

N.B.: Si la population est finie et que l'on fait un échantillonnage sans remise, pour appliquer ces théorèmes, on doit avoir $N > 20\,n$.

La preuve de ce théorème fait appel à des techniques mathématiques avancées et ne sera pas donnée ici. Il est également difficile de donner des exemples permettant de vérifier ce théorème. En effet, il serait très long de calculer les moyennes de tous les échantillons de taille 30 pris dans une population de 1 000 individus. Nous aurions dans ce cas ${}_{1000}C_{30} \simeq 2{,}5 \times 10^{54}$ échantillons possibles sans remise. Les exemples précédents ne permettent pas d'illustrer, car ils sont de taille 2, donc pas assez grands, car selon le théorème $n \geq 30$. On doit donc admettre ce théorème et faire un acte de foi.

9.4 UTILISATION DE LA DISTRIBUTION DE LA MOYENNE ÉCHANTILLONNALE

Un psychologue veut mesurer le temps de réaction de 30 étudiants à une impulsion électrique et estimer le temps de réaction de tous les individus à une telle impulsion. Un manufacturier veut trouver la vie moyenne en heures de 100 ampoules électriques et estimer la vie moyenne de toutes les ampoules qu'il fabrique. Ce sont deux exemples d'inférence statistique qui seront résolus avec l'aide de la distribution de la moyenne échantillonnale et le théorème de la limite centrale. Pour l'instant limitons-nous à un cas ou nous avons un échantillon aléatoire et où les paramètres de la population sont connus.

Exemple 1

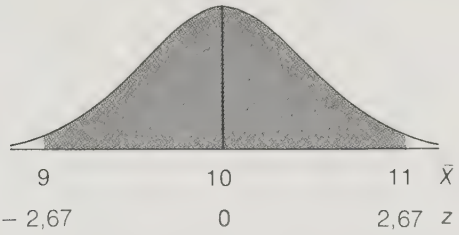
Un échantillon aléatoire de taille 64 est choisi d'une population de taille 100 000 dont la moyenne est 10 et l'écart-type 3. Quelle est la probabilité que la moyenne de l'échantillon soit entre 9 et 11?

La distribution de la moyenne échantillonnale est approximativement normale, car n > 30.

Sa moyenne est $\mu_{\bar{x}} = \mu = 10$, puisque $N = 100\,000$ et est plus grand que $20n = 20 * 64$.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{64}} = \frac{3}{8} = 0{,}375$$

Représentation de la distribution de la moyenne échantillonnale



Noter que l'axe horizontal est noté $\bar{X}$ puisque l'on a une distribution de moyennes.

(a) $p(9 < \bar{X} < 11) = ?$ C'est un problème familier où il faut trouver l'aire sous une courbe normale. Pour transformer en cote z, on soustrait la moyenne de la distribution puis, on divise par son écart-type. Voici les équations.

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma/\sqrt{n}}.$$

Quand $\bar{X} = 11, z = \frac{11-10}{3/\sqrt{64}} = 2{,}67$ quand $\bar{X} = 9, z = \frac{9-10}{3/\sqrt{64}} = -2{,}67; p(9 < \bar{X} < 11)$

$= p\ (-2{,}67 < z < 2{,}67)) = 2 * 0{,}496 = 0{,}99.$

Exemple 2

Tous les élèves de 3e année de la ville de Montréal passent un test d'aptitudes manuelles. Le résultat moyen est 200 avec un écart-type de 20. Quelle est la probabilité qu'un échantillon de 16 élèves choisis au hasard en 3e année à l'école Christ-Roy ait une moyenne supérieure à 205?

●●Dans ce problème, il y a deux pièges, essayez de les découvrir avant de poursuivre la lecture.

Premièrement, l'échantillon n'est pas aléatoire et n'est pas représentatif de la population (tous les élèves de 3e année de la ville de Montréal). Deuxièmement, la taille de l'échantillon $n = 16 < 30$ est telle que la distribution de la moyenne échantillonnale sera normale seulement si la distribution de la population est normale. Or ceci n'est pas explicité dans la donnée du problème, et on ne pourra pas apprécier l'erreur commise en prenant pour acquis que la population est normale alors qu'elle pourrait ne pas l'être.

Exemple 3

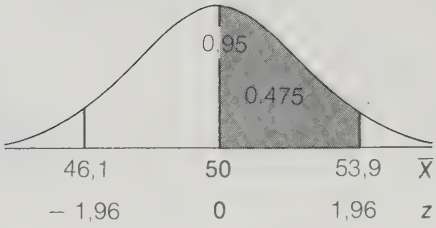
Une population normale a une moyenne de 50 et un écart-type de 10. Dans quel intervalle la moyenne de 95% des échantillons de taille 25 se situera-t-elle? (Posons que l'intervalle a pour milieu 50).

L'aire sous la courbe de l'intervalle cherché est 0,95 donc l'aire à droite de $z = 0$ est 0,475. Si $A(z) = 0{,}475$ alors $z = 1{,}96$

$$\mu = 50,\ \sigma = 10,\ n = 25,\ \sigma_{\bar{X}} = \frac{10}{\sqrt{25}} = 2$$

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - 50}{2}$$

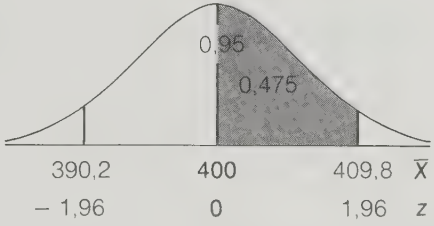
$$\pm 1{,}96 = \frac{\bar{X} - 50}{2}.$$



$\bar{X} = 50 \pm 3{,}92 = 53{,}92$ ou 46,08; 95% des échantillons de taille 25 auront une moyenne comprise entre 46,1 et 53,9.

9.5 INFÉRENCE STATISTIQUE

Nous devrons être vite familiers avec un énoncé du type «Il y a une probabilité de 0,95 que la moyenne d'un échantillon de taille 100, prélevé au hasard d'une population où $\mu = 400$ et $\sigma = 50$, sera comprise entre 390,2 et 409,8».



La principale préoccupation du statisticien est l'étude d'un échantillon puis, par inférence, prévoir le comportement de la population d'où l'échantillon est prélevé. Par exemple, le ministère des Transports veut évaluer le nombre moyen de kilomètres qu'un camion parcourt pendant une semaine. Un économiste veut connaître la proportion des charpentiers qui gagnent plus de $10 000 par année, ou la différence entre les salaires moyens des métallos et des monteurs d'automobiles dans une chaîne de production. Un psychologue veut établir le temps moyen de réaction d'un homme à un certain stimulus ou la proportion des enfants dont les parents n'ont pas fait d'études collégiales. Un médecin étudie le poids moyen des enfants à la naissance, ou la proportion des leucémiques qui survivent pendant 5 ans après avoir débuté un traitement spécial ou la différence des espérances de vie entre les fumeurs et les non-fumeurs. Un politicologue veut connaître la proportion des gens qui voteront pour le parti libéral lors des prochaines élections.

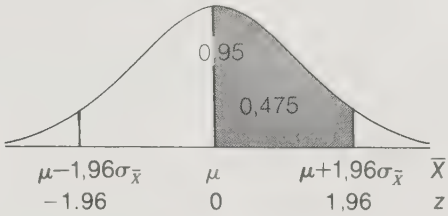
Trois cas peuvent se présenter selon les exemples donnés ci-haut: un premier s'intéresse à la moyenne, un autre à la proportion, un troisième à la différence des moyennes. Dans tous ces cas, le chercheur doit généraliser après avoir examiné un échantillon pris d'une population; il doit faire de l'inférence statistique.

9.6 ESTIMATION DE LA MOYENNE DE LA POPULATION QUAND σ EST CONNU.

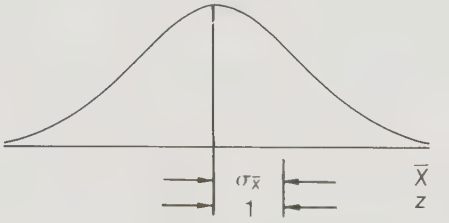
Supposons que la moyenne d'un échantillon aléatoire de taille 36 est 100 et que

l'écart-type de la population (très grande) est 24. Une estimation de la moyenne de la population peut être présentée de deux façons. Par une estimation **ponctuelle** en donnant à μ une valeur numérique («J'estime que μ = ____»), ou par un **intervalle de confiance de la moyenne,** en disant que la moyenne est entre deux valeurs («J'estime que μ est compris entre _____ et _____»). Dans le cas précédent, une bonne estimation ponctuelle pour μ serait de prendre $\mu = \bar{X} = 100$; pour le moment, nous ne pouvons donner la précision de cette estimation, nous y reviendrons plus tard. Deux psychologues peuvent obtenir des estimations ponctuelles différentes d'une même population tout en ayant utilisé des échantillons différents. Nous verrons, après avoir trouvé la précision d'une estimation ponctuelle, qu'en réalité on obtient un intervalle de confiance pour la moyenne.

Sachant qu'une distribution de la moyenne échantillonnale est normale, on peut alors dire que 95% des échantillons auront une moyenne dont la cote z sera comprise entre $-1{,}96$ et $+1{,}96$,



Si μ est inconnu, alors 95% des moyennes d'échantillonnage seront entre $1{,}96\sigma_{\bar{X}}$ unités de part et d'autre de μ. Dans l'exemple précédent, $\sigma = 24$, $n = 36$ et $\sigma_{\bar{X}} = \frac{24}{\sqrt{36}} = 4$ donc $1{,}96\ \sigma_{\bar{X}} = 7{,}84$.



Voici le raisonnement que l'on peut faire. Sur 10 000 échantillons choisis dans cette population, environ 9 500 auront une moyenne se situant à moins de 7,8 unités de la moyenne et 500 auront une moyenne se situant à plus de 7,8 unités de la moyenne, de la population. On ne sait pas dans quel groupe notre échantillon se trouve, mais comme il a été choisi au hasard on peut affirmer **que la probabilité que l'intervalle [92,2; 107,8] contienne la moyenne de la population est de 95%.**

L'intervalle 92,2 à 107,8 est appelé **intervalle de confiance** pour μ.

Voici deux énoncés équivalents:

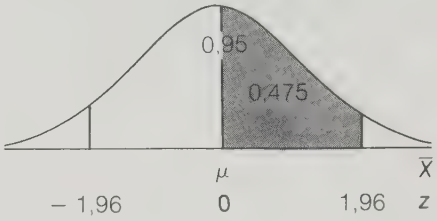
1. J'ai une certitude de 95% que l'intervalle [92,2; 107,8] contienne la moyenne de la population.
2. Un intervalle de confiance pour μ au niveau de 95% est 92,2 à 107,8.

La probabilité que l'intervalle 92,2 à 107,8 contienne la moyenne de la population est appelée **niveau de confiance.** Quand ce niveau est 0,95 par exemple, et que plusieurs échantillons sont choisis et qu'on détermine leur intervalle de confiance, alors 95% de ces intervalles contiendront la moyenne μ.

Exemple 1

La moyenne d'un échantillon aléatoire de taille 36 est 100 et l'écart-type de la population est 24. Trouver un intervalle de confiance pour μ au niveau de 95%. Avec un coefficient de 0,95 on aura:

$z = \pm 1{,}96;\ \sigma = 24,\ n = 36;\ \sigma_{\bar{X}} = \dfrac{24}{\sqrt{36}} = 4.$



$\bar{X} = 100$

$z = \dfrac{\bar{X} - \mu}{\sigma_{\bar{X}}}$

$\pm 1{,}96 = \dfrac{100 - \mu}{4}$

$\mu = 100 \pm 7{,}8 = 92{,}2$ ou $107{,}8$

L'intervalle de confiance pour μ est donc 92,2 à 107,8.

Toujours dessiner une courbe normale représentant la distribution de la moyenne échantillonnale et y inscrire l'aire appropriée, les échelles et les valeurs.

On n'aura pas toujours un niveau de 95%, essayons avec d'autres niveaux.

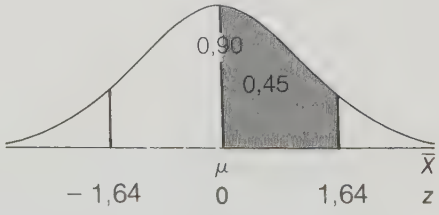
Exemple 2

La moyenne d'un échantillon aléatoire de taille 36 est 100 et l'écart-type d'une grande population est 24. Trouver un intervalle de confiance pour μ au niveau (a) de 90% (b) de 98%.

(a) avec un coefficient de 0,90 $z = \pm 1{,}64$; $\sigma_{\bar{X}} = 4$, $\bar{X} = 100$

$z = \dfrac{\bar{X} - \mu}{\sigma_{\bar{X}}}$

$\pm 1{,}64 = \dfrac{100 - \mu}{4}$



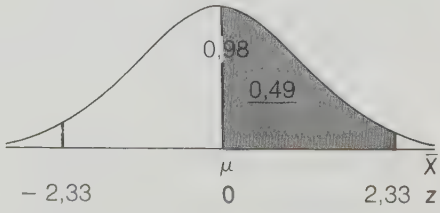
$\mu = 100 \pm 6{,}6 = 93{,}4$ ou $106{,}6$

L'intervalle de confiance pour μ au niveau de 90% est donc [93,4; 106,6].

(b) avec 0,98 $z = \pm 2{,}33$

$\pm 2{,}33 = \dfrac{100 - \mu}{4}$

$\mu = 100 \pm 9{,}32 = 90{,}7$ ou $109{,}3$



L'intervalle de confiance pour μ au niveau de 98% est donc [90,7; 109,3].

Voici un résumé des intervalles de confiance trouvés pour différents niveaux:

Niveau	Intervalle de confiance
90%	93,4 à 106,6
95%	92,2 à 107,8
98%	90,7 à 109,3

Il est à noter qu'au fur et à mesure que le niveau de confiance augmente, l'intervalle de confiance s'élargit. (Plus l'intervalle est large, plus la probabilité que l'intervalle contienne la moyenne de la population est grande.)

Règle générale pour trouver un intervalle de confiance pour la moyenne μ quand σ est connu.

Note: Cette procédure peut être appliquée seulement si la population est normale ou si la taille de l'échantillon est suffisamment grande ($n \geqslant 30$).

1. Trouver la moyenne de l'échantillon $\overline{X}$

2. Trouver l'écart-type de la distribution de la moyenne échantillonnale $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$
N.B.: Si la population est finie et que l'on fait un échantillonnage sans remise, on doit avoir $N > 20\,n$.

3. Déterminer z pour le niveau de confiance donné. Les plus utilisés sont

Niveau de confiance %:	68,3	90	95	98	99	99,7
z:	± 1,00	± 1,64	± 1,96	± 2,33	± 2,58	± 3,00

4. Trouver l'intervalle de confiance en résolvant: $\pm z = \dfrac{\overline{X} - \mu}{\sigma_{\overline{X}}}$ pour μ et $\mu = \overline{X} \pm z\,\sigma_{\overline{X}}$.

Exemple 3

Un échantillon aléatoire de 36 étudiants, choisi sans remise d'une classe de 72 étudiants, a un poids moyen de 60 kg. On sait que l'écart-type de tous les étudiants de la classe est 2 kg. Trouver un intervalle de confiance pour le poids moyen de la classe au niveau de 90%.

1. $\overline{X} = 60$

2. $\sigma_{\overline{X}} = 2/\sqrt{36}\ \sqrt{\dfrac{72-36}{72-1}} = 1/3\ \sqrt{36/71} = 0{,}24$

3. l'échantillon est trop petit.

Il est à espérer que vous n'avez pas appliqué la procédure mécaniquement, mais que vous avez raisonné chacune des étapes. Au niveau de 90% $A(z) = 0{,}45$ et $z = \pm 1{,}64$, on a alors une probabilité de 90% que l'intervalle $60 \pm 1{,}64 * \sigma_{\bar{x}}$ contienne la moyenne μ.

Les exemples précédents ne semblent pas logiques puisque σ est connu et alors pourquoi pas μ? Il y a plusieurs occasions où on peut connaître σ par un moyen rapide, mais pas μ comme dans l'exemple ci-dessous.

Exemple 4

À chaque décade depuis 1900, la grandeur des recrues de l'armée fut déterminée. La grandeur moyenne a continuellement augmenté, mais l'écart-type des grandeurs à chaque décade est demeuré constant à 7 cm. Un échantillon aléatoire de 900 recrues fut choisi en 1970 et leur grandeur moyenne fut de 1,78 m. Trouver un intervalle de confiance au niveau de 95% pour la grandeur moyenne de toutes les recrues de 1970.
On prend pour acquis que $\sigma = 7$ cm en 1970, $\bar{X} = 178$, $n = 900$ $\sigma_{\bar{x}} = 7/\sqrt{900}$
$= 7/30 = 0{,}23$ au niveau de 95%, $z = \pm 1{,}96$
donc $\mu = \bar{X} \pm 1{,}96\ \sigma_{\bar{x}} = 178 \pm 1{,}96 * 0{,}23$
$\mu = 178 \pm 0{,}45$.

On est certain à 95% que l'intervalle [177,55; 148,45] contient la grandeur moyenne de toutes les recrues en 1970.

Dans l'exemple 1, ($\bar{X} = 100$, $n = 36$, $\sigma = 24$) nous avons donné à μ une estimation ponctuelle de 100 et trouvé un intervalle de confiance au niveau de 95%. On ne peut préciser la justesse de ce choix pour une estimation ponctuelle, mais l'intervalle de confiance donne une idée de la précision de ce choix.

9.7 ESTIMATEURS BIAISÉS ET EFFICACES

La prochaine étape consiste à trouver un intervalle de confiance pour μ quand σ est inconnu. Peut-on utiliser l'écart-type de l'échantillon s et procéder comme avant? (Dans certains cas oui; examinons quand et pourquoi.)

À chaque échantillon statistique on peut faire correspondre les paramètres de la population: $\bar{X}$ à μ, s^2 à σ^2, s à σ; il en sera de même avec un échantillon de proportions. On espère pouvoir utiliser $\bar{X}$, s^2, s, p pour estimer μ, σ^2, σ, P etc. Nous venons de voir qu'on peut utiliser la moyenne $\bar{X}$ de l'échantillon pour estimer la moyenne μ de la population. Ceci ne veut pas dire que la moyenne de chaque échantillon va égaler la moyenne de la population, mais que $\mu_{\bar{x}} = \mu$ c'est-à-dire que la moyenne de toutes les moyennes des échantillons donnera μ. De plus la moyenne d'un échantillon pris au hasard ne s'écartera pas trop de la moyenne. On dit alors que $\bar{X}$ est un estimateur efficace de μ.

> ***Définition:*** Un paramètre d'un échantillon est un **estimateur efficace** d'un paramètre de la population, si la moyenne de la distribution d'échatillonnage de ce paramètre est égale au paramètre de la population.

Est-ce que s^2 est un estimateur efficace de σ^2? La réponse est oui. Pour s'en convaincre, revoir la discussion de la section 4.6 et le problème 19 de la section 4.10. Si σ est inconnu et n **est au moins 30,** $\sigma_{\bar{x}}$ est approximé par $\frac{s}{\sqrt{n}}$. Si σ et s sont connus, utiliser $\frac{\sigma}{\sqrt{n}}$ de préférence à $\frac{s}{\sqrt{n}}$. Le cas où σ est inconnu, et n plus petit que 30 sera résolu au chapitre 12.

9.8 INTERVALLE DE CONFIANCE POUR μ QUAND σ EST INCONNU ET $n \geqslant 30$

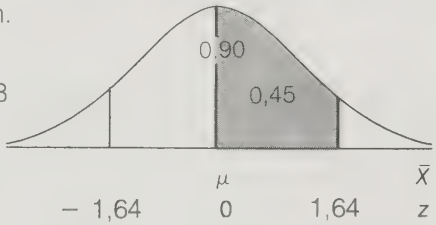
Exemple 1

Trouver un intervalle de confiance au niveau de 90% pour la grandeur de 2 200 étudiants si un échantillon aléatoire de 100 étudiants a une grandeur moyenne de 170 cm et un écart-type de 8 cm.

$\bar{X} = 170, s = 8, n = 100, N = 2\,200$

σ est inconnu donc $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = 8/\sqrt{100} = 0{,}8$

au niveau de 90%, $z = \pm 1{,}64$



$\mu = 170 \pm 1{,}64 * 0{,}8 = 170 \pm 1{,}31$ cm

La probabilité que l'intervalle [168,69; 171,31] contienne la grandeur moyenne des 2 200 étudiants est de 90%.

Exemple 2

Le poids moyen d'un échantillon de 36 étudiants d'université est 72 kg avec un écart-type de 7 kg. On sait que lors de tests prédédents, on a trouvé que l'écart-type de tous les étudiants de cette université était de 8 kg. Trouver le poids moyen de tous les étudiants au niveau de 95%.

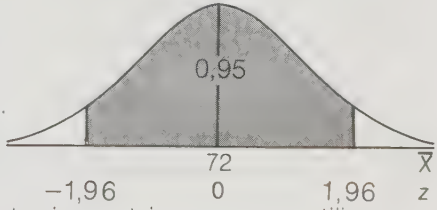
$\bar{X} = 72, n = 36, s = 7, \sigma = 8$, quand s et σ sont connus, on utilise

$\sigma_{\bar{x}} = \sigma / \sqrt{n} = 8/\sqrt{36} = 8/6 = 1{,}33$

à 95% $z = \pm 1{,}96$

$\mu = 72 \pm 1{,}96 * 1{,}33$

(69,39 kg $\leqslant \mu \leqslant$ 74,61 kg) au niveau de 95%.



Répétons la règle donnée précédemment: si σ est inconnu, on utilise $s/\sqrt{n}$ comme approximation de $\sigma_{\bar{x}}$ **seulement si** $n \geqslant 30$.

9.9 VOCABULAIRE ET SYMBOLES

distribution d'échantillonnage des moyennes
$\sigma_{\bar{x}}$
estimation ponctuelle
niveau de confiance
estimateur biaisé.
$\mu_{\bar{x}}$
théorème de la limite centrale
intervalle de confiance
estimateur efficace

9.10 EXERCICES

1. (a) Donner les résultats de la moyenne échantillonnale pour les échantillons de taille 3 pris sans remise d'une population dont les membres A; B; C; D; E ont les résultats suivants 1; 1; 2; 2; 3 respectivement.

(b) Calculer la moyenne et l'écart-type de cette distribution des moyennes.

(c) Calculer μ et σ.

(d) Comparer b et c.

2. (a) Montrer que la probabilité est 0,50 que la moyenne d'un échantillon de taille 40 choisi au hasard d'une population, dont la moyenne est 10 et l'écart-type 6, sera comprise entre $10 - 0{,}67 * 6/\sqrt{40}$ et $10 + 0{,}67 * 6/\sqrt{40}$.

(b) Montrer que la probabilité est 0,50 que la moyenne d'un échantillon de taille n choisi au hasard d'une population dont la moyenne est μ et l'écart-type est σ, sera comprise entre $\mu - 0{,}67\sigma_{\bar{x}}$ et $\mu + 0{,}67\sigma_{\bar{x}}$. Le nombre $0{,}67\sigma_{\bar{x}}$ est appelé l'**erreur probable de la moyenne.**

(c) Trouver l'erreur probable de la moyenne de l'exemple 1, de la section 9.4.

3. Quelle proportion des échantillons aléatoires de taille 64 pris d'une population dont la moyenne est 1 000 et l'écart-type 160, auront une moyenne (a) comprise entre 975 et 1 025 (b) plus grande que 1 040?

4. Le poids de 10 000 billes de roulement est normalement distribué avec une moyenne de 1 g et un écart-type de 0,007 g. Si 100 échantillons de 49 billes sont choisis au hasard, combien d'échantillons auront un poids total (a) entre 48,9 et 49,1 g (b) plus de 49,1 g?

5. Que veut-on dire par l'expression:

(a) **distribution de la médiane échantillonnale;**

(b) **l'écart-type de la médiane.**

6. Des échantillons de taille 3 sont choisis au hasard sans remise de la population 3; 6; 9; 12.

(a) Trouver la moyenne de chacun des 4 échantillons de taille 3, et donner la distribution des moyennes.

(b) Montrer que $\mu_{\bar{x}} = 7{,}5$, $\sigma_{\bar{x}} = \sqrt{5/4}$.

(c) Montrer que $\mu = 7{,}5$ et $\sigma = \sqrt{45/4}$.

(d) Est-ce que $\mu_{\bar{x}} = \mu$? et $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$?

7. (a) Une population de taille 10 est formée des nombres 0;1, . . . 9. Calculer μ et σ.

(b) En prenant le dernier chiffre de chaque nombre dans la 3e colonne de la table des nombres aléatoires, on obtient 50 échantillons de taille 1 de cette population dont la moyenne μ et l'écart-type σ ont été trouvés en (a). Calculer la moyenne et l'écart-type des moyennes de ces 50 échantillons. (La moyenne d'un échantillon de taille 1 est le résultat lui-même.)

8. Quel est l'effet sur l'écart-type de la moyenne de changer la taille de l'échantillon de a) 25 à 100 (b) 25 à 2 500 si la population est très grande?

9. Si $\sigma = 100$, $N = 1\,000$, $n = 100$ et $\bar{X}$ pour un échantillon aléatoire pris avec remise est 480. Trouver un intervalle de confiance pour μ au niveau de 95%.

10. Trouver un intervalle de confiance pour la grandeur de 1 200 étudiants au niveau de 90% si un échantillon aléatoire de 49 étudiants a une moyenne de 170 cm et que l'écart-type pour l'ensemble des étudiants est de 7,6 cm.

11. Un économiste a trouvé que le nombre moyen de km parcourus chaque semaine par un échantillon de 64 camions de livraison est de 400 km avec un écart-type de 100 km. Quel sera le nombre moyen de km parcourus chaque semaine par l'ensemble des camions de livraison au niveau de 95%?

12. Un psychologue a découvert que le temps moyen de réaction à un choc électrique de 18 volts donné à 36 rats est de 0,45 seconde avec un écart-type de 0,06 seconde. Trouver un intervalle de confiance du niveau de 90% pour le temps moyen de réaction de tous les rats au même stimulus.

13. Un criminologue a trouvé que le temps moyen écoulé entre l'arrestation et le procès de 81 hommes choisis au hasard parmi ceux qui furent arrêtés à Montréal pour cause criminelle est de 300 jours avec un écart-type de 30 jours. Trouver un intervalle de confiance pour le temps moyen d'attente pour l'ensemble des causes criminelles à Montréal avec une certitude de 98%.

14. Un agent de location de chambres d'étudiants veut connaître le loyer mensuel moyen de location de son secteur. Il choisit au hasard 30 des 300 chambres de son secteur et trouve un loyer mensuel moyen de $60 avec un écart-type de $10. Quel est le loyer mensuel moyen de l'ensemble des 300 chambres avec une certitude de 95%.

15. Une vérification de 36 ouvriers choisis au hasard a montré que le nombre moyen d'unités assemblées par ouvrier par heure est de 16,2 avec un écart-type de 2,1. Quel est l'intervalle de confiance du nombre moyen d'unités assemblées par heure par l'ensemble des 700 ouvriers de l'usine au niveau de 95%?

16. Un magazine a 2 500 000 abonnés. L'examen de 1 600 abonnés choisis au hasard montre que le revenu annuel moyen est de $12 100 avec un écart-type de $1 000. Trouver, à un niveau de 98%, l'intervalle de confiance du revenu moyen annuel de tous les abonnés.

17. L'examen de 49 des 250 nouveaux procureurs de la ville de Montréal donne une moyenne de 13,7 années d'expérience avec un écart-type de 2,8. Trouver un intervalle de confiance au niveau de 90% pour le nombre moyen d'années d'expérience pour l'ensemble des nouveaux procureurs.

18. Une vérification de 100 accidents d'automobiles, où il y eut des morts, montre que l'âge moyen des conducteurs était de 24,5 ans avec un écart-type de 3,5 ans. Quel est l'âge moyen de tous les conducteurs impliqués dans des accidents mortels au niveau de 95%?

19. Il y a 2 150 super-marchés dans une région métropolitaine. Un échantillon aléatoire de 64 indique qu'il y a en moyenne 35 super-marchés par semaine qui réduisent le prix des aliments pour bébé avec un écart-type de 6 super-marchés. Trouver le nombre moyen de super-marchés par semaine qui feront un rabais sur la nourriture de bébé au niveau de 95%.

20. Un constructeur d'automobiles vérifie au hasard 81 voitures neuves de sa production et trouve une consommation moyenne de 10 ℓ au 100 km avec un écart-type de 1,5 ℓ. Quel est l'intervalle de confiance pour la consommation moyenne de toutes les voitures produites par le constructeur au niveau de 95%?

21. Une étude portant sur 50 familles de 4 personnes de la région de Montréal, choisies au hasard, montre qu'elles dépensent en moyenne $76 par semaine pour se nourrir avec un écart-type de $3. Trouver le coût moyen par semaine pour la nourriture de l'ensemble des familles de 4 personnes au niveau de 96%.

22. Les pneus XYZ ont une vie moyenne de 30 000 km avec un écart-type de 3 km. Le fabricant apporte des modifications et espère ainsi augmenter la vie moyenne sans changer l'écart-type. Il vérifie au hasard 60 pneus modifiés et obtient une vie moyenne de 33 000 km avec un écart-type de 3 200 km. Trouver une estimation au niveau de 90% du kilométrage moyen des pneus modifiés.

23. Un consultant en sécurité vérifie 36 pare-chocs choisis au hasard sur des modèles chevrolets, en frappant un mur de béton à une vitesse de 8 km/h. Le coût moyen des réparations fut de $70 avec un écart-type de $15. Avec quelle certitude peut-il affirmer que le coût moyen des réparations pour un choc de 8 km/h sur des chevrolets sera compris entre $65 et $75?

24. Le salaire moyen de 10 travailleurs de Montréal est de $240 par semaine avec un écart-type de $50. Trouver un intervalle de confiance au niveau de 95% du salaire moyen hebdomadaire pour l'ensemble des travailleurs du Québec. (Attention)●

25. Il y a 120 compagnies de scouts dans la région. 36 sont choisies au hasard et on trouve qu'il y a en moyenne 28 garçons par compagnie avec un écart-type de 3,5 garçons. Quel est le nombre moyen de garçons pour l'ensemble des compagnies de scouts de la région au niveau de 95%.

26. L'examen des dépenses faites par semaine par 64 étudiants d'un collège pour l'achat de livres et disques est en moyenne de $8,10 avec un écart-type de $2,50. Quel est le montant moyen dépensé par semaine par l'ensemble des étudiants de ce collège au niveau de 95%?

27. Trente-six soldats d'une compagnie sont choisis au hasard pour participer à une course à obstacles. Leur temps moyen fut 36 secondes avec un écart-type de 4 secondes. Quel sera le temps moyen pour l'ensemble des soldats de la compagnie au niveau de 90%?

28. 3 600 étudiants au niveau collégial sont choisis au hasard. Il ressort d'une étude qu'ils écoutent la télévision une moyenne 10,5 heures par semaine avec un écart-type de 2 heures. Trouver le temps moyen d'écoute par semaine pour l'ensemble des étudiants du collégial au niveau de 98%.

29. Trouver un intervalle de confiance au niveau de 90% pour l'âge moyen des finissants de 600 étudiants du collégial si un échantillon aléatoire de 100 étudiants a un âge moyen de 18,1 ans avec un écart-type de 0,4 an.

30. Trouver un intervalle de confiance au niveau de 98% pour le nombre de radios par famille si un échantillon aléatoire de 900 familles montre une moyenne 2,5 radios par famille avec un écart-type de 0,33.

31. 120 étudiants suivent un cours de dactylographie. Un échantillon aléatoire de 49 étudiants écrivent à la vitesse moyenne de 26 mots/minute avec un écart-type de 5 mots/minute. Trouver au niveau de 95% la vitesse moyenne de toute la classe.

32. Une vérification faite auprès de 81 des 640 détenus à la prison fédérale indique en moyenne 2,1 incarcérations antérieures par prisonnier avec un écart-type de

0,4. Quel est le nombre moyen d'incarcérations antérieures pour l'ensemble des détenus au niveau de 90%?

33. Une enquête auprès de 25 docteurs en sciences sociales donne une moyenne de 4,1 ans d'études post-graduées avec un écart-type de 1,2 an. Combien d'années post-graduées sont nécessaires pour obtenir un Ph D en sciences sociales au niveau de 90%?

RÉPONSES

1. (a) Échantillons: A; B; C. A; B; D. A; B; E. A; C; D. A; C; E. A; D; E. B; C; D. B; C; E. B; D; E. C; D; E. Résultats: 1; 1; 2. 1; 1; 2. 1; 1; 3. 1; 2; 2. 1; 2; 3. 1; 2; 3. 1; 2; 2. 1; 2; 3. 1; 2; 3. 2; 2; 3. Moyenne (résultats de la distribution d'échantillonnage des moyennes) $\frac{4}{3}$; $\frac{4}{3}$; $\frac{5}{3}$; $\frac{5}{3}$; 2; 2; $\frac{5}{3}$; 2; 2; $\frac{7}{3}$

(b) $\mu_{\bar{x}} = 1{,}8$; $\sigma_{\bar{x}} = \sqrt{7/75} = 0{,}306$

(c) $\mu = 9/5 = 1{,}8$; $\sigma = \sqrt{14/25} = 0{,}748$

(d) $\mu_{\bar{x}} = \mu$.

$$\sqrt{7/75} = \frac{\sqrt{14/25}}{\sqrt{3}} * \sqrt{\frac{5-3}{5-1}} = \sqrt{\frac{14}{75} * \frac{2}{4}} \left(\text{ou } 0{,}306 = \frac{0{,}748}{\sqrt{3}} * \sqrt{\frac{5-3}{5-1}}\right), \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

3. $\sigma_{\bar{X}} = \frac{160}{\sqrt{64}} = 20$

$$z = \frac{\bar{X} - 1000}{20}$$

$\bar{X}$	z
975	− 1,25
1025	+ 1,25
1040	2,00

(a) $p\ (975 < \bar{X} < 1025) = p\ (-1{,}25 < z < 1{,}25) = 2 \quad 0{,}3944 = 0{,}789$; 78,9% des échantillons auront une moyenne comprise entre 975 et 1025.

(b) La distribution de la médiane échantillonnale possède sa propre moyenne et son écart-type.

5. (a) La distribution de toutes les médianes de tous les échantillons de taille n pris dans une population.

(b) La distribution de la médiane échantillonnale possède sa propre moyenne et son écart-type.

7. (a) $\mu = 4{,}5$; $\sigma = 2{,}9$

(b) $\mu_{\bar{x}} = 4{,}5$; $\sigma_{\bar{x}} = 2{,}8$ (approximativement égal car on a seulement 50 échantillons).

9. $\pm 1{,}96 = \frac{480 - \mu}{100/\sqrt{100}}$; 460,4 à 499,6

11. $\pm 1{,}96 = \frac{400 - \mu}{100/\sqrt{64}}$; 375,5 à 424,5 km par semaine

13. $\pm 2{,}33 = \frac{300 - \mu}{30/\sqrt{81}}$; 292,2 à 307,8 jours

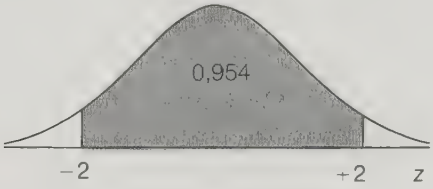
15. Ici, on ne peut faire le problème, car $N \leqslant 20\,n$.

17 Ici, on ne peut faire le problème, car $N \leqslant 20\,n$.

19. $\pm 1{,}96 = \dfrac{35 - \mu}{6/\sqrt{64}}$; 33,5 à 36,5 super-marchés

21. $\pm 2{,}05 = \dfrac{76 - \mu}{3/\sqrt{50}}$; \$76,14 à \$76,86 par semaine

23. $z = \dfrac{70 - 65}{15/\sqrt{36}} = 2{,}00$ où $z = \dfrac{70 - 75}{15/\sqrt{36}} = -2{,}00$; $p\,(-2{,}00 < z < 2{,}00) = 0{,}954$
95,4% de confiance.



25. $\sigma_{\bar{x}} = \dfrac{3{,}5}{\sqrt{36}}\sqrt{\dfrac{120 - 36}{120 - 1}} = 0{,}490$; mais, ici on ne peut faire le problème, car $N \leqslant 20\,n$.

27. $\pm 1{,}64 = \dfrac{36 - \mu}{4/\sqrt{36}}$; 34,9 à 37,1 secondes

29. Ici, on ne peut faire le problème, car $N \leqslant 20\,n$.

31. Ici, on ne peut faire le problème, car $N \leqslant 20\,n$.

33. Attention: $n < 30$ et la distribution d'échantillonnage n'est pas normale, on étudiera ce cas au chapitre 12.

Chapitre 10

Intervalles de confiance de proportions, différences de moyennes.

Nous avons appris à trouver un intervalle de confiance pour la moyenne d'une population après avoir étudié la distribution de la moyenne échantillonnale. Maintenant, nous apprendrons comment trouver trois autres estimations par intervalle de confiance, en étudiant préalablement leur distribution d'échantillonnage. Voici les notations utilisées; P proportion de la population, p pour la proportion de l'échantillon, σ_p pour l'écart-type de la distribution des proportions et ainsi de suite. Il est à remarquer que les distributions d'échantillonnage qu'on utilisera dans ce chapitre sont ***toujours*** *normales; on utilisera la cote z pour trouver l'aire sous la courbe normale. La formule pour trouver z change d'apparence selon qu'on utilise l'une ou l'autre des distributions, mais conserve ses caractéristiques fondamentales.*

Distribution de	*Résultats de la distribution*	*Moyenne*	*Écart-type*	*Cote z*
Moyennes	$\overline{X}$	μ	$\sigma_{\overline{X}}$	$z = \dfrac{\overline{X} - \mu}{\sigma_{\overline{X}}}$
Proportions	p	P	σ_p	$z = \dfrac{p - P}{\sigma_p}$
Différence de moyennes	$\overline{X} - \overline{Y}$	$\mu_X - \mu_Y$	$\sigma_{(\overline{X}-\overline{Y})}$	$z = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\overline{X}-\overline{Y})}}$
Différence de proportions	$p_X - p_Y$	$P_X - P_Y$	$\sigma_{(p_X - p_Y)}$	$z = \dfrac{(p_X - p_Y) - (P_X - P_Y)}{\sigma_{(p_X - p_Y)}}$

Nous constaterons qu'il n'est pas plus difficile de trouver un intervalle de confiance pour une différence de moyennes ou de proportions que pour une moyenne, comme nous l'avons fait au chapitre 9.

10.1 DISTRIBUTION DE LA PROPORTION ÉCHANTILLONNALE

Notre première préoccupation sera de trouver une estimation ponctuelle et un intervalle pour estimer la proportion d'une population à partir de la distribution de la proportion échantillonnale. Voici un cas type: Une enquête est menée pour trouver combien de gens voteront pour le candidat A à la présidence. Un échantillon de 1 000 votants est choisi et on trouve que 520 voteront pour le candidat A. Évaluer le pourcentage de la population qui votera pour le candidat A avec un degré de confiance de 95%.

Il est évident que l'estimation ponctuelle de la proportion est $\frac{520}{1000} = 52\%$. C'est la meilleure valeur pour estimer la proportion de la population, mais quelle certitude avons-nous que c'est une bonne estimation?

Avant de trouver un intervalle de confiance, examinons un peu la distribution de la proportion échantillonnale.

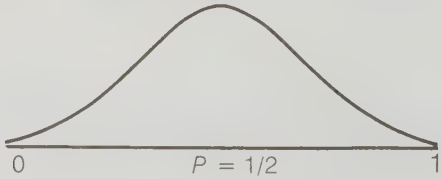
Encore une fois, on trouve tous les échantillons de taille n que l'on peut prendre d'une population de taille N, on trouve la proportion de chaque échantillon et on forme la distribution de ces proportions.

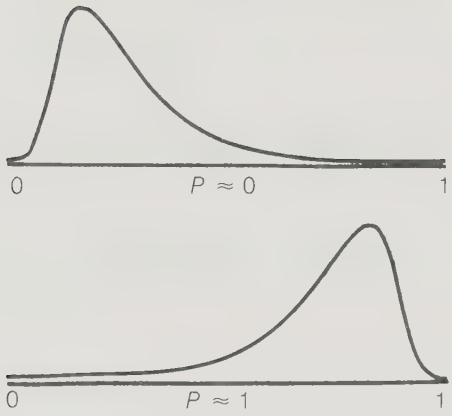
Voici les notations utilisées:

Notation:
- P = *proportion de la population (un paramètre)*
- $Q = 1 - P$
- p = proportion de l'échantillon (une statistique). Attention p n'est pas une probabilité mais une proportion
- μ_p = moyenne de la distribution de la proportion échantillonnale
- σ_p = écart-type de la distribution de la proportion échantillonnale.
- N = taille de la population
- n = taille de l'échantillon

La distribution de la proportion échantillonnale possède les caractéristiques suivantes:

1. $\mu_p = P$.
2. $\sigma_p = \sqrt{\frac{PQ}{n}} * \sqrt{\frac{N-n}{N-1}}$ si la population est finie et l'échantillon est choisi sans remise avec $N \leqslant 20n$; $\sigma_p = \sqrt{\frac{PQ}{n}}$ dans les autres cas. Comme précédemment, dans la plupart des cas, on peut utiliser cette dernière formule car on a souvent $N > 20\,n$.
3. Elle est approximativement normale si nP et nQ sont tous les deux au moins égaux à 15. On doit faire attention puisque p est une proportion qui est toujours comprise entre 0 et 1. Si $P \approx 1/2$, la distribution d'échantillonnage est approximativement normale si $n \geqslant 30$. Mais si p est près de 0 (ou 1) la distribution a tendance à avoir une grande queue vers la droite (ou vers la gauche) et n devra être augmenté pour obtenir une approximation par une loi normale.





Exemple 1

Dans une association, il y a trois membres A; B; C. A votera pour monsieur M comme président, B et C ne voteront pas pour monsieur M. Considérer la distribution de la proportion échantillonnale des échantillons de taille 2 pris **avec remise.**

Soit P la proportion de ceux qui voteront pour monsieur M.
$P = 1/3$, $Q = 2/3$, $N = 3$, $n = 2$.

Échantillon	p
A, A	1
A, B	1/2
A, C	1/2
B, A	1/2
B, B,	0
B, C	0
C, A	1/2
C, B	0
C, C	0

Distribution de la proportion échantillonnale

p	f	fp	fp^2
1	1	1	1
1/2	4	2	1
0	4	0	0
	9	3	2

$$\mu_p = 3/9 = 1/3$$

$$\sigma_p = \sqrt{\frac{2 - 3^2/9}{9}} = \sqrt{\frac{1}{9}} = \frac{1}{3}$$

(a) $\mu_p = P\ (= 1/3)$.

(b) $\sqrt{\frac{PQ}{n}} = \sqrt{\frac{\frac{1}{3} * \frac{2}{3}}{2}} = \sqrt{\frac{1}{9}} = \frac{1}{3}$, donc $\sigma_p = \sqrt{\frac{PQ}{n}}$.

(c) La distribution de la proportion échantillonnale n'est pas normale; $nP = 2/3$ est trop petit.

Examinons maintenant un exemple sans remise.

Exemple 2

Une population est formée de 6 individus D; E; F; G; H; I dont deux (D; E) portent des lunettes et les autres pas.

P = proportion de ceux qui portent des lunettes = 2/6 = 1/3.
Q = proportion de ceux qui ne portent pas de lunettes = $1 - P = 1 - 1/3 = 2/3$.

Considérons tous les échantillons de taille 3 sans remise pris dans la population. Il y en a ${}_6\dot{C}_3 = 20$. Pour chaque échantillon, p = proportion de ceux qui portent des lunettes.

Échantillon	p	f	fp	$p - \mu_p$	$(p - \mu_p)^2$	$f(p - \mu_p)^2$
FGH, FGI, FHI, GHI	0	4	0	$-\frac{1}{3}$	$\frac{1}{9}$	$\frac{4}{9}$
D FG, D FH, D FI, D GH, D GI, D HI, E FG, E FH, E FI, E GH, E GI, E HI	$\frac{1}{3}$	12	4	0	0	0
DE F, DE G, DE H, DE I	$\frac{2}{3}$	4	$\frac{8}{3}$	$\frac{1}{3}$	$\frac{1}{9}$	$\frac{4}{9}$
		20	$\frac{20}{3}$			$\frac{8}{9}$

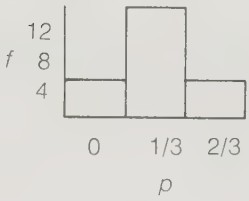
Le rectangle représente la distribution de la proportion échantillonnale.

1. $\mu_p = \dfrac{20/3}{20} = 1/3 = P.$

2. $\sigma_p = \sqrt{\dfrac{\Sigma f(p - \mu_p)^2}{\Sigma f}} = \sqrt{\dfrac{8/9}{20}} = \sqrt{\dfrac{8}{9 * 20}} = \sqrt{\dfrac{2}{45}}$

$$\sigma_p = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{1/3 * 2/3}{3}} * \sqrt{\frac{6-3}{6-1}} = \sqrt{\frac{2}{9 * 3}} * \sqrt{\frac{3}{5}} = \sqrt{\frac{2}{45}}.$$

3. Histogramme de la distribution de la proportion échantillonnale (n est trop petit pour que la distribution soit normale).

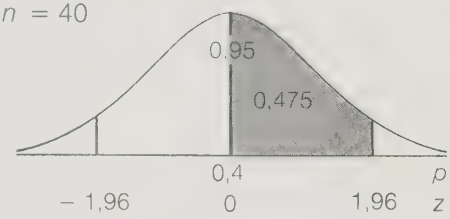


10.2 INTERVALLE DE CONFIANCE POUR UNE PROPORTION

En utilisant la représentation d'une courbe normale, vous pouvez résoudre le problème suivant:

Exemple 1

Si $P = 0{,}4$ et $N = 1\,000$, trouver les valeurs entre lesquelles p sera compris avec une probabilité de 0,95? (Trouver un intervalle de confiance pour p au niveau de 95%.) pour un échantillon de taille $n = 40$



$P = 0{,}4$; $Q = 0{,}6$; $N = 1\,000$; $n = 40$

$$\sigma_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0{,}4 * 0{,}6}{40}} = \sqrt{\frac{0{,}24}{40}} = 0{,}077$$

La probabilité = 0,95 donc $A(z) = 0{,}475$ et $z = \pm 1{,}96$.

Comme d'habitude, la transformation permettant d'obtenir un résultat standardisé est donnée par $z = \dfrac{p - P}{\sigma_p}$ où p joue le rôle de $\overline{X}$ et P joue le rôle de μ,

en transformant on obtient $p = P \pm z\ \sigma_P$

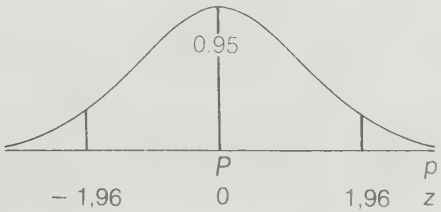
$p = 0{,}4 \pm 1{,}96 * 0{,}077 = 0{,}40 \pm 0{,}15 = 0{,}25$ ou $0{,}55$.

Finalement, $0{,}25 \leqslant p \leqslant 0{,}55$ avec une certitude de 95%.

Exemple 2

Un sondage est fait pour savoir combien de personnes voteront pour le candidat A à la présidence. Un échantillon de 1000 votants est choisi et on trouve que 520 ont l'intention de voter pour le candidat A. Trouver un intervalle de confiance pour la proportion de ceux qui voteront pour le candidat A au niveau de 95%.

Débutons comme si on cherchait un intervalle de confiance pour la moyenne: 95% des résultats d'une distribution normale sont compris entre – 1,96 et 1,96.



La distribution de la proportion échantillonnale est approximativement

distribuée normalement ($np = 1000 * 0{,}52 > 5$ de même que $nq = 1000 * 0{,}48$). Donc il y a une probabilité de 0,95 qu'un échantillon aléatoire de même taille ait une proportion $p = \frac{520}{1000} = 0{,}52$ comprise entre $-1{,}96\ \sigma_p$ et $1{,}96\sigma_p$, par rapport à la moyenne P inconnue.

$$z = \frac{p - P}{\sigma_p}$$

$$\pm 1{,}96 = \frac{0{,}52 - P}{\sigma_p}$$

$$P = 0{,}52 \pm 1{,}96\ \sigma_p$$

Calculons la valeur de σ_p. Ici l'écart-type dépend de P: $\sigma_p = \sqrt{\frac{Pq}{n}}$. peut être utilisé comme approximation de σ_p même si P est inconnu, à la condition que n soit suffisamment grand ($n \geqslant 30$).
Nous aurons:

$$P = 0{,}52 \pm 1{,}96\ \sigma_p = 0{,}52 \pm 1{,}96 * \sqrt{\frac{0{,}52 * 0{,}48}{1000}}$$

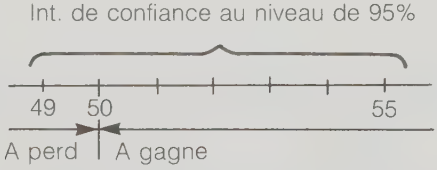
$0{,}52 \pm 1{,}96 * \sqrt{0{,}0002497}$

$0{,}52 \pm 1{,}96 * 0{,}0158$

$0{,}52 \pm 0{,}03$

On peut décrire que $49\% \leqslant P \leqslant 55\%$ avec une certitude de 95%.

Il est à noter que s'il y a seulement deux candidats à la présidence, ce résultat ne peut prédire de vainqueur puisque le 50% nécessaire pour gagner est à l'intérieur de l'intervalle de confiance.



Essayons maintenant le même problème mais cette fois en prenant un échantillon de 4 000 votants où 52% ont l'intention de voter pour le candidat A.

Réponse

Si $n = 4\,000$, $\sigma_p = \sqrt{\frac{0{,}52 * 0{,}48}{4000}} = 0{,}0079$, alors $P = 0{,}52 \pm 1{,}96 * 0{,}0079$ $= 0{,}52 \pm 0{,}015$. L'intervalle de confiance au niveau de 95% pour P est alors $50{,}5\% \leqslant P \leqslant 53{,}5\%$; on a 95% des chances de ne pas se tromper en affirmant que le candidat A va gagner son élection.

Exemple 3

Mille personnes sont invitées à assister à un spectacle. Vous constatez que sur les 90 premières personnes à arriver pour la représentation, 42% sont des femmes. Trouver un intervalle de confiance au niveau de 95% pour le pourcentage de femmes qui assisteront à la représentation.

● Attention, il y a un piège! Lequel? Est-ce un échantillon représentatif? Tout ce que nous avons fait précédemment était basé sur le choix d'un échantillon aléatoire. Peut-être que les femmes ont l'habitude d'arriver un peu plus tard. Il ne faut pas abuser des statistiques!

Exemple 4

Faire le problème 3 en choisissant un échantillon aléatoire de 90 personnes dont 42 sont des femmes.

1. $p = \frac{42}{90}$, $q = \frac{48}{90}$; $n = 90$.

2. $\sigma_p = \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}} =$

$$\sqrt{\frac{42/90 * 48/90}{90}} \sqrt{\frac{1000-90}{1000-1}} = 0{,}0502.$$

3. Ici, on ne peut trouver un intervalle de confiance car la population ne peut être considérée infinie, en effet, on fait un échantillonnage sans remise d'une population finie où $N \leqslant 20n$.

10.3 DISTRIBUTION DE LA DIFFÉRENCE DES MOYENNES ÉCHANTILLONNALES

Considérons deux populations différentes (une donnant le poids de tous les hommes âgés de plus de 17 ans au Canada, l'autre donnant le poids de toutes les femmes de plus de 17 ans au Canada). La première population (X) a une moyenne μ_X et un écart-type σ_X, la seconde (Y) a une moyenne μ_Y et un écart-type σ_Y (ne pas confondre μ_X et $\mu_{\bar{X}}$; dans le 1er cas, l'indice X permet de distinguer deux populations, dans le second $\bar{X}$ indique la moyenne de la distribution de la moyenne échantillonnale). Dans la première population, on choisit un échantillon de taille n_X et on calcule sa moyenne $\bar{X}$. Pour la seconde population, on choisit un échantillon de taille n_Y et on calcule sa moyenne $\bar{Y}$; puis on détermine $\bar{X} - \bar{Y}$. On fait ceci pour toutes les paires d'échantillons qu'on peut choisir indépendamment

dans les deux populations. Les différences $\overline{X} - \overline{Y}$ sont les résultats d'un nouvel ensemble qui devient une **distribution de la différence des moyennes échantillonnales.** Clarifions par un exemple.

Exemple 1

Population *X:* 1; 3; 5.
Population *Y:* 10; 20; 30; 40.

(a) Choisir des échantillons de taille 2 sans remise dans la 1re population et de taille 1 dans la seconde, puis former la distribution de la différence des moyennes échantillonnales.

(b) Calculer μ_X, σ_X^2, $\sigma_{\overline{X}}^2$, μ_Y, σ_Y^2, $\sigma_{\overline{Y}}^2$ (Noter qu'on demande les variances et non les écarts-types.) La moyenne de la distribution de la différence des moyennes échantillonnales $\mu_{(\overline{X}-\overline{Y})}$ et l'écart-type des différences de moyennes $\sigma_{(\overline{X}-\overline{Y})}$. (On écrira les résultats sous forme fractionnaire).

(a)

Échantillons de la Population *X*	Échantillons de la Population *Y*	$\overline{X}$	$\overline{Y}$	$\overline{X} - \overline{Y}$
1; 3	10	2	10	− 8
1; 3	20	2	20	−18
1; 3	30	2	30	−28
1; 3	40	2	40	−38
1; 5	10	3	10	− 7
1; 5	20	3	20	−17
1; 5	30	3	30	−27
1; 5	40	3	40	−37
3; 5	10	4	10	− 6
3; 5	20	4	20	−16
3; 5	30	4	30	−26
3; 5	40	4	40	−36

$n_X = 2, n_Y = 1$. La distribution d'échantillonnage des différences de moyennes pour des échantillons de tailles 2 et 1 pris dans les populations *X* et *Y* respectivement est donnée dans le rectangle ci-dessus à droite.

(b) $\mu_X = 3, \sigma_X^2 = 8/3, n_X = 2, \sigma_{\overline{X}}^2 = \dfrac{8/3}{2} * \dfrac{3-2}{3-1} = 2/3.$ (l'échantillonnage est fait sans remise, donc on doit utiliser le facteur de correction $\dfrac{N-n}{N-1}$; noter que le facteur de correction est élevé au carré car on demande la variance et non l'écart-type).

$$\mu_Y = 25, \sigma_Y^2 = 125, n_Y = 1, \sigma_{\overline{Y}}^2 = \frac{125}{1} * \frac{4-1}{4-1} = 125$$

$$\mu_{\overline{X}} - \mu_{\overline{Y}} = -22, \sigma_{(\overline{X}-\overline{Y})} = \sqrt{377/3}$$

Il serait bon que vous revisiez les calculs précédents. Il faut se rappeler

que pour une distribution d'échantillonnage, le dénominateur de l'écart-type est le nombre de résultats dans la distribution tout comme dans une population.

Vérifions les relations suivantes. :

1. $\mu_{(\bar{X}-\bar{Y})} = \mu_X - \mu_Y$. $(-22 = 3 - 25.)$

2. $\sigma_{(\bar{X}-\bar{Y})} = \sqrt{\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2}$, puisque $\sqrt{377/3} = \sqrt{2/3 + 125.}$

3. n_X et n_Y sont trop petits pour espérer avoir une distribution normale. De plus, $N_X \leqslant 20n_X$ et $N_Y \leqslant 20n_Y$.

Les caractéristiques d'une distribution de la différence des moyennes échantillonnales sont les suivantes:

1. La moyenne de la distribution de la différence des moyennes échantillonnales est égale à la différence des moyennes des populations:

$$\mu_{(\bar{X}-\bar{Y})} = \mu_X - \mu_Y$$

2. L'écart-type de la distribution de la différence des moyennes échantillonnales est donné par:

$$\sigma_{(\bar{X}-\bar{Y})} = \sqrt{\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2}$$

où $\sigma_{\bar{X}}$ est l'écart-type de la moyenne de la 1re population et $\sigma_{\bar{Y}}$ celui de la seconde population.

N.B. Comme $\sigma_{\bar{X}}^2 = \dfrac{\sigma_X^2}{n_X}$ il est inutile d'extraire la racine carrée puisqu'on doit faire la **somme** des variances.

3. La distribution est approximativement normale si $n_X \geqslant 30$ et $n_Y \geqslant 30$ ou encore si les deux populations sont normales. Ne pas oublier de vérifier si $N_X > 20n_X$ et $N_Y > 20n_Y$ dans le cas où l'on a des populations finies et que l'on fait des échantillonnages sans remise pour que la distribution soit normale.

10.4 UTILISATION DE LA DISTRIBUTION DE LA DIFFÉRENCE DES MOYENNES ÉCHANTILLONNALES

Dans l'utilisation de cette distribution on doit s'assurer que les deux échantillons sont choisis indépendamment dans chacune des populations. Si un échantillon de 36 étudiants est choisi pour vérifier le temps de réaction avant et après avoir bu 25 cm³ d'alcool, on a deux séries de mesure, mais on n'a pas deux échantillons indépendants. Les techniques pour faire de l'inférence à partir de deux ensembles de données seront étudiées aux chapitres 12, 15, 16 et 17.

Une autre limitation de cette distribution est due au fait que n_X et n_Y doivent tous les deux être au moins 30 si $s^2{}_X$ et $s^2{}_Y$ sont utilisés pour trouver $\sigma_{\bar{X}}^2$ et $\sigma_{\bar{Y}}^2$. Si cette condition n'est pas respectée, nous utiliserons une autre méthode qui sera présentée au chapitre 12.

Exemple 1

Le Q.I. moyen de 1 200 étudiants du collège St-Joseph est 122 avec un écart-type de 6 alors que 2 000 élèves du collège Bernard ont un Q.I. moyen de 118 avec un écart-type de 5. Quelle est la probabilité que le Q.I. moyen d'un

échantillon aléatoire de 36 étudiants de St-Joseph sera au moins 6 points plus haut que le Q.I. moyen d'un échantillon aléatoire de 49 étudiants du collège Bernard? Essayer de résoudre avant de lire les solutions.

$$\mu_X = 122,\ \sigma_X = 6,\ n_X = 36,\ \sigma_{\bar{X}}^2 = \frac{6^2}{36} = 1$$

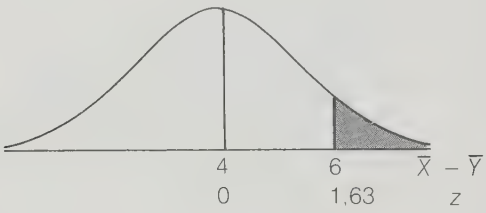
$$\mu_Y = 118,\ \sigma_Y = 5,\ n_Y = 49,\ \sigma_{\bar{Y}}^2 = \frac{5^2}{49} = \frac{25}{49}$$

$$\mu_X - \mu_Y = 122 - 118 = 4$$

$$\sigma_{(\bar{X}-\bar{Y})} = \sqrt{1 + \frac{25}{49}} = 1{,}23$$

$$\overline{X} - \overline{Y} = 6$$

$$z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\bar{X}-\bar{Y})}} = \frac{6 - 4}{1{,}23} = 1{,}63$$



$$p(\overline{X} - \overline{Y} > 6) = p(z > 1{,}63) = 0{,}5 - p\ (0 < z < 1{,}63) = 0{,}5 - 0{,}448 = 0{,}052.$$

Voici un problème intéressant. Si les moyennes de deux échantillons choisis indépendamment et au hasard de deux populations différentes sont connues, et si les écart-types de deux populations ou des échantillons sont connus, quelle affirmation peut-on faire à propos de la différence des moyennes de deux populations d'où on a pris les échantillons?

Exemple 2

Un psychologue vérifie que 36 rats dont le régime alimentaire ne contient pas de vitamine A prennent en moyenne 4,0 minutes pour franchir un labyrinthe avec un écart-type de 0,5 minute; il choisit 50 autres rats, dont le régime est riche en vitamine A, et trouve qu'ils prennent en moyenne 3,2 minutes pour franchir le même labyrinthe avec un écart-type de 0,4 minute. Au niveau de 95% que peut-il affirmer sur la différence des temps moyens pris pour franchir le labyrinthe par les rats sans vitamine A et ceux qui ont un régime riche en vitamine A?

Population *X* (rats sans vitamine A):

$$\overline{X} = 4{,}0;\ n_X = 36;\ \sigma_X \text{ est inconnu; } s_X = 0{,}50;\ \sigma_{\bar{X}}^2 = \frac{s_X^2}{n_X} = \frac{0{,}50^2}{36} = 0{,}0069$$

Population *Y* (rats avec vitamine A):

$$\overline{Y} = 3{,}2;\ n_Y = 50;\ s_Y = 0{,}40;\ \sigma_{\bar{Y}}^2 = \frac{s_Y^2}{n_Y} = \frac{0{,}40^2}{50} = 0{,}0032$$

$$\overline{X} - \overline{Y} = 4{,}0 - 3{,}2 = 0{,}8$$

$$\sigma_{(\bar{X}-\bar{Y})} = \sqrt{\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2} = \sqrt{0{,}0069 + 0{,}0032} = 0{,}10$$

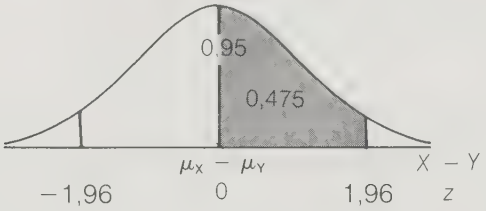
Au niveau de 95%, $z = \pm 1{,}96$

$$z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\bar{X}-\bar{Y})}}$$

$$\pm 1{,}96 = \frac{0{,}8 - (\mu_X - \mu_Y)}{0{,}10}$$

$$\mu_X - \mu_Y = 0{,}8 \pm 0{,}196 = 0{,}60 \text{ ou } 1{,}00$$

L'intervalle de confiance au niveau de 95% pour la différence des temps moyens pris par les rats avec et sans vitamine A est 0,6 à 1,0 minute (36 à 60 secondes).



Exemple 3

Si un échantillon aléatoire de 50 non-fumeurs a une vie moyenne de 76 ans avec un écart-type de 8 ans et qu'un échantillon de 65 fumeurs a une vie moyenne de 68 ans avec un écart-type de 9 ans, trouver, au niveau de 95%, un intervalle de confiance pour la différence des vies moyennes des non-fumeurs et des fumeurs?

Population X (non-fumeurs):

$$n_X = 50,\ \bar{X} = 76,\ s_X = 8,\ \sigma_{\bar{X}}^2 = \frac{8^2}{50} = 1{,}28$$

Population Y (fumeurs):

$$n_Y = 65,\ \bar{Y} = 68,\ s_Y = 9,\ \sigma_{\bar{Y}}^2 = \frac{9^2}{65} = 1{,}25$$

$$\bar{X} - \bar{Y} = 76 - 68 = 8$$

$$\sigma_{(\bar{X}-\bar{Y})} = \sqrt{1{,}28 + 1{,}25} = 1{,}59$$

Au niveau de 95%, $z = \pm 1{,}96$

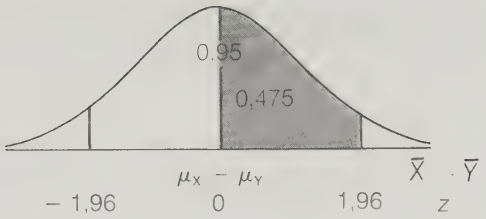
$$z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\bar{X}-\bar{Y})}}$$

$$\pm 1{,}96 = \frac{8 - (\mu_X - \mu_Y)}{1{,}59}$$

$$\pm 3{,}1 = 8 - (\mu_x - \mu_y)$$

$$\mu_x - \mu_y = 8 \pm 3{,}1$$

$$\mu_x - \mu_y = 4{,}9 \quad \text{ou} \quad 11{,}1$$



L'intervalle de confiance au niveau de 95% est $4{,}9 \leq \mu_x - \mu_y \leq 11{,}1$. On peut affirmer avec une certitude de 95% que la vie moyenne des non-fumeurs est plus longue de 4,9 à 11,1 ans que celle des fumeurs.

10.5 DISTRIBUTION DE LA DIFFÉRENCE DES PROPORTIONS ÉCHANTILLONNALES

Considérons deux populations différentes (par exemple des votants de Laval et de Québec). La première population a une proportion P_x qui favorise telle opinion et une proportion $Q_x = 1 - P_x$ qui ne favorise pas cette opinion (à Laval P_x représente la proportion de ceux qui voteront pour le parti A, Q_x la proportion de ceux qui ne voteront pas pour le parti A). Les mêmes proportions pour la seconde population sont P_y et Q_y. On prend un échantillon de taille n_x dans la population X (les votants de Laval) et on trouve p_x de ceux qui favorisent l'opinion (ceux qui votent pour le parti A). On choisit un échantillon **indépendant** de taille n_y de la population Y (votants de Québec) puis on trouve p_y qui favorisent l'opinion (le même parti). On trouve la différence $p_x - p_y$, qui est **un** résultat de la distribution de la différence des proportions échantillonnales. Pour trouver tous les résultats on doit procéder de la même manière avec tous les échantillons possibles de tailles n_x et n_y pris dans chaque population. Ce procédé est identique à celui énoncé à l'article précédent. On peut écrire alors les résultats suivants:

1. La moyenne de la distribution de la différence des proportions échantillonnales $\mu_{(p_x - p_y)}$ est égale à la différence des proportions P_x et P_y des deux populations:

$$\mu_{(p_x - p_y)} = P_x - P_y.$$

2. $\sigma_{(p_x - p_y)}$, l'écart-type de cette distribution est égal à $\sqrt{\sigma_{p_x}^2 + \sigma_{p_y}^2}$, où σ_{p_x} est l'écart-type de la distribution de l'échantillonnage de proportion pour X: $\sigma_{p_x} = \sqrt{\dfrac{P_x Q_x}{n_x}}$; ou $\sqrt{\dfrac{P_x Q_x}{n_x}} \sqrt{\dfrac{N_x - n_x}{N_x - 1}}$

si $N_x \leq 20\, n_x$; $\sigma_{p_y} = \sqrt{\dfrac{P_y Q_y}{n_y}}$ ou $\sqrt{\dfrac{P_y Q_y}{n_y}} \sqrt{\dfrac{N_y - n_y}{N_y - 1}}$ si $N_y \leq 20\, n_y$

Noter que

$$\left(\sqrt{\frac{P_x Q_x}{n_x}}\right)^2 + \left(\sqrt{\frac{P_y Q_y}{n_y}}\right)^2 = \frac{P_x Q_x}{n_x} + \frac{P_y Q_y}{n_y}$$

donc il n'est pas nécessaire de trouver les racines carrées. Finalement:

$$\sigma_{(p_x - p_y)} = \sqrt{\frac{P_x Q_x}{n_x} + \frac{P_y Q_y}{n_y}}$$

3. Cette distribution est approximativement normale si n_X et $n_Y \geqslant 30$. Ne pas oublier de vérifier si $N_X > 20n_X$ et $N_Y > 20n_Y$ dans le cas où on a des populations finies et que l'on fait des échantillonnages sans remise pour que cette distribution soit normale.

Exemple 1

Il y a trois votants A; B; C dans la rue des Lilas; A votera pour M comme maire, B et C ne voteront pas pour M. Dans la rue des Pinsons il y a également trois votants D; E; F; D et E voteront pour M, F non. Considérer la distribution d'échantillonnage de différences de proportions, si chaque échantillon est formé de 2 votants pris sur chaque rue.

$$P_X = 1/3,\ Q_X = 2/3,\ N_X = 3,\ n_X = 2,\ \sigma_{p_X} = \sqrt{\frac{P_X Q_X}{n_X}}\sqrt{\frac{N_X - n_X}{N_X - 1}}$$

$$= \sqrt{\frac{1/3 * 2/3}{2}}\sqrt{\frac{3-2}{3-1}} = \sqrt{\frac{1}{18}}$$

$$P_Y = 2/3,\ Q_Y = 1/3,\ N_Y = 3,\ n_Y = 2,\ \sigma_{p_Y} = \sqrt{\frac{P_Y Q_Y}{n_Y}}\sqrt{\frac{N_Y - n_Y}{N_Y - 1}}$$

$$= \sqrt{\frac{2/3 * 1/3}{2}}\sqrt{\frac{3-2}{3-1}} = \sqrt{\frac{1}{18}}$$

Échantillons		Proportions pour M.		
rue des Lilas	rue des Pinsons	des Lilas p_1	des Pinsons p_2	$p_1 - p_2$
AB	DE	1/2	1	−1/2
AB	DF	1/2	1/2	0
AB	EF	1/2	1/2	0
AC	DE	1/2	1	−1/2
AC	DF	1/2	1/2	0
AC	EF	1/2	1/2	0
BC	DE	0	1	−1
BC	DF	0	1/2	−1/2
BC	EF	0	1/2	−1/2

La distribution de la différence des proportions échantillonnales est donnée dans le rectangle ci-dessus. On peut la réécrire plus simplement, comme ci-dessous.

$p_X - p_Y$	f	$f(p_X - p_Y)$	$f(p_X - p_Y)^2$
−1	1	−1	1
−1/2	4	−2	1
0	4	0	0
	9	−3	2

1. $\mu_{(p_X - p_Y)} = \frac{-3}{9} = \frac{-1}{3}$ $P_X - P_Y = \frac{1}{3} - \frac{2}{3} = -\frac{1}{3}$; $\mu_{(p_X - p_Y)} = P_X - P_Y$.

2. $\sigma_{(p_X - p_Y)} = \sqrt{\frac{2 - (-3)^2/9}{9}} = \sqrt{2/9 - 1/9} = 1/3$;
$\sqrt{\sigma_{p_X}^2 + \sigma_{p_Y}^2} = \sqrt{1/18 + 1/18} = 1/3$; $\sigma_{(p_X - p_Y)} = \sqrt{\sigma_{p_X}^2 + \sigma_{p_Y}^2}$.

3. Comme précédemment n_X et n_Y sont trop petits pour que la distribution soit normale. De plus, $N_X \leq 20n_X$ et $N_Y \leq 20n_Y$.

Exemple 2

Deux groupes sont formés, chacun de 100 malades qui ont la tuberculose. Un nouveau médicament est donné au premier groupe mais pas au second (groupe témoin). On obtient 75 guérisons dans le premier groupe et seulement 60 dans le second. Trouver au niveau de 95% un intervalle de confiance pour la différence de proportions de tous les malades qui ont la tuberculose et qui guérissent.

$$p_X = 0{,}75 \quad q_X = 0{,}25 \quad n_X = 100, \quad \sigma_{p_X}^2 = \frac{p_X q_X}{n_X} = \frac{0{,}75 * 0{,}25}{100} = 0{,}001875$$

$$p_Y = 0{,}60 \quad q_Y = 0{,}40 \quad n_Y = 100, \quad \sigma_{p_Y}^2 = \frac{p_Y q_Y}{n_Y} = \frac{0{,}60 * 0{,}40}{100} = 0{,}002400$$

$$p_X - p_Y = 0{,}75 - 0{,}60 = 0{,}15$$

$$\sigma_{(p_X - p_Y)} = \sqrt{\sigma_{p_X}^2 + \sigma_{p_Y}^2} = \sqrt{0{,}001875 + 0{,}00240} = 0{,}065$$

Au niveau de 95%, $z = \pm 1{,}96$.

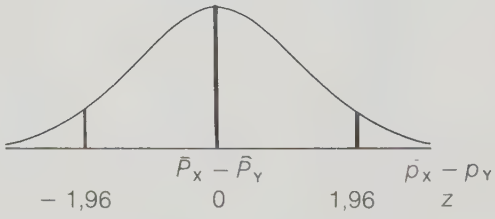
$$\pm z = \frac{(p_X - p_Y) - (P_X - P_Y)}{\sigma_{(p_X - p_Y)}}$$

$$\pm 1{,}96 = \frac{0{,}15 - (P_X - P_Y)}{0{,}065}$$

$$\pm 0{,}13 = 0{,}15 - (P_X - P_Y)$$

$$P_X - P_Y = 0{,}15 \pm 0{,}13$$

$$P_X - P_Y = 0{,}02 \text{ ou } 0{,}28$$



1. Il y a une probabilité de 0,95 que l'intervalle [0,02; 0,28] contienne la différence des proportions entre les malades ayant pris le nouveau médicament et qui ont guéri et ceux qui ne l'ont pas pris.

2. Il y a une probabilité de 0,95 que entre 2 et 28% plus de malades guérissent en prenant le nouveau médicament.

3. L'intervalle de confiance au niveau de 95% pour la différence de proportions des malades qui ont pris le nouveau médicament et qui ont guéri et ceux qui ne l'ont pas pris est entre 2% et 28%.

10.6 RÉSUMÉ DES FORMULES UTILISÉES DANS L'INFÉRENCE STATISTIQUE

Inférence avec la	Formule	Écart-type †
Moyenne		
si σ est connu	$z = \dfrac{\overline{X} - \mu}{\sigma_{\overline{X}}}$	$\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$
si σ est inconnu	$z = \dfrac{\overline{X} - \mu}{\sigma_{\overline{X}}}$	$\sigma_{\overline{X}} = \dfrac{s}{\sqrt{n}}$ si $n \geq 30$
Proportion		
si P est connu	$z = \dfrac{p - P}{\sigma_p}$	$\sigma_p = \sqrt{\dfrac{PQ}{n}}$
si P est inconnu	$z = \dfrac{p - P}{\sigma_p}$	$\sigma_p = \sqrt{\dfrac{pq}{n}}$ si $n \geq 30$
Différences de Moyennes	$z = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\overline{X} - \overline{Y})}}$	$\sigma_{(\overline{X} - \overline{Y})} = \sqrt{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2}$
Différences de Proportions	$z = \dfrac{(p_X - p_Y) - (P_X - P_Y)}{\sigma_{(p_X - p_Y)}}$	$\sigma_{(p_X - p_Y)} = \sqrt{\sigma_{p_X}^2 + \sigma_{p_Y}^2}$

10.7 TAILLE DE L'ÉCHANTILLON

Dans les problèmes qu'on a résolus jusqu'à maintenant, la taille de l'échantillon était donnée. En pratique, on réalise que plus la taille de l'échantillon est grande, plus la précision augmente. Si toute la population est interrogée, on n'a plus besoin d'inférence statistique. Quelle taille d'échantillon est-il nécessaire de prendre pour obtenir un certain degré de précision? C'est une question fondamentale pour le statisticien. Combien de personnes devra prendre un politicologue, s'il veut estimer la proportion de ceux qui voteront pour le candidat X, avec une précision de 1% et un degré de confiance de 98%? Combien de rats devra choisir un psychologue, s'il veut connaître leur temps de réaction moyen à 0,6 seconde près et avec une certitude de 95%?

Il est à noter que dans chaque cas, on mentionne la précision du résultat (avec 1%, à 0,6 seconde près). Avant de déterminer la taille de l'échantillon, on doit établir le degré de précision que l'on veut. Même après avoir pris toutes précautions, on ne peut garantir le degré de précision. Au niveau de 98% ou de 95%, il peut arriver que sur une centaine d'échantillons que l'on peut choisir il y en ait 2 ou 5 qui possèdent un intervalle de confiance qui ne contienne pas le paramètre de la population; dans ce cas, le degré de précision consiste à mesurer toute la population.

Nous étudierons maintenant comment déterminer la taille d'un échantillon pour estimer la moyenne, la proportion, les différences de moyennes et les différences de proportions de populations. Pour chacun des quatre cas, nous avons la démarche suivante:

1. Déterminer le degré de précision, ou ce qui revient au même, déterminer l'erreur E que l'on admet.
2. Écrire une équation dans laquelle on a la taille n et l'erreur permise E. Cette équation devra contenir le paramètre de la population que l'on veut estimer.
3. Résoudre cette équation pour n.

10.8 TAILLE DE L'ÉCHANTILLON POUR ESTIMER LA MOYENNE D'UNE POPULATION

Si on mesure la ligne suivante à un cm près, _____ on peut dire qu'elle mesure 5 cm ± 0,5 cm. Ceci veut dire que sa vraie longueur est comprise entre 4,5 et 5,5 cm ou encore, que sa vraie longueur est 5 cm avec une erreur possible de 0,5 cm.

De façon similaire, on a trouvé les limites de confiance pour μ, par exemple: en écrivant $\mu = 70 \pm 0{,}8$; la moyenne de la population est 70 avec une erreur possible de 0,8 dans chaque direction. En général, $\mu = \overline{X} \pm z\sigma_{\overline{X}} = \overline{X} \pm z\dfrac{\sigma}{\sqrt{n}}$ où z a deux valeurs possibles (une positive et l'autre négative) on peut alors écrire $\mu = \overline{X} \pm |z|\dfrac{\sigma}{\sqrt{n}}$. L'erreur possible pour μ est alors $|z|\dfrac{\sigma}{\sqrt{n}}$.

Un chercheur ou un analyste décide et fixe l'erreur maximum E qu'il juge acceptable puis résoud l'équation, $|z| * \sigma_{\overline{X}} = E$ pour trouver la taille de l'échantillon nécessaire. Comme $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, on obtient:

$$|z|\frac{\sigma}{\sqrt{n}} = E$$

$$z^2\frac{\sigma^2}{n} = E^2$$

$$n = \frac{Z^2\sigma^2}{E^2}$$

n doit être un entier et doit être complété selon le cas (si $n = 44{,}2$ on prendra alors $n = 45$).

Exemple 1

Un manufacturier d'ampoules électriques sait que l'écart-type de la vie moyenne de ses ampoules électriques est de 100 heures. Quelle taille d'échantillon devra-t-il prendre s'il veut être certain à 95% que l'erreur dans son estimation de la vie moyenne est inférieure à 10 heures?

$$n = \frac{z^2\sigma^2}{E^2} = \frac{(1{,}96)^2 100^2}{10^2} = 384{,}2;\ n = 385$$

Il semble que le nombre d'ampoules électriques à vérifier est grand. Il ne faut pas oublier que ceci dépend de l'erreur (E − 10 h) qui est passablement petite. Si la vie moyenne est de 2 000 heures, 10 heures représentent une erreur de 0,5%. Si on avait choisi une erreur deux fois plus grande (20 h), la taille de l'échantillon aurait été quatre fois plus petite soit 97 ampoules électriques.

Exemple 2

Notre manufacturier fabrique toujours des ampoules électriques, mais ne connaît pas l'écart-type de la vie moyenne. Quelle taille d'échantillon doit-on prendre si on veut être certain à 95% que l'erreur dans l'estimation de la vie moyenne de toutes les ampoules électriques soit inférieure à $0{,}1\sigma$?

$$|z|\frac{\sigma}{\sqrt{n}} = 0{,}1\sigma = E$$

$$n = \frac{z^2\sigma^2}{E^2} = \frac{(1{,}96)^2\sigma^2}{(0{,}1)^2\,\sigma^2} = \frac{3{,}842\sigma^2}{0{,}01\sigma^2} = 384{,}2;\ n = 385$$

Si σ est inconnu, la taille de l'échantillon peut être déterminée si l'erreur est donnée en unités d'écart-type (en fonction de σ). La difficulté provient du fait qu'on ne connaît pas l'erreur en unités d'écart-type avant d'avoir choisi un échantillon puis déterminé «s» alors $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$, ainsi que E peut être évalué; $\sigma \approx s$ et $E = 0{,}1\ \sigma$ peut être calculé.

Exemple 3

Reprenons le cas précédent, avec σ inconnu. Quelle taille d'échantillon doit-on prendre pour être certain à 95% que l'erreur dans le calcul de la vie moyenne sera moins de 10 heures?

$$n = \frac{(1{,}96)^2\,\sigma^2}{10^2}$$

On ne peut procéder comme précédemment car il y a ici deux inconnus σ et n. Que devons-nous faire? Si on peut estimer σ en se basant sur des expériences précédentes, on pourra déterminer n. Sinon, on devra mener une expérience pilote en prenant, par exemple, un échantillon de 30 ampoules électriques, déterminer s puis s'en servir comme approximation de σ. Supposons que les 30 ampoules électriques sont testées et ont un écart-type de 90 heures, alors

$$n \approx \frac{(1{,}96)^2\ 90^2}{10^2} = 311{,}2;\ n = 312$$

Il est à noter qu'on ne se sert pas de X. On a déjà vérifié 30 ampoules électriques, il en reste donc 282 autres à vérifier. Si on n'a pas le temps ou l'argent pour poursuivre l'expérience, on ne pourra atteindre la précision demandée en diminuant la taille de l'échantillon.

10.9 TAILLE DE L'ÉCHANTILLON AVEC DES PROPORTIONS

En utilisant le même raisonnement que pour les moyennes, l'équation $\pm z = \frac{p - P}{\sigma_p}$ peut être résolue pour P: $P = p \pm z\sigma_p$, l'erreur est le terme $|z|\sigma_p$, on a donc à

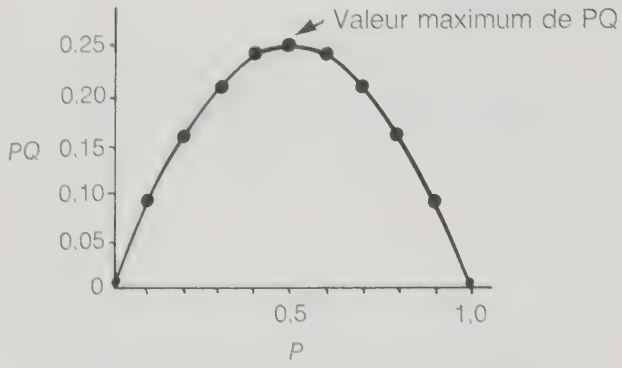
résoudre l'équation $|z|\sqrt{\frac{PQ}{n}} = E$ pour n;

$$n = \frac{z^2 PQ}{E^2}$$

Encore ici, on a un paramètre de la population P qui est connu seulement après avoir complété l'expérience. On peut faire trois choses:

1. Si on peut évaluer l'ordre de grandeur de P en se basant sur des expériences precedentes, on utilisera la valeur maximum de PQ. Si P est près de 1/2 on prendra $PQ = 1/4$. (Si on trouve que P varie entre 0 et 0,3, on posera $P = 0,3$ et $Q = 0,7$, $PQ = 0,21$.)
2. On peut mener une enquête rapide et déterminer p de l'échantillon puis utiliser pq comme approximation de PQ.
3. La valeur maximum de PQ est 1/4, sachant que $PQ = P(1 - P)$ traçons le graphique de cette équation du second degré.

P	PQ
0	0,00
0,1	0,09
0,2	0,16
0,3	0,21
0,4	0,24
0,5	0,25 = 1/4
0,6	0,24
0,7	0,21
0,8	0,16
0,9	0,09
1,0	0,00



Si $PQ \leq 1/4$ alors $n = \frac{z^2 PQ}{E^2} \leq \frac{z^2}{4E^2}$; donc

$$n = \frac{z^2}{4E^2} \text{ est une valeur maximum}$$

Si la valeur de p est trop eloignee de 0,5 la valeur de n sera trop grande; il est alors preferable de prendre l'une ou l'autre des deux premières methodes. L'exemple suivant illustre bien cette situation.

Exemple 1

Trouver la taille d'un échantillon de votants si on veut trouver le pourcentage de ceux qui appuieront le parti A avec une certitude de 95% et une erreur inférieure à 1%.

Au niveau de 95%, $z = \pm 1,96$; $E = 0,01$ $n = \frac{(1,96)^2}{4\,(0,01)^2} = 9\,604$. Si on avait mene une enquête auprès de 200 votants et que 10% aient eu l'intention d'appuyer le parti A, nous aurions pour le même degré de confiance

$$PQ \approx pq = 0{,}1 * 0{,}9 = 0{,}09$$

$$n = \frac{(1{,}96)^2 * 0{,}09}{(0{,}01)^2} = 3\,458$$

Un échantillon plus petit tout en ayant une erreur inférieure à 1%.

10.10★ TAILLE DE L'ÉCHANTILLON POUR UNE DIFFÉRENCE DE MOYENNES

Supposons que les deux échantillons ont la même taille $n_X = n_Y\ (=n)$; $z = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\overline{X}-\overline{Y})}}$, ou $\mu_X - \mu_Y = (\overline{X} - \overline{Y}) \pm z\sigma_{(\overline{X}-\overline{Y})}$, l'erreur est donnée par $|z| * \sigma_{(\overline{X}-\overline{Y})}$,

$$|z|\sigma_{(\overline{X}-\overline{Y})} = z\sqrt{\sigma_{\overline{X}}^2 + \sigma_{\overline{Y}}^2} = z\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}} = \frac{z\sqrt{\sigma_X^2 + \sigma_Y^2}}{\sqrt{n}} = E,$$

$$\text{Puis } n = \frac{z^2(\sigma_X^2 + \sigma_Y^2)}{E^2}$$

Exemple 1

Un économiste veut connaître la différence des salaires entre des professeurs du système public et du secteur privé. Se basant sur des études antérieures, il sait que les écarts-types sont \$900 et \$1 200 respectivement. Quelle sera la taille de son échantillon, s'il veut déterminer la différence à \$300 près avec une certitude de 98%?

Au niveau de 98%, $z = \pm 2{,}33$.

$$n = \frac{(2{,}33)^2(900^2 + 1200^2)}{300^2}$$

$$= \frac{(2{,}33)^2(3^2 * 300^2 + 4^2 * 300^2)}{300^2}$$

$$= (2{,}33)^2(9 + 16) = 135{,}7;\ n = 136$$

Si σ_X et σ_Y sont inconnus, comme précédemment, on devra mener une enquête pilote par exemple, auprès de 30 individus (ou plus) pour évaluer σ_X et σ_Y. Ce choix de 30 ou plus est important pour s'assurer que la distribution sera normale.

10.11★ TAILLE DE L'ÉCHANTILLON POUR UNE DIFFÉRENCE DE PROPORTIONS

Supposons encore une fois que les deux échantillons ont la même taille $n_X = n_Y = n$;

$$z = \frac{(p_X - p_Y) - (P_X - P_Y)}{\sigma_{(p_X - p_Y)}}, \text{ ou } P_X - P_Y = (p_X - p_Y) \pm z\sigma_{(p_X - p_Y)};$$

$$|z|\sigma_{(p_X - p_Y)} < E.$$

L'erreur est donnée par: $|z|\sigma_{(p_X - p_Y)}$.

$$z\sqrt{\frac{P_X Q_X}{n} + \frac{P_Y Q_Y}{n}} = E, \text{ ou}$$

$$n = \frac{z^2 (P_X Q_X + P_Y Q_Y)}{E^2}$$

Trois cas peuvent se présenter:

1. On peut évaluer P_X et P_Y; si on trouve une étendue pour chacun, on prend la valeur la plus près de 1/2 (si on pense que P_X varie de 0,1 à 0,3 et P_Y varie de 0,6 à 0,8 on prendra $P_X = 0{,}3$, $P_Y = 0{,}6$ donc $Q_X = 0{,}7$ et $Q_Y = 0{,}4$).
2. Mener une enquête pilote pour estimer P_X et P_Y par p_X et p_Y.
3. Prendre la valeur maximum de PQ soit $P_X Q_X = P_Y Q_Y = 1/4$; la formule devient alors

$$n = \frac{z^2}{2E^2}$$

Exemple 1

Un éditeur suppose qu'il y a un plus grand pourcentage de gens qui lisent les pages des sports parmi ceux qui achètent un journal du matin que parmi ceux qui achètent un journal du soir. Quelle taille d'échantillon devra-t-il choisir s'il veut être certain à 95% qu'il connaîtra la différence à 10% près? Au niveau de 95%, $z = \pm 1{,}96$; $E = 0{,}10$ $n = \dfrac{(1{,}96)^2}{2\,(0{,}1)^2} = 192{,}1$

donc $n = 193$; chacun des échantillons devra avoir un minimum de 193 lecteurs.

Exemple 2

À une joute de football, on trouve que dans un échantillon de 600 personnes, il y a 200 femmes. À une joute de baseball, on trouve 20 femmes sur 100 personnes. (a) Les échantillons sont-ils assez grands pour être certain, à 95%, qu'on connaîtra la différence des proportions des femmes aux joutes de football et de baseball avec une erreur de moins de 5%? (b) Quelle est la taille minimum de chaque échantillon pour avoir une erreur inférieure à 5% avec une certitude de 95%?

(a) $n_X = 600$, $p_X = 1/3$; $n_Y = 100$, $p_Y = 1/5$.

Au niveau de 95%, $z = \pm 1{,}96$.

et $|z|\sigma_{(p_X - p_Y)} < 0{,}05$?

$$|z|\sigma_{(p_X - p_Y)} = 1{,}96\sqrt{\frac{1/3 * 2/3}{600} + \frac{1/5 * 4/5}{100}} = 1{,}96\sqrt{\frac{2}{5400} + \frac{4}{2500}} = 0{,}09 > 0{,}05$$

Les échantillons ne sont pas assez grands.

$$\text{(b)}\quad n = \frac{1{,}96^2(1/3 \cdot 2/3 + 1/5 \cdot 4/5)}{(0{,}05)^2} = 588$$

10.12 VOCABULAIRE ET SYMBOLES

P	q	Q
p	μ_p	σ_p
N	E	$\sigma\,(_{PX} - _{PY})$

10.13 EXERCICES

1. Une population est formée de 5 personnes: A, B, C qui conduisent des Renauds 5 et D, E des Cadillacs. Soit p la proportion de celles qui conduisent des Renauds 5.

(a) Trouver la distribution de la proportion échantillonnale pour des échantillons de taille 2 sans remise dans la population.

(b) Trouver μ_p et P et montrer qu'ils sont égaux.

(c) Calculer l'écart-type des proportions σ_p et montrer qu'il est égal à $\sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$.

2. Si 80% de tous les étudiants de 1re année au collégial suivent un cours d'anglais, quelle est la probabilité que dans un échantillon aléatoire de 100 étudiants il y ait entre 75 et 85% des étudiants qui suivent un cours d'anglais?

3. Un psychologue veut étudier la proportion des étudiants de collège dont les parents n'ont pas fait d'études collégiales. Il veut estimer cette proportion avec une certitude de 95%. Sur 4 000 étudiants choisis au hasard, il y en a 400 dont les parents n'ont pas fait d'études collégiales. Aidez-le à faire son estimation.

4. Un échantillon aléatoire de 200 travailleurs manuels montre que 80 ont un revenu annuel supérieur à $9 000. Trouver au niveau de 90% un intervalle de confiance pour le pourcentage de tous les travailleurs manuels qui gagnent plus de $9 000 par an.

5. (a) Une carte réponse fut adressée à 10 000 000 de votants leur demandant leur choix pour le Président. Un cinquième répondit, 56% des réponses favorisant M.L. Trouver au niveau de 68% un intervalle de confiance pour la proportion de tous les votants favorisant M.L.

(b) Dans l'élection qui suivit, seulement 34% votèrent pour M.L. Quelle est la probabilité que dans un échantillon aléatoire de 2 000 000 de personnes, 56% ou plus votèrent pour M.L.?

(c) Pouvez-vous expliquer la contradiction entre (a) et (b)?

6. Un médecin vérifie un nouveau médicament sur 100 malades ayant la leucémie. Après six ans, 60 sont encore vivants. Quelle proportion de tous les leucémiques seront encore vivants après six ans de traitement avec ce nouveau médicament, avec une certitude de 95%?

7. L'éditeur d'un journal étudiant interroge 200 étudiants sur la qualité des articles. 155 se disent insatisfaits. Avec une certitude de 90%, quelle proportion des étudiants sont insatisfaits des articles?

8. Dans un échantillon aléatoire de 100 pharmaciens, 77 gagnent plus de $30 000. Trouver un intervalle de confiance au niveau de 95% pour la proportion des pharmaciens qui gagnent plus de $30 000.

9. Un centre de traitement essaie une nouvelle méthode de désintoxication sur 100 héroïnomanes choisis au hasard. De ceux-ci, 36 eurent une désintoxication sans grands problèmes. Trouver un intervalle de confiance au niveau de 95% pour la proportion de tous les héroïnomanes qui seront désintoxiqués sans difficultés.

10. Un échantillon aléatoire de 150 travailleurs de General Motors montre que 105 possèdent une voiture G.M. Quelle proportion de tous les travailleurs possèdent un véhicule G.M. avec une certitude de 95%?

11. Un économiste trouve qu'un échantillon aléatoire de 100 métallos a un salaire hebdomadaire moyen de $300 avec un écart-type de $50 alors que 80 travailleurs de l'automobile ont un salaire moyen de $250 avec un écart-type de $45. Trouver au niveau de 95% un intervalle de confiance pour la différence des salaires des métallos et des travailleurs de l'automobile.

12. Deux femmes qui mesurent respectivement 160 cm et 168 cm habitent dans la maison X. Trois hommes qui mesurent 175 cm, 178 cm et 182 cm habitent dans la maison Y. Des échantillons de taille 1 sont pris dans chacune des maisons et la différence des grandeurs est calculée.

(a) Trouver la distribution de la différence des moyennes échantillonnales pour la grandeur.

(b) Calculer $\mu_{(\bar{X}-\bar{Y})}$, μ_X, μ_Y puis montrer que $\mu_{(\bar{X}-\bar{Y})} = \mu_X - \mu_Y$.

(c) Calculer $\sigma_{(\bar{X}-\bar{Y})}$, $\sigma_{\bar{X}}^2$, $\sigma_{\bar{Y}}^2$, puis montrer que $\sigma_{(\bar{X}-\bar{Y})} = \sqrt{\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2}$.

13. Un politicologue trouve que l'âge moyen des 50 votants qui favorisent J.P. à la direction du parti libéral est de 36 ans avec un écart-type de 4 ans, alors que 80 votants qui favorisent M.A. à la direction du même parti ont un âge moyen de 46 ans avec un écart-type de 5 ans. Trouver la différence des âges entre ceux qui favorisent J.P. et M.A. à la direction du parti avec une certitude de 98%.

14. Un professeur de psychologie donne deux cours identiques, l'un à 08:00 et l'autre à 13:00. Un échantillon aléatoire de 50 étudiants qui suivent le cours du matin passent un examen avec une moyenne de 82 et un écart-type de 8. Un échantillon de 60 étudiants qui suivent le cours de l'après-midi obtient une moyenne de 87 avec un écart-type de 6. Trouver un intervalle de confiance au niveau de 95% pour la différence des résultats entre les deux classes.

15. À l'Université de Montréal, 200 étudiants choisis au hasard suivent en moyenne 15 crédits par semestre avec un écart-type de 0,4. À l'Université du Québec, 100 étudiants choisis au hasard suivent en moyenne 16 crédits avec un écart-type de 0,3. Quelle est la différence du nombre moyen de crédits par semestre entre les deux universités au niveau de confiance de 90%?

16. L'âge moyen de 50 auditeurs de la station de radio CFTQ est 62 ans, avec un écart-type de 12 ans, alors que l'âge moyen de 60 auditeurs de la station CWAQ est 19 ans avec un écart-type de 3 ans. Trouver au niveau de 95% un intervalle de confiance pour la différence des âges des deux groupes d'auditeurs.

17. La durée moyenne de 50 chansons des Beatles choisies au hasard est de 3 minutes avec un écart-type de 30 secondes. Un échantillon aléatoire de 55 chansons des

Rolling Stones a une durée moyenne de 4 minutes 6 secondes avec un écart-type de 45 secondes. Au niveau de 98%, quelle est la différence de durée entre les chansons des Beatles et celles des Rolling Stones?

18. Cent joueurs de basketball ont une grandeur moyenne de 1,90 m avec un écart-type de 12 cm alors que 100 joueurs de hockey ont 1,80 m avec un écart-type de 9 cm. Trouver au niveau de 98% la différence de grandeur entre les joueurs de basketball et les joueurs de hockey.

19. La population X est formée des quatre nombres: 1; 2; 4; 8. La population Y est formée des nombres 5; 10; 20. Soient P_X et P_Y les proportions de nombres pairs dans ces populations.

(a) Trouver la distribution de la différence de proportions échantillonnales et des échantillons de taille 2 sont pris dans la population X et des échantillons de taille 1 dans la population Y. Échantillons pris indépendamment et avec remise.

(b) Calculer $\mu_{(p_X - p_Y)}$ ainsi que $P_X - P_Y$ puis montrer qu'ils sont égaux.

(c) Calculer $\sigma_{(p_X - p_Y)}$ et montrer qu'il est égal à $\sqrt{\frac{P_X Q_X}{n_X} + \frac{P_Y Q_Y}{n_Y}}$.

20. Un échantillon aléatoire de 1 000 libéraux montre que 400 sont en faveur de la peine de mort alors qu'un échantillon de 500 conservateurs montre que 250 sont en faveur de la peine de mort. Trouver un intervalle de confiance au niveau de 90% pour la différence de proportions entre les libéraux et les conservateurs qui favorisent la peine de mort.

21. Des échantillons de 800 adultes et de 400 adolescents ont été prélevés à 21:00 un soir de novembre. 200 adultes et 50 adolescents regardaient la T.V. (a) Trouver la différence de proportions entre les adultes et les adolescents qui regardaient la T.V. à ce moment avec un niveau de confiance de 95%.(b) Avec quel niveau de confiance peut-on affirmer que la différence entre les proportions est $0{,}125 \pm 0{,}05$

22. Dans un échantillon de 800 citadins, 480 sont en faveur des centres locaux de services communautaires (C.L.S.C.) alors que sur 500 banlieusards, 200 les favorisent. Trouver un intervalle de confiance au niveau de 98% pour la différence des proportions entre les citadins et les banlieusards qui favorisent l'implantation des C.L.S.C.

23. L'étude d'un échantillon de 1 200 bacheliers en chimie montre que 480 veulent travailler pour une multinationale immédiatement après la fin des études alors que seulement 60 bachelières sur 300 veulent faire de même. Trouver la différence de proportions entre bacheliers et bachelières qui veulent travailler pour une multinationale avec une certitude de 90%.

24. Deux villages voisins veulent fusionner leurs services municipaux. Chaque village fait une consultation populaire. Un échantillon de 1 000 personnes est choisi dans chaque village. Dans le premier, la proposition est adoptée dans une proportion de 9 à 1, alors que dans le second elle fut adoptée dans une proportion de 3 à 1. Trouver la différence de proportions de ceux qui favorisent la fusion avec une certitude de 95%.

25. Dans un échantillon de 100 Égyptologues Britanniques amateurs, 5 croient que les pyramides furent construites par lévitation. Cette croyance est partagée également par 4 Égyptologues amateurs Canadiens pris dans un échantillon de 56. Au niveau de 98%, quelle est la différence de proportions entre les Britanniques et les Canadiens qui croient que les pyramides furent construites par lévitation?

26. À une convention nationale du club Lions des échantillons aléatoires de 330 délégués venant des provinces de l'ouest et 300 venant de l'est, furent choisis. Parmi ceux venant de l'ouest, 231 veulent augmenter le nombre de représentants, alors que 180 de l'est s'opposent à cette mesure. Trouver, au niveau de 95%, un intervalle de confiance pour la différence de proportions des membres qui préfèrent ne pas augmenter le nombre de représentants.

●27. Un citoyen mécontent fait circuler une pétition parmi les 212 propriétaires et 44 commerçants de la rue. Seulement 10% des commerçants et 90% des propriétaires signèrent. La pétition demande de fermer la rue aux camions. Trouver un intervalle de confiance au niveau de 95% pour la différence de proportions de ceux qui approuvent la pétition.

28. Le fabricant de la nourriture pour chien (Fido plus) veut déterminer le gain moyen de poids par mois pour des chiens Esquimos de 1 an nourris au «Fido plus» avec une certitude de 95% et une erreur inférieure à 60 g. Il sait que l'écart-type de l'augmentation de poids pour ces chiens est de 150 g. Combien de chiens devra-t-il vérifier?

29. Le directeur du département de français du collège Rosemont veut comparer ses étudiants avec ceux des autres collèges. Un examen provincial a une moyenne de 500 et un écart-type de 100. Combien d'étudiants devra-t-il choisir dans son collège s'il veut connaître leur moyenne à 0,1 σ près avec une certitude de 95%?

30. Actuellement seulement 60% des leucémiques survivent plus de 6 ans. Un médecin a développé un nouveau médicament et veut savoir avec une certitude de 95% et une erreur inférieure à 20% le taux de survie après 6 ans. Combien de patients devra-t-il choisir pour vérifier son nouveau médicament?

31. Un politicologue veut connaître la différence de proportions des libéraux qui sont convaincus et ceux qui ne le sont pas. Quelle taille d'échantillons devra-t-il choisir s'il veut déterminer cette différence à 10% près et avec une certitude de 90%?

32. Un test de lecture doit être donné à des échantillons de garçons et de filles des écoles de Montréal. Combien de garçons et de filles doit-on prendre si leurs résultats moyens doivent se comparer à 4 points près avec une certitude de 90%?

33. Le régistraire de l'Université Laval veut comparer les résultats moyens des étudiants mariés avec ceux des célibataires. Il veut connaître la différence à 0,1 près avec une certitude de 95%. Il sait que l'an passé le résultat moyen de 100 étudiants mariés choisis au hasard était de 2,85 avec un écart-type de 0,4 alors que 100 célibataires choisits au hasard obtinrent 2,73 avec un écart-type de 0,3. Trouver la taille des échantillons qu'il doit choisir.

34. Le nombre de vols commis aux magasins de l'Ouest et à ceux de l'Est doit être comparé. On sait que l'an passé, 12% des magasins de l'Ouest et 15% de ceux de l'Est furent cambriolés. Trouver le nombre de magasins qu'on doit choisir cette année

(a) pour déterminer avec une erreur inférieure à 10% et une certitude de 90% la différence de proportions entre les magasins de l'Ouest et ceux de l'Est qui furent cambriolés.

(b) pour déterminer la différence entre le nombre de magasins cambriolés dans l'Ouest et dans l'Est.

35. On soupçonne un dé de ne pas être régulier et on veut le vérifier. On lance le dé et on enregistre le nombre de 6 obtenus. Combien de fois devra-t-on le lancer?

RÉPONSES

1. (a)

	p	f	fp	fp^2
AB, AC, BC	1	3	3	3
AD, AE, BD, BE, CD, CE	1/2	6	3	3/2
DE	0	1	0	0
		10	6	9/2

(b) $\mu_p = 6/10 = 3/5 = P.$

(c) $\sigma_p = \sqrt{\dfrac{9/2 - 6^2/10}{10}} = \sqrt{\dfrac{9}{100}}\sqrt{\dfrac{PQ}{n}}\sqrt{\dfrac{N-n}{N-1}} = \sqrt{\dfrac{3/5 * 2/5}{2} * \dfrac{5-2}{5-1}}$

$= \sqrt{\dfrac{9}{100}}.$

3. $\pm 1{,}96 = \dfrac{0{,}10 - P}{\sqrt{\dfrac{0{,}10 * 0{,}90}{4000}}}$; 9,1 à 10,9%

5. (a) Ces chiffres sont à peu près ceux obtenus en 1936 lors de la campagne électorale «Landon vs Roosevelt» aux États-Unis.

$$\pm\ 0{,}99 = \frac{0{,}56 - P}{\sqrt{\dfrac{0{,}56 * 0{,}44}{2\,000\,000}}};\ 55{,}97 \text{ à } 56{,}03 \text{ pour cent.}$$

(b) $z = \dfrac{0{,}56 - 0{,}34}{\sqrt{\dfrac{0{,}34 * 0{,}66}{2\,000\,000}}} = 657.$ La probabilité est trop petite.

(c) L'échantillonnage n'était pas aléatoire. Le sondage fut fait en consultant les personnes possédant une automobile ou le téléphone pendant la grande dépression économique.

7. $\pm 1.64 = \dfrac{0{,}775 - P}{\sqrt{\dfrac{0{,}775 * 0{,}225}{200}}}$; 73 à 82 % sont insatisfaits (il faut s'assurer que l'échantillon est aléatoire; une étude basée sur des réponses volontaires peut biaiser le résultat vers l'insatisfaction).

9. $\pm 1{,}96 = \dfrac{0{,}36 - P}{\sqrt{\dfrac{0{,}36 * 0{,}64}{100}}}$

L'intervalle de confiance au niveau de 95% de la proportion de tous ceux qui sont

désintoxiqués sans difficulté est [0,27; 0,45].

11. $\pm 1.96 = \dfrac{(300 - 250) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{50^2}{100} + \dfrac{45^2}{80}}}$; \$36 à \$64.

13. $\pm 2.33 = \dfrac{(36 - 46) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{4^2}{50} + \dfrac{5^2}{80}}}$; Ceux qui favorisent J.P. sont entre 8,1 et 11,9 ans plus jeunes que ceux qui favorisent M.A. avec un degré de confiance de 98%.

15. $\pm 1.64 = \dfrac{(15 - 16) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{0{,}4^2}{200} + \dfrac{0{,}3^2}{100}}}$; À l'université du Québec, les étudiants prennent entre 0,93 et 1,07 plus de crédits par semestre avec une certitude de 90%.

17. $\pm 2.33 = \dfrac{(3 - 4{,}1) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{0{,}5^2}{50} + \dfrac{0{,}75^2}{55}}}$; J'ai une certitude de 98% que les chansons des Beatles sont entre 0,81 et 1,39 minute (49 à 83 s) plus courtes que celles des Rolling Stones.

19.

Échantillons			Échantillons			Échantillons		
X	Y	$p_X - p_Y$	X	Y	$p_X - p_Y$	X	Y	$p_X - p_Y$
1;1	5	0	1;1	10	− 1,0	1;1	20	− 1,0
1;2	5	0,5	1;2	10	− 0,5	1;2	20	− 0,5
1;4	5	0,5	1;4	10	− 0,5	1;4	20	− 0,5
1;8	5	0,5	1;8	10	− 0,5	1;8	20	− 0,5
2;1	5	0,5	2;1	10	− 0,5	1;8	20	− 0,5
2;2	5	1,0	2;2	10	0	2;2	20	0
2;4	5	1,0	2;4	10	0	2;4	20	0
2;8	5	1,0	2;8	10	0	2;8	20	0
4;1	5	0,5	4;1	10	− 0,5	4;1	20	− 0,5
4;2	5	1,0	4;2	10	0	4;2	20	0
4;4	5	1,0	4;4	10	0	4;4	20	0
4;8	5	1,0	4;8	10	0	4;8	20	0
8;1	5	0,5	8;1	10	− 0,5	8;1	20	− 0,5
8;2	5	1,0	8;2	10	0	8;2	20	0
8;3	5	1,0	8;4	10	0	8;4	20	0
8;4	5	1,0	8;8	10	0	8;8	20	0

(a)

$p_X - p_Y$	f	$f(p_X - p_Y)$	$f(p_X - p_Y)^2$
− 1,0	2	−2	2
− 0,5	12	−6	3
0,0	19	0	0
0,5	6	3	3/2
1,0	9	9	9
	48	4	31/2

(b) $\mu_{(p_X - p_Y)} = 4/48 = 1/12$; $P_X - P_Y = 3/4 - 2/3 = 1/12$.

(c) $\sigma_{(p_X - p_Y)} = \sqrt{\frac{31/2 - 4^2/48}{48}} = \sqrt{\frac{93/6 - 2/6}{48}} = \sqrt{\frac{91}{288}}$

$$\sqrt{\frac{P_X Q_X}{n_X} + \frac{P_Y Q_Y}{n_Y}} = \sqrt{(3/4 * 1/4)/2 + (2/3 * 1/3)/1} = \sqrt{91/288}.$$

21. (a) $\pm 1{,}96 = \dfrac{1/8 - (P_X - P_Y)}{\sqrt{\frac{1/4 * 3/4}{800} + \frac{1/8 * 7/8}{400}}}$; 8,1 à 16,9%

(b) $|z| * \sigma_p = |z| * 0{,}023 = 0{,}05$
$|z| = 2{,}17$
97% certain

23. $p_X = 0{,}40$, $n_X = 1200$, $p_Y = 0{,}20$, $n_Y = 300$, $z = \pm 1{,}64$

$$\sigma(p_X - p_Y) = \sqrt{\frac{0{,}40 * 0{,}60}{1200} + \frac{0{,}20 * 0{,}80}{300}} = 0{,}027$$

$$\pm 1{,}64 = \frac{0{,}20 - (P_X - P_Y)}{0{,}027}$$

$$P_X - P_Y = 0{,}20 \pm 0{,}044 = -0{,}117 \text{ ou } 0{,}075$$

entre 15,6 et 24,4% plus de chimistes veulent travailler pour une multinationale.

25. $p_X = 0{,}05$, $p_Y = 0{,}071$, $n = 100$, $n_Y = 56$, $z = \pm 2{,}33$.

$$\sigma_{(p_X - p_Y)} = \sqrt{\frac{0{,}05 * 0{,}95}{100} + \frac{0{,}071 * 0{,}929}{56}} = \sqrt{0{,}00165} = 0{,}041$$

$$\pm 2{,}33 = \frac{(0{,}05 - 0{,}071) - (P_X - P_Y)}{0{,}041}$$

$$P_X - P_Y = -0{,}021 \pm 0{,}0955 = -0{,}177 \text{ ou } 0{,}075$$

Avec une certitude de 98%, la différence dans le pourcentage entre les Égyptologues Britanniques et Canadiens est − 12 à + 8%. Les échantillons sont trop petits.

27. J'espère que vous n'avez pas trop perdu de temps sur ce problème, les échantillons ne sont pas aléatoires.

29. $E = 0{,}1\ \sigma$ donc $n = \dfrac{1{,}96^2\ \sigma^2}{(0{,}1)^2\ \sigma^2} = 385$. Si le nombre d'étudiants est trop grand, on peut augmenter l'erreur ou diminuer le degré de confiance (au niveau de 75% avec une erreur $E = 0{,}2\ \sigma$, on obtient $n = 34$ par exemple).

31. $n = \dfrac{z^2}{4E^2} = \dfrac{1{,}64^2}{4 * 0{,}1^2} = 68$ dans chaque échantillon.

33. En posant que les écarts-types n'ont pas changé, on a $\sigma_X = 0{,}4$ $\sigma_Y = 0{,}3$

$$n = \frac{z^2(\sigma_X^2 + \sigma_Y^2)}{E^2} = \frac{1{,}96^2\,(0{,}4^2 + 0{,}3^2)}{0{,}1^2} = 97 \text{ dans chaque échantillon.}$$

35. On ne précise pas l'erreur et le degré de confiance, on ne peut donc pas

résoudre ce problème. (Au chapitre 11, on verra comment vérifier l'hypothèse «le dé est régulier» après l'avoir lancé un certain nombre de fois.)

Chapitre 11

Vérification d'hypothèse

Dans ce chapitre, nous apprendrons à formuler et vérifier une hypothèse sur la moyenne ou la proportion d'une population, ou sur la différence de moyennes ou de proportions de deux populations. Nous apprendrons comment formuler une hypothèse rivale qui sera acceptée si la première hypothèse est rejetée, comment décider si la première sera rejetée et quelle est la probabilité d'erreur. Cela ne sera pas difficile si nous réalisons que nous utiliserons une distribution d'échantillonnage à d'autres fins.

11.1 INTRODUCTION

Dans les chapitres 9 et 10, nous avons appris à faire une inférence pour la moyenne ou la proportion d'une population à partir de la moyenne ou la proportion d'un échantillon. Par exemple, si un échantillon aléatoire de taille 50 a une moyenne de 100 et un écart-type de 5, on peut alors affirmer avec une certitude de 95% que l'intervalle [98,6; 101,4] contient la moyenne de la population.

Nous allons maintenant procéder autrement. Nous ferons une affirmation ou **hypothèse** sur l'un ou l'autre des paramètres de la population (moyenne, proportion, différence de moyennes ou de proportions), puis vérifier si cette hypothèse est vraie en examinant un échantillon aléatoire pris dans la population. Un exemple, on veut faire l'hypothèse qu'une pièce de monnaie est régulière ($P = 1/2$). Si on lance la pièce 100 fois et qu'on obtient 50 faces, on dira probablement «J'accepte une légère fluctuation, car si je lance la pièce 100 autres fois, j'obtiendrai peut-être alors 51 faces». Donc on acceptera sans hésiter l'hypothèse $P = 1/2$. Si on obtient seulement 1 face en 100 jets, on rejettera sans hésiter l'hypothèse $P = 1/2$. Il est à noter que si $P = 1/2$ (la pièce est régulière), la probabilité d'obtenir une face en 100 jets n'est pas nulle, en effet P (1 face) $= {}_{100}C_1\,(1/2)^1\,(1/2)^{99}$, qui est très petite mais non nulle. Il y a donc une petite probabilité qu'on se trompe si on rejette l'hypothèse $P = 1/2$ quand on obtient exactement 1 face en 100 jets. Intuitivement on acceptera l'hypothèse que $P = 1/2$ si on obtient 49;50 ou 51 faces en 100 jets et on la rejettera si on obtient 1 ou 99 faces. Que ferons-nous si on obtient 37 ; 59 ou 43 faces? Il est évident qu'on doit prendre une décision à savoir, accepter ou rejeter l'hypothèse $P = 1/2$.

En général, un test d'hypothèse comporte les étapes suivantes:

1. Choisir l'hypothèse à vérifier.
2. Choisir une hypothèse rivale qui sera acceptée si l'hypothèse originale est rejetée.
3. Se fixer une règle pour prendre la décision d'accepter ou de rejeter l'hypothèse.
4. Choisir un échantillon aléatoire dans la population et calculer la statistique appropriée; moyenne, proportion, variances et autres.
5. Prendre la décision.

11.2 CHOIX DE L'HYPOTHÈSE À VÉRIFIER (H_0)

L'hypothèse qu'on veut vérifier est appelée hypothèse nulle, ce qui veut dire que son acceptation indique que le procédé n'a pas d'effet ou que la différence est nulle. Elle est désignée par H_0 (lire H zéro).

Dans l'expérience du jet d'une pièce de monnaie de la section précédente, l'hypothèse que la pièce est régulière peut être exprimée symboliquement par:

$$H_0 : P = 1/2.$$

H_0 est toujours une affirmation concernant un aspect (moyenne, proportion, etc.) d'une population, jamais d'une statistique, d'un échantillon. H_0 peut être $\mu = 100$ ou $\mu_1 - \mu_2 = 4$ ou $P = 0{,}6$ ou $P_1 - P_2 = 0$, jamais $\bar{X} = 4$ ou $p = 0{,}36$,etc.«Un échantillon aléatoire de taille 50 a une moyenne $\bar{X} = 100$ et $s = 4{,}5$. Au niveau de signification de 0,05, est-ce que μ peut être 101,5? Les nombres 100 et 4,5 ne seront jamais utilisés dans H_0. Ici H_0 sera $\mu = 101{,}5$.»

H_0 est habituellement facile à déterminer.

Il y a une convention qu'on doit accepter pour le moment (nous justifierons à la section 11.7). H_0 est une égalité ($\mu = 14$) et non une inégalité du type ($\mu \geqslant 14$ ou $\mu < 14$). Pour le moment, il y a quatre choix possibles pour H_0 :

$$\mu = ____$$
$$P = ____$$
$$\mu_X - \mu_Y = ____$$
$$P_X - P_Y = ____$$

Plus loin dans ce chapitre, nous apprendrons comment vérifier d'autres types d'hypothèses. Nous pourrons vérifier, par exemple, si une population a un certain type de distribution ou encore si deux populations ont la même variance. Revenons à nos hypothèses sur la moyenne ou la proportion.

Exemple 1

Un manufacturier de hamacs achète sa corde de la compagnie Nylon Inc. Un échantillon de 36 cordes choisies au hasard de la dernière livraison a une force de rupture moyenne de 58 kg avec un écart-type de 3 kg. La force de rupture désirée est de 60 kg. Acceptera-t-il cette dernière livraison avec un niveau de signification de 0,05?

Quel est H_0?

Il ne faut pas prendre $\bar{X} = 58$ ni $\mu = 58$ ni $\mu \geqslant 60$. Ici $H_0 : \mu = 60$.

Déterminer H_0 dans les exemples suivants:

Exemple 2

Actuellement seulement 60% des leucémiques survivent plus de 6 ans. Un médecin produit un nouveau médicament. Il le vérifie sur 40 patients choisis au hasard et après 6 ans, 26 sont encore vivants. Le nouveau médicament est-il supérieur au traitement précédent avec un niveau de signification de 0,10?

$p = 26/40$ provient de l'échantillon; la réponse $H_0 : P = 0{,}60$.

Exemple 3

Un botaniste a développé une nouvelle variété de riz qui d'après lui augmentera la production de 1 m³ par 4 000 m² de surface. Des semences de la nouvelle variété et de l'ancienne furent sélectionnées au hasard, puis un champ de 240 000 m² présentant des caractéristiques (fertilité, humidité, etc.) uniformes fut divisé en 60 lots de 4 000 m². Trente de ces lots furent choisis au hasard puis ensemencés par la nouvelle variété. La récolte de la nouvelle variété fut de 6,3 m³ par lot avec un écart-type de 0,15 m³. Les 30 autres lots ensemencés par l'ancienne variété donna une récolte de 5,4 m³ par lot avec un écart-type de 0,17 m³. Est-ce que la nouvelle variété de riz produit au moins 1 m³ de plus par 4 000 m² avec un niveau de confiance de 0,05?

Si X représente la nouvelle variété et Y l'ancienne alors $\overline{X} - \overline{Y} = 0{,}9$ ne peut être utilisé comme hypothèse de même que $\mu_X \geqslant \mu_Y + 1$, car on a une inéquation. On posera $H_0 : \mu_X - \mu_Y = 1$.

Exemple 4

Dans une laiterie, une embouteilleuse doit verser 1 litre (1 000 cm³) de lait dans chaque contenant avec un écart-type de 2 cm³. Lors d'une vérification journalière, 36 contenants sont choisis au hasard et on trouve qu'ils contiennent en moyenne 1,003 l. (1 003 cm³). La machine est-elle bien ajustée au niveau de signification de 0,05?

L'hypothèse sera $H_0 : \mu = 1\,000$ cm³.

Exemple 5

On croit que le salaire moyen des professeurs des collèges publics est de \$1 000 supérieur à celui des professeurs des collèges privés. Un échantillon aléatoire de 100 professeurs de collèges publics montre que ceux-ci reçoivent un salaire moyen de \$15 000 avec un écart-type de \$1 000, alors qu'un échantillon de 100 professeurs de collèges privés, choisis au hasard, montre que ceux-ci reçoivent un salaire moyen de \$14 200 avec un écart-type de \$900. Au niveau de signification de 0,01, est-ce que le salaire moyen des professeurs des collèges publics est de \$1 000 supérieur à celui des professeurs des collèges privés?

$H_0 : \mu_X - \mu_Y = 1\,000$ (X: collèges publics et Y: collèges privés).

11.3 CHOIX DE L'HYPOTHÈSE RIVALE

La notation H_1 (lire H un) est utilisée pour l'hypothèse qui sera acceptée si H_0 est rejetée. H_1 doit également être déterminée avant de faire le test. Tout comme H_0, H_1 ne dépend pas des valeurs de l'échantillon. Si on s'interroge sur la régularité d'une pièce de monnaie, H_0 est $P = 1/2$ et $H_1 : P \neq 1/2$ mais on peut avoir d'autres hypothèses rivales, $P < 1/2, P > 1/2$ ou $P = 0{,}1$. Le nombre d'hypothèses est assez limité; voici les choix possibles:

Si H_0 est:	Alors H_1 sera:
$\mu = A$	$\mu \neq A$ ou $\mu < A$ ou $\mu > A$
$P = B$	$P \neq B$ ou $P < B$ ou $P > B$
$\mu_X - \mu_Y = C$	$\mu_X - \mu_Y \neq C$ ou $\mu_X - \mu_Y < C$ ou $\mu_X - \mu_Y > C$
$P_X - P_Y = D$	$P_X - P_Y \neq D$ ou $P_X - P_Y < D$ ou $P_X - P_Y > D$

où A; B; C; D sont des constantes.

Pourquoi doit-on choisir une hypothèse rivale avant de vérifier H_0 qu'on a déjà? Pourquoi ne peut-on pas accepter ou rejeter l'hypothèse nulle puis, penser à l'hypothèse rivale seulement dans le cas d'un rejet de H_0?

La raison est que la formulation de l'hypothèse rivale affectera la décision d'accepter ou de rejeter l'hypothèse nulle. Cette partie sera discutée en détail dans la section 11.4. Essayons quand même de formuler l'hypothèse rivale des exemples de la section précédente (11.2).

Réponses

Exemple 1

$H_0 : \mu = 60$; $H_1 : \mu < 60$. Si la moyenne de l'échantillon est supérieure à 60, on sera heureux d'accepter la livraison, mais si la moyenne de l'échantillon est inférieure à la moyenne désirée, on refusera la livraison.

Exemple 2

$H_0 : P = 0{,}6$; $H_1 : P > 0{,}6$. Le médecin veut savoir si son médicament est supérieur au traitement précédent. Si $P > 0{,}6$ de façon significative, il acceptera H_1. Si $P \leq 0{,}6$, il acceptera H_0 et alors il devra poursuivre ses expériences ou chercher un autre médicament.

Exemple 3

H_0: $\mu_x - \mu_y = 1$; H_1: $\mu_x - \mu_y > 1$ car le botaniste tentera de prouver que son produit est meilleur, sinon, par défaut, il choisira H_0. Si H_1 est acceptée, ceci veut dire que chaque 4 000 m² de surface produira au moins 1 m³ plus de riz. Si H_0 est acceptée, le botaniste n'encouragera pas la production de la nouvelle variété de riz.

Exemple 4

$H_0 : \mu = 1\,000$ cm³; $H_1 : \mu \neq 1\,000$ cm³. On veut savoir si la machine est bien ajustée; or si $\mu \neq 1\,000$ cm³, elle est mal ajustée.

Exemple 5

$H_0 : \mu_X - \mu_Y = 1\,000$; $H_1 : \mu_X - \mu_Y > 1\,000$.

11.4 RÈGLE DE DÉCISION

On doit avoir une méthode permettant de prendre une décision; H_0 est acceptée (et H_1 est rejetée) ou H_0 est rejetée (et H_1 est acceptée). Chaque échantillon choisi au hasard n'aura pas la même moyenne ou proportion que la population d'où il est prélevé. Il semble raisonnable d'accepter l'hypothèse nulle si la valeur de l'échantillon est relativement près de la valeur prévue par l'hypothèse nulle. Par contre, si la valeur de l'échantillon est loin de la valeur prévue et que ceci n'est pas dû au hasard, on dira que la différence est significative et on rejettera l'hypothèse nulle. La méthode permettant de prendre une décision doit être établie **avant** de vérifier sur l'échantillon. Comment définir une différence significative?

Une hypothèse nulle est vraie ou fausse et est soit acceptée ou rejetée. On ne fait pas d'erreur en acceptant une hypothèse vraie ou en rejetant une hypothèse fausse. On commet une erreur si on rejette une hypothèse vraie ou accepte une hypothèse fausse.

Définitions: Une **erreur de 1re espèce** est commise si on rejette H_0 alors que celle-ci est vraie.

Une **erreur de 2e espèce** est commise si on accepte H_0 et que celle-ci est fausse.

On utilise deux lettres grecques α (alpha) et β (bêta)

Notations: α est la probabilité de commettre une erreur de 1re espèce; α est aussi le **niveau de signification.**

β est la probabilité de commettre une erreur de 2e espèce.

La table suivante résume les relations.

		Si H_0 est acceptée	Si H_0 est rejetée
Quand H_0 est	Vraie	Pas d'erreur	Erreur de 1re espèce (α)
	Fausse	Erreur de 2e espèce (β)	Pas d'erreur

Supposons qu'on veut vérifier si une pièce de monnaie est régulière ($H_0: P = 1/2$; $H_1: P \neq 1/2$) en la lançant 100 fois.On doit décider si on accepte ou rejette H_0. Nous apprendrons bientôt comment prendre une bonne décision. Pour le moment, considérons quatre choix arbitraires différents et examinons l'effet sur (α) et (β).

Exemple 1

$H_0 : P = 1/2$ est toujours acceptée quel que soit le nombre de faces

obtenues en 100 jets. (a) Trouver α. (b) Trouver β si P est égal à: (i) 0,35; (ii) 0,40; (iii) 0,45.

(a) $\alpha = 0$ (puisqu'on ne rejette jamais H_0, la probabilité est 0 que H_0 soit rejetée quand elle est vraie).

(b) $\beta = 1{,}00$ (quelle que soit la valeur de P). Puisqu'on accepte toujours H_0 qu'elle soit vraie ou fausse, on est certain de l'accepter quand elle sera fausse.

Exemple 2

$H_0 : P = 1/2$ est acceptée si on obtient un nombre de faces compris entre 40 et 60 inclusivement et rejetée dans les autres cas. (a) Trouver α. (b) Trouver β si P est égal à (i) 0,35 (ii) 0,40 (iii) 0,45.

(a) La probabilité d'obtenir moins de 40 ou plus de 60 faces quand $P = 1/2$ est 0,04 (voir problème no 15, section 7.13). En effet, $P = 1/2$; $n = 100$; $P(40 \leqslant X \leqslant 60) = ?$ $\mu = np = 100 * 1/2 = 50$; $\sigma = \sqrt{npq} = \sqrt{25} = 5$ comme la distribution binômiale est approximée par une normale $P\ (39{,}5 < X < 60{,}5) = P(-2{,}10 < z < 2{,}10) = 0{,}964$ donc $P\ (X < 40$ et $X > 60) = 1 - P\ (40 \leqslant X \leqslant 60) = 1 - 0{,}964 = 0{,}036$. Donc $\alpha = 0{,}04$ en arrondissant.

(b) (i) si $P = 0{,}35$ alors $\mu = 0{,}35 * 100 = 35$, $\sigma = \sqrt{100 * 0{,}35 * 0{,}65} = 4{,}77$ et $\beta = P\ (39{,}5 < X < 60{,}5)$
$\beta = P\ (0{,}94 < z < 5{,}35) = 0{,}5 - 0{,}33 = 0{,}17$;
(ii) si $P = 0{,}40$ alors $\mu = 40$, $\sigma = \sqrt{100 * 0{,}4 * 0{,}6} = 4{,}90$ et
$\beta = P\ (39{,}5 < X < 60{,}5) = P\ (-0{,}1 < z < 4{,}18) = 0{,}5 + 0{,}04$ $\beta = 0{,}54$.
(iii) si $P = 0{,}45$ alors $\mu = 45$, $\sigma = \sqrt{100 * 0{,}45 * 0{,}55} = 4{,}97$ et $\beta = P\ (39{,}5 < X < 60{,}5) = P\ (-1{,}11 < z < 3{,}12) = 0{,}87$.

Exemple 3

$H_0 : P = 1/2$ est acceptée si on obtient entre 43 et 57 faces inclusivement. (a) Trouver α. (b) Trouver β si P est: (i) 0,35; (ii) 0,40; (iii) 0,45.

(a) La probabilité d'obtenir moins de 43 et plus de 57 faces si $P = 1/2$ est $1 - 0{,}87 = 0{,}13$ (voir (a) problème précédent).

(b) (i) $\beta = P\ (42{,}5 < X < 57{,}5) = P\ (1{,}57 < z < 4{,}72) = 0{,}50 - 0{,}44 = 0{,}06$.
(ii) $\beta = P\ (42{,}5 < X < 57{,}5) = P\ (0{,}51 < z < 3{,}57) = 0{,}50 - 0{,}20 = 0{,}30$;
(iii) $\beta = P\ (42{,}5 < X < 57{,}5) = P\ (-0{,}5 < z < 2{,}52) = 0{,}192 + 0{,}494 = 0{,}69$.

Exemple 4

$H_0 : P = 1/2$ est toujours rejetée quel que soit le nombre de faces obtenues. (a) Trouver α. (b) Trouver β.

(a) $\alpha = 1{,}00$ (si on rejette toujours H_0, alors on est certain de la rejeter quand elle sera vraie).

(b) $\beta = 0$ (comme on n'accepte jamais H_0, on ne l'acceptera jamais quand elle sera fausse).

Ces exemples vous ont permis de revoir la distribution binômiale et son approximation par une distribution normale mais encore ils illustrent certains points importants. Voici un résumé des résultats précédents:

Exemple	Accepter H_0 et rejeter H_1	(a) α	(b) β $H_1: P = 0{,}35$	(c) β $H_1: P = 0{,}4$	(d) β $H_1: P = 0{,}45$
1	Toujours	0	1,0	1,0	1,0
2	si 40 — 60 faces	0,04	0,17	0,54	0,87
3	si 43 — 57 faces	0,13	0,06	0,30	0,69
4	jamais	1,00	0	0	0

Il est à noter:

1. Après que H_0 et H_1 sont choisies, il reste à prendre une décision. Comment décider si on accepte ou rejette l'hypothèse nulle? Nous allons découvrir comment dans ce qui suit. Dans cet exemple des choix arbitraires furent faits pour montrer l'effet sur α et β.

2. α peut être facilement déterminé. β peut être déterminé seulement si H_1 est connue ($P = 0{,}4$; $P = 0{,}45$ etc.) et varie si H_1 varie; la valeur de β ne peut être déterminée si H_1 est un énoncé général «$P \neq 1/2$», (exceptions faites des cas ou H_0 est toujours acceptée ou jamais acceptée indépendamment du résultat de l'expérience. Il est inutile alors de faire l'expérience).

3. Si α augmente, β décroît. Comparer la colonne a avec les colonnes b; c; d. Au fur et à mesure que la probabilité de rejeter H_0 quand elle est vraie augmente, la probabilité d'accepter H_0 quand elle est fausse diminue. Il n'est pas vrai en général que $\beta = 1 - \alpha$, car α est calculé en supposant que la population a une distribution binômiale avec $P = 0{,}5$ alors que β est calculé sous différentes valeurs de la population ($P = 0{,}35$; $P = 0{,}40$; $P = 0{,}45$). La relation entre α et β est plus compliquée et sera étudiée à la section 11.7.

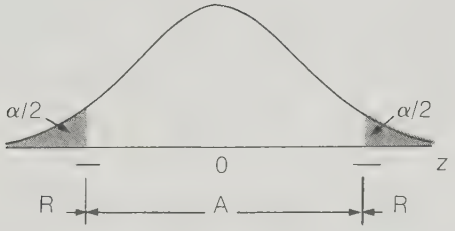
Dans cet exemple différentes décisions furent faites arbitrairement. Il a été décidé d'accepter H_0 si le nombre de faces était entre deux limites et de rejeter H_0 si le nombre de faces était à l'extérieur de ces limites. Puis α fut déterminé pour chacun de ces choix. En pratique, on procède autrement: α est choisi arbitrairement puis, les limites d'acceptation de H_0 sont calculées; si la statistique de l'échantillon est à l'extérieur de ces limites, H_0 est rejetée (et H_1 est acceptée).

La formulation de H_1 déterminera le genre de limites à calculer. Examinons comment prendre une décision sur l'acceptation ou le rejet de H_0 quand H_1 comporte le symbole « $\neq$ ». H_1 a la forme $\mu \neq \ldots$, $P \neq \ldots$, $\mu_X - \mu_Y \neq \ldots$, $P_X - P_Y \neq \ldots$, la décision est prise de la façon suivante:

1. α est arbitrairement choisi petit (habituellement 0,01; 0,05 ou 0,10). On discutera ce choix à la section 11.6. α est le **niveau de signification.**

2. Les valeurs de z sont déterminées de sorte que l'aire ombrée sous la courbe normale, ici les queues de la distribution, aient une aire de $\alpha / 2$. Les valeurs les plus fréquentes de z sont:

α	0,10	0,05	0,01
z	±1,64	±1,96	±2,58



La formulation de H_1 ($\mu \neq$ ... etc.) que l'on considère maintenant nous amène à un test bilatéral. À la section 11.5, on verra que si H_1 a la forme $\mu <$ ..., $P <$ ..., $\mu_X - \mu_Y <$ ... etc., nous aurons un test unilatéral. Il faut se rappeler que H_1 est fixée avant de faire l'expérience et ainsi on saura s'il faut faire un test bilatéral ou unilatéral.

3. L'expérience est faite puis la valeur de z de la statistique ($\overline{X}$, p, $\overline{X} - \overline{Y}$ ou $p_X - p_Y$) est déterminée. Si cette valeur de z est à l'intérieur des limites mentionnées ci-haut, H_0 est acceptée. Si cette valeur de z tombe à l'extérieur des limites, H_0 est rejetée (et H_1 est acceptée). Les limites qui déterminent l'acceptation ou le rejet de H_0 sont appelées **valeurs critiques de z**.

Étudier attentivement les exemples suivants pour découvrir la logique de ce test d'hypothèses.

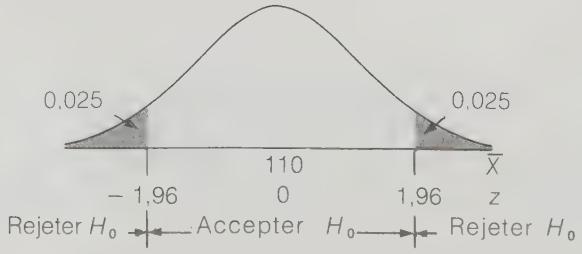
Exemple 1

On croit que le quotient intellectuel moyen de tous les étudiants du niveau collégial est 110 avec un écart-type de 10. Un échantillon aléatoire de 49 étudiants du collège Laval présente un Q.I. moyen de 112. Est-ce que le Q.I. moyen des étudiants du collège Laval diffère de celui de tous les étudiants du collégial au niveau de signification de 0,05?

H_0 : $\mu = 110$ (la population de tous les étudiants du collégial).
H_1: $\mu \neq 110$ (le Q.I. moyen des étudiants de Laval est différent de celui de tous les étudiants).

$\alpha = 0{,}05$

$n = 49$, $\overline{X} = 112$



Se basant sur la distribution d'échantillonnage de la moyenne, l'erreur-type $s = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{49}} = \frac{10}{7}$. La probabilité de rejeter H_0 quand elle est vraie est $\alpha = 0{,}05$.

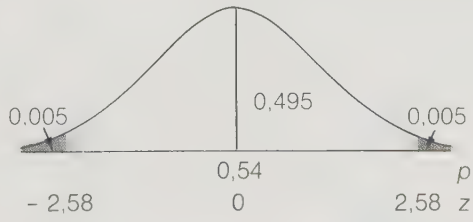
Chaque aire ombrée sera 0,025 et les valeurs correspondantes de z seront $\pm$ 1,96. Ceci veut dire que si la valeur z de la moyenne de l'échantillon tombe en dehors de l'intervalle (− 1,96; 1,96), H_0 est rejetée. La probabilité d'obtenir un tel échantillon est si faible (moins de 5%) qu'il est peu probable qu'elle provienne de la fluctuation de l'échantillonnage; donc elle provient du fait que H_0 est fausse. Si la moyenne de l'échantillon a une valeur z qui est entre − 1,96 et 1,96, alors voici le raisonnement que l'on peut faire: chaque échantillon ne peut avoir une moyenne égale à la moyenne de la population. Si H_0 est vraie, il y a une probabilité de 0,95 que la moyenne de l'échantillon aura un z qui sera compris entre − 1,96 et 1,96. Dans ces conditions, H_0 est acceptée. La cote z pour l'échantillon de 49 étudiants de Laval est $\frac{112 - 110}{10/\sqrt{49}} = 1{,}40$ donc l'hypothèse est acceptée; les étudiants du collège Laval ont le même Q.I. que tous les autres étudiants du collégial.

Exemple 2

Un politicologue sait qu'un certain comté a voté à 54% libéral lors des dernières élections. Il sait aussi que depuis ce temps des changements importants sont survenus dans ce comté et croit que ceux-ci changeront la proportion de ceux qui voteront libéral aux prochaines élections. Mais il ne sait pas dans quel sens ira le changement. Il fait un sondage auprès de 1 000 votants de ce comté et trouve que 480 ont l'intention de voter libéral. Son appréhension est-elle fondée au niveau de signification de 0,01?

$$H_0 : P := 0{,}54;\ H_1 : P \neq 0{,}54.$$

Ici $\alpha = 0{,}01$; les régions de rejets sont telles que la probabilité de rejeter H_0 quand elle est vraie est 0,01. Ceci veut dire que la distribution d'échantillonnage a une proportion de 0,54 (ce qui est le cas quand H_0 est vraie) alors on doit trouver des régions de sorte que la probabilité que la proportion d'un échantillon aléatoire sera dans celles-ci est de 0,01. Donc H_1 a la forme de $P \neq 0{,}54$. On trouve 2 régions ayant chacune une aire de 0,01/2 = 0,005.

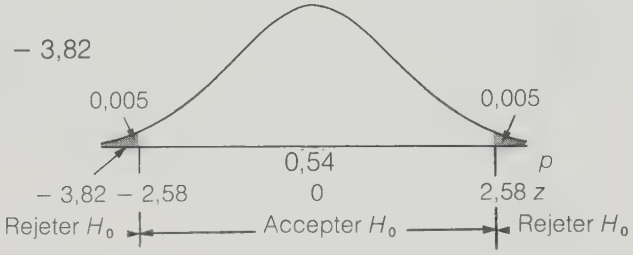


On trouve les valeurs de z (dans la table 4, appendice C) correspondant à une aire de 0,5 − 0,005 = 0,495; on obtient $z = \pm$ 2,58. L'hypothèse H_0 sera rejetée (et H_1 acceptée) si la valeur z de la proportion de l'échantillon est à l'extérieur de l'intervalle [−2,58; 2,58] et H_0 sera acceptée si z est à l'intérieur de [−2,58; 2,58].

Dans notre échantillon, on a une proportion qui votera libéral de:
$p = 480/1\,000 = 0{,}48$; $n = 1\,000$

$$\sigma_P = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0{,}54 * 0{,}46}{1{,}000}} = 0{,}0158$$

$$z = \frac{p - P}{\sigma_p} = \frac{0{,}48 - 0{,}54}{0{,}0157} = -3{,}82$$



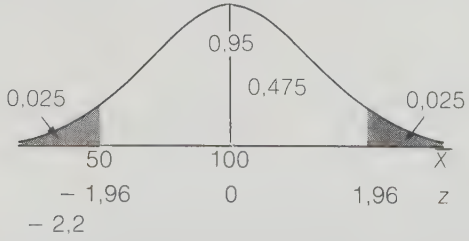
La valeur de z tombe dans la région de rejet de H_0. Il peut conclure qu'il y a un changement significatif de la proportion de ceux qui voteront libéral. Il est à noter que l'on a une probabilité de 1% de se tromper si H_0 est vraie (pas de changement). Or dans ce cas, la valeur de z qui tombe dans la zone de rejet a une probabilité faible d'être due au hasard, ce qui entraîne le rejet de H_0.

Exemple 3

En 1965, un économiste a comparé les salaires des électriciens et des menuisiers au Québec. Il a découvert qu'un électricien gagne \$100 de plus par mois qu'un menuisier. Il reprend son étude en 1975 et se demande si la même différence persiste. Il trouve que le salaire moyen de 100 électriciens choisis au hasard est de \$1 000 avec un écart-type de \$200 alors qu'un échantillon de 81 menuisiers a un salaire moyen de \$950 avec un écart-type de \$100. Au niveau de signification de 0,05 peut-il affirmer que la différence des salaires entre électriciens et menuisiers n'a pas changé?

$H_0 : \mu_X - \mu_Y = 100$ (différence de salaires)

$H_1 : \mu_X - \mu_Y \neq 100$ (on ne sait pas si la différence des salaires a augmenté ou diminué, on veut savoir si elle a varié).



On doit faire un test bilatéral car H_1 a la forme de ($\mu_X - \mu_Y \neq$). Le niveau de signification ($\alpha = 0{,}05$) nous indique de trouver la valeur de z correspondant à une aire de 0,475; $z = 1{,}96$. On doit utiliser ici une distribution de différences de moyennes. Sa moyenne est $\mu_X - \mu_Y = 100$ (provient de H_0).

$$n_X = 100,\ s_X = 200,\ \sigma_{\bar{X}} = \frac{200}{\sqrt{100}} = 20$$

$$n_Y = 81,\ s_Y = 100,\ \sigma_{\bar{Y}} = \frac{100}{\sqrt{81}} = 11{,}1$$

$$\sigma_{(\bar{X}-\bar{Y})} = \sqrt{20^2 + 11{,}1^2} = \sqrt{523{,}21} = 22{,}9$$

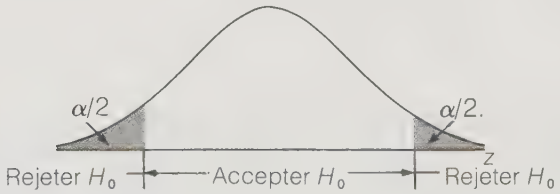
$$\bar{X} = 1000,\ \bar{Y} = 950$$

$$z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\bar{X}-\bar{Y})}} = \frac{(1000 - 950) - 100}{22{,}9} = \frac{-50}{22{,}9} = -2{,}2$$

H_0 est rejetée, car $-2{,}2 < -1{,}96$. On accepte H_1 à savoir que la différence des salaires entre électriciens et menuisiers n'est plus de \$100.

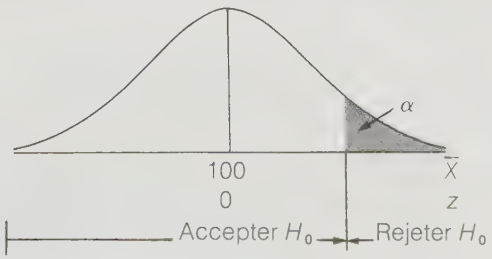
11.5 TEST UNILATÉRAL

Dans la section précédente, nous avons appris comment établir une hypothèse nulle H_0 et une hypothèse rivale H_1 de la forme $\mu \neq ...$, $P \neq ...$, etc., et comment prendre une décision. Ceci nous a permis de faire un test bilatéral. La règle était la suivante: chaque zone de rejet a une aire de $\alpha / 2$, ceci permet de trouver les valeurs critiques de z qui donnent une zone d'acceptation et deux zones de rejets. La valeur z de la statistique de l'échantillon est calculée, puis H_0 est acceptée si celle-ci est à l'intérieur de la zone d'acceptation, dans le cas contraire H_0 est rejetée.



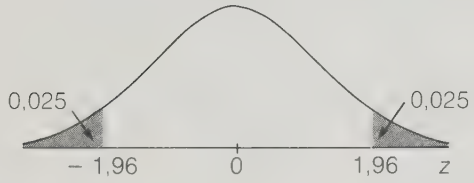
Examinons maintenant un test unilatéral dans lequel H_1 a la forme de $\mu < ..., P > ..., \mu_1 - \mu_2 > ...$ etc.

Supposons que H_0 ait la forme $\mu = 100$. Si H_1 a la forme de $\mu > 100$, la région de rejet est alors à l'extrémité droite de la distribution des moyennes au lieu d'être séparée en deux comme dans un test bilatéral. L'aire de cette région est alors α et la valeur critique de z en conformité.



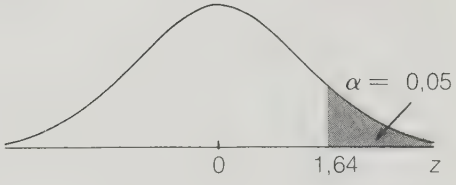
Test bilatéral

$H_0 : \mu = 100$; $H_1 : \mu \neq 100$; $\alpha = 0{,}05$



Test unilatéral

$H_0 : \mu = 100$; $H_1 : \mu > 100$; $\alpha = 0{,}05$



Les valeurs les plus utilisées pour α et les valeurs critiques correspondantes de z sont les suivantes:

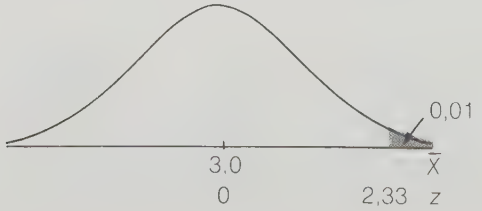
		z		
		Bilatéral	Unilatéral, <	Unilatéral, >
$\alpha =$	0,10	± 1,64	– 1,28	1,28
	0,05	± 1,96	– 1,64	1,64
	0,01	± 2,58	– 2,33	2,33

Il faut se rappeler que α est choisi arbitrairement comme précédemment.

Exemple 1

On sait qu'un chien d'un an gagne en moyenne 1,4 kg par mois avec un écart-type de 180 g. Une diète spéciale riche en protéines est donnée à 50 chiens d'un an choisis au hasard et leur augmentation moyenne de poids pour un mois fut de 1,47 kg avec un écart-type de 140 g. La diète spéciale, riche en protéines, favorise-t-elle l'augmentation de poids avec un niveau de signification de 0,01?

$H_0 : \mu = 1\,400$ g
$H_1 : \mu > 1\,400$ g
$\alpha = 0{,}01$ test unilatéral
Rejeter H_0 si $z > 2{,}33$



$$z = \frac{1\,470 - 1\,400}{180 / \sqrt{50}} = 2{,}7.$$

H_0 est rejetée donc H_1 acceptée (la diète est efficace).

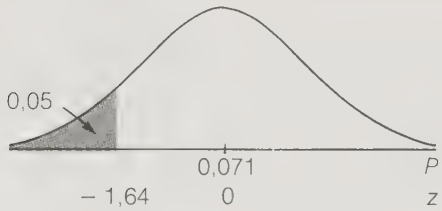
Exemple 2

Un économiste qui travaille pour le ministère du travail sait que le mois dernier il y avait un taux de chômage de 7,1%. Ce mois-ci, il trouve que sur un échantillon aléatoire de 5 000 travailleurs il y a 350 chômeurs. Au niveau de signification de 0,05, est-ce que le taux de chômage a diminué ce mois-ci?

$H_0 : P = 0{,}071$; $H_1 : P < 0{,}071$; $\alpha = 0{,}05$; test unilatéral.

Rejeter H_0 si $z < -1{,}645$

$$z = \frac{0{,}070 - 0{,}071}{\sqrt{0{,}071 * 0{,}929/5\,000}} = -0{,}28$$



On accepte H_0 : le taux de chômage n'a pas changé de façon significative ce mois-ci.

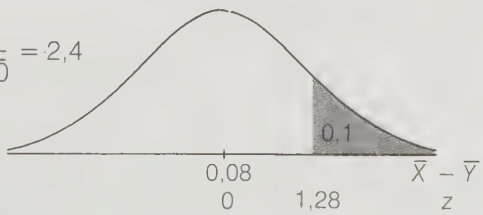
Exemple 3

Le régistraire de l'Université Laval veut comparer les résultats des étudiants mariés avec ceux des célibataires. Il choisit au hasard 100 étudiants mariés et trouve que leur résultat moyen est de 2,85 avec un écart-type de 0,4. En procédant de la même façon avec 100 célibataires, il trouve une moyenne de 2,73 avec un écart-type de 0,3. Au niveau de signification de 0,10, est-ce que les étudiants mariés ont des meilleurs résultats que les célibataires?

$H_0 : \mu_X - \mu_Y = 0$; $H_1 : \mu_X - \mu_Y > 0$; $\alpha = 0{,}10$; test unilatéral.

H_0 est rejetée si $z > 1{,}28$

$$z = \frac{(2{,}85 - 2{,}73) - 0}{\sqrt{0{,}4^2 / 100 + 0{,}3^2 / 100}} = 2{,}4$$



H_0 est rejetée, les étudiants mariés ont des meilleurs résultats.

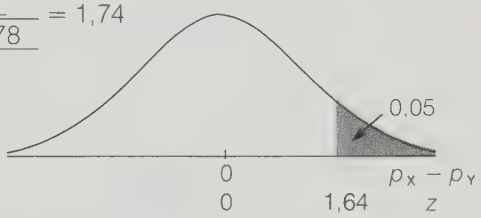
Exemple 4

Un politicologue veut connaître la différence de proportions entre les libéraux qui sont convaincus et ceux qui ne le sont pas. (Les convaincus sont membres du parti, les autres pas.) Il choisit au hasard un échantillon de 1 000 votants dont les familles ont un revenu supérieur à $14 000 et 1 000 votants dont les familles ont un revenu inférieur à $8 000. Dans le premier groupe, il en trouve 25,3% qui sont membres du parti, alors que dans le second, 22,0% sont membres. Au niveau de signification de 0,05 est-ce que la proportion des membres du parti libéral est supérieure dans le premier groupe en regard du second?

$H_0: P_X - P_Y = 0$; $H_1: P_X > P_Y$; $\alpha = 0{,}05$; test unilatéral.

H_0 est rejetée si $z > 1{,}64$

$$z = \frac{0{,}253 - 0{,}22}{\sqrt{\frac{0{,}253 * 0{,}747}{1\,000} + \frac{0{,}22 * 0{,}78}{1\,000}}} = 1{,}74$$



H_0 est rejetée; il y a plus de membres du parti libéral dans le premier groupe que dans le second.

11.6 ERREURS DE 1re ESPÈCE ET DE 2e ESPÈCE

Exemple 1

Un accusé qui doit être jugé peut être coupable ou non coupable. Le juré peut le déclarer coupable ou non coupable. Les quatre possibilités peuvent être illustrées de la façon suivante:

		L'accusé est ?	
		Non coupable	Coupable
Le juré le déclare ?	Non coupable	*A*	*B* (2e espèce : β)
	Coupable	*C* (1re espèce: α)	*D*

Justice est rendue si les événements *A* et *D* se produisent: on commet une erreur si *B* et *C* se produisent. L'hypothèse nulle est que l'accusé est non coupable. L'événement *C* (rejet de l'hypothèse nulle quand celle-ci est vraie) nous conduit à une erreur de 1re espèce; α est la probabilité d'une telle erreur.

L'événement B (acceptation de l'hypothèse nulle quand elle est fausse) est une erreur de 2e espèce: β est la probabilité de cette erreur.

Le procureur de la défense ne sera pas très heureux si le juré fait une erreur de 1re espèce. Il préférera avoir un α petit et ne sera pas fâché si β est grand. D'un autre côté le procureur de la couronne préférera que la probabilité que le juré trouve l'accusé non coupable, quand celui-ci a commis le crime, soit petite (le procureur de la couronne préférera β petit).

Un manufacturier d'appareils de télévision doit rejeter la dernière livraison de tubes images, parce que, après avoir vérifié un échantillon, il trouve que leur vie moyenne est trop courte. Il fera une erreur de 1re espèce, si ce n'est pas le cas; il peut avoir des problèmes de production et peut avoir dans le futur des relations difficiles avec son fournisseur. D'un autre côté, s'il accepte la livraison et que effectivement elles ont une vie moyenne inférieure aux normes (erreur de 2e espèce), il peut avoir des difficultés à respecter la garantie qu'il donne aux clients.

Ces exemples avaient pour but d'illustrer que les erreurs de 1re et de 2e espèces dépendent des hypothèses à vérifier. Dans notre convention qui consiste à vérifier H_0 en utilisant une statistique d'échantillonnage, seule l'erreur de 1re espèce nous importe (et la probabilité est donnée par le niveau de signification α) β (l'erreur de 2e espèce) est ignorée. Quelques exemples illustreront la relation qui existe entre α et β. En étudiant ces exemples il faudra se rappeler les résultats suivants:

1. Si α augmente (ou diminue), β diminue (ou augmente), la relation entre α et β n'est pas linéaire.
2. Pour une valeur de α, β est facilement déterminé si H_1 est de la forme $\mu = ..., P = ...$, etc. Si H_1 est de la forme $\mu < ..., P \neq ...$, etc., on ne peut déterminer β.
3. Pour une valeur fixe de α, la valeur de β peut être diminuée en augmentant la taille de l'échantillon.

Exemple 2

Un manufacturier annonce que ses ampoules électriques ont une vie moyenne de 1 000 heures avec un écart-type de 40 heures. Un gros détaillant vérifie un échantillon aléatoire de 64 lumières prises dans la dernière livraison.

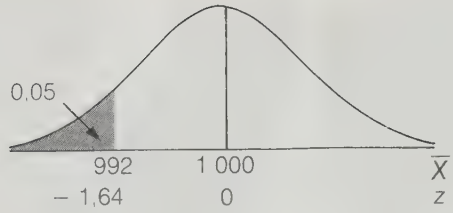
(a) Trouver la valeur minimale de vie moyenne qui permet d'accepter la livraison si le niveau de signification est (i) 0,05 (ii) 0,01.

(b) Quelle est la probabilité d'accepter la livraison, si de fait la vie moyenne des lumières est de 987 heures avec un niveau de signification de (i) $\alpha = 0{,}05$ (ii) $\alpha = 0{,}01$? (Supposer que l'écart-type demeure 40 heures.)

Réponse

(a) $H_0: \mu = 1000$; $H_1: \mu < 1000$; test unilatéral

(i) $\alpha = 0{,}05$; alors $z = -1{,}64$; $\sigma_{\bar{x}} = \dfrac{40}{\sqrt{64}} = 5$.

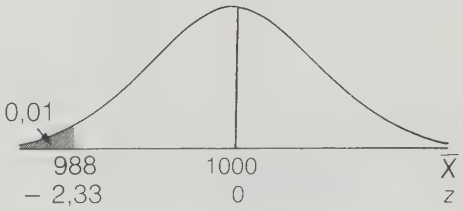


$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}; -1{,}64 = \frac{\overline{X} - 1000}{5}; \overline{X} = 992$$

La livraison est acceptée si $\overline{X} > 992$.

(ii) $\alpha = 0{,}01$; alors $z = -2{,}33$.

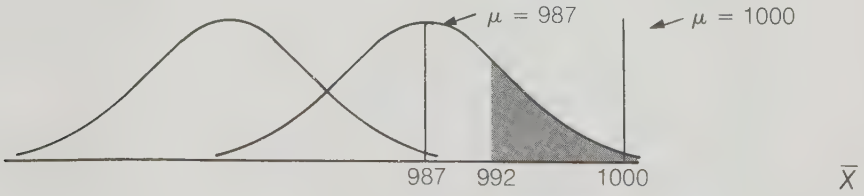
$$-2{,}33 = \frac{\overline{X} - 1000}{5}; \overline{X} = 988.$$

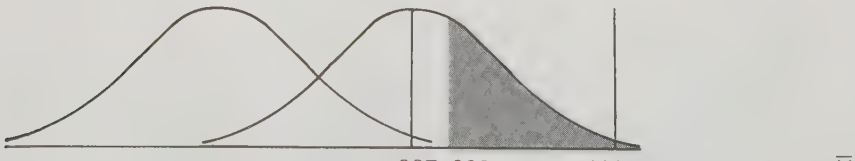


La livraison est acceptée si $\overline{X} > 988$.

Il est à noter que jusqu'à maintenant nous avions trouvé la valeur critique de z puis nous trouvions la valeur statistique pour savoir si celle-ci tombait dans la zone d'acceptation ou non. Ici la valeur critique est exprimée non pas en terme de z mais en fonction de la variable $\overline{X}$.

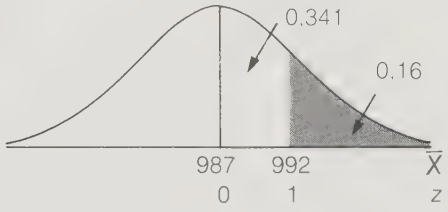
(b) $H_0 : \mu = 1\,000, H_1 : \mu = 987$. Les distributions d'échantillonnages des moyennes correspondant à ces deux hypothèses sont normalement distribuées et se chevauchent comme ci-dessous.



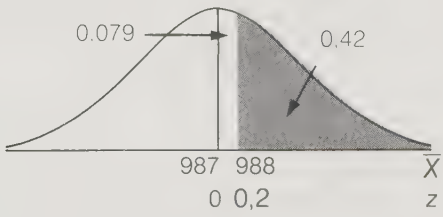


La probabilité d'accepter la livraison quand la vie moyenne des lumières est de 987 heures est représentée par l'aire ombrée dans la figure (a) quand $\alpha = 0{,}05$ et par la figure (b) quand $\alpha = 0{,}01$. Cette aire ombrée représente la probabilité d'une erreur de 2e espèce β si l'hypothèse nulle dit que $\mu = 1\,000$ et H_1 est $\mu = 987$.

(i) Si $\alpha = 0{,}05$, β est l'aire ombrée de la figure (a) à droite de $\overline{X} = 992$. Comme $\sigma_{\overline{x}} = 5$ et que la moyenne de cette courbe est 987, un résultat $\overline{X}$ de 992 correspond à un résultat z de $\dfrac{992 - 987}{5} = 1$; alors $\beta = 0{,}5 - 0{,}341 = 0{,}16$.



(ii) Si $\alpha = 0{,}01$, β est l'aire ombrée à droite de 988 dans la figure (b). Un résultat de 988 correspond à un résultat z de $\dfrac{988 - 987}{5} = 0{,}2$ donc $\beta = 0{,}5 - 0{,}079 = 0{,}42$.



α	β
0,05	0,16
0,01	0,42

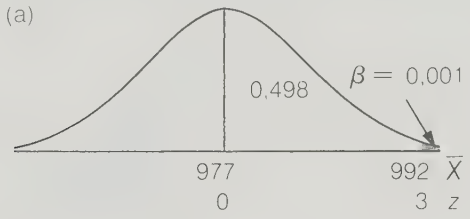
Remarquer que β augmente quand α diminue.

Exemple 3

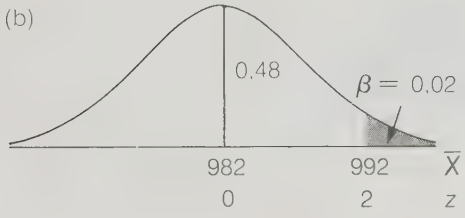
Le manufacturier affirme toujours que ses ampoules électriques ont une vie moyenne de 1 000 heures avec un écart-type de 40 heures. Le détaillant vérifie un échantillon aléatoire de 64 ampoules électriques avec un niveau de signification de 0,05. Calculer β si en réalité la vie moyenne des ampoules électriques est (a) 977 (b) 982 (c) 987 (d) 992 (e) 997 (f) 1 002 heures.

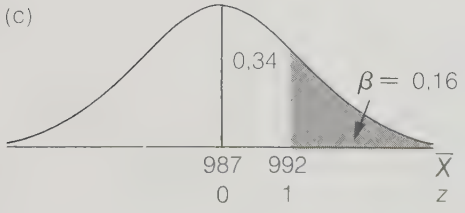
Réponses

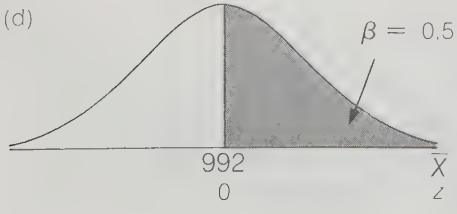
(a) À un niveau de signification de 0,05, le détaillant accepte la livraison si la vie moyenne de l'échantillon est plus grande que 992 heures (voir exemple précédent). Mais si la vraie moyenne de la population est 977, alors 992 correspond à un résultat z de $\frac{992 - 977}{5} = 3{,}0$ et $\beta = 0{,}5 - 0{,}499 = 0{,}001$.

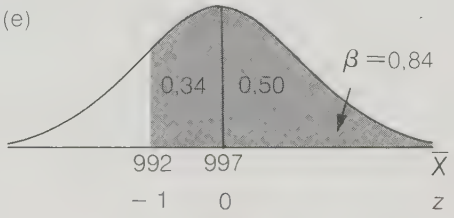


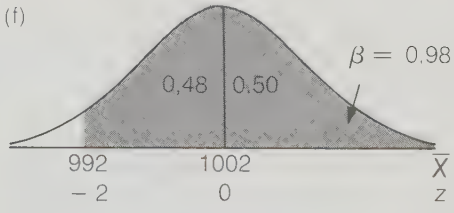
Les autres parties de ce problème sont résolues de la même manière, nous ne donnerons ici que les diagrammes.





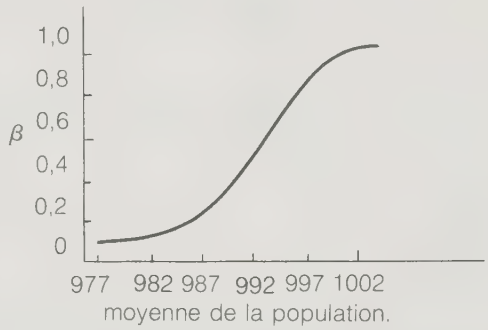






On peut résumer les résultats précédents en traçant un graphique qui montre la probabilité d'une erreur de 2e espèce β pour différentes hypothèses H_1 si H_0 est $\mu = 1\,000$ et $\alpha = 0{,}05$. C'est la courbe caractéristique du test.

H_1: $\mu =$	β
977	0,001
982	0,02
987	0,16
992	0,50
997	0,84
1002	0,98



On voit par cette courbe que la probabilité d'une erreur de 2e espèce est près de 1 si la vraie moyenne de la population est supérieure à 997, mais diminue rapidement si la moyenne est inférieure à 997; β est presque 0 si la vraie moyenne de la population est inférieure à 977.

La valeur de β dépend alors de l'hypothèse rivale qu'on acceptera si H_0 est rejetée. En pratique, on utilise plus fréquemment α car on rencontre plus souvent des hypothèses rivales de la forme H_1: $\mu > 100$ ce qui élimine l'utilisation de β. Un exemple, un nouveau médicament est rejeté seulement si ce n'est pas une amélioration marquée plutôt qu'acceptée parce que c'est peut-être une amélioration.

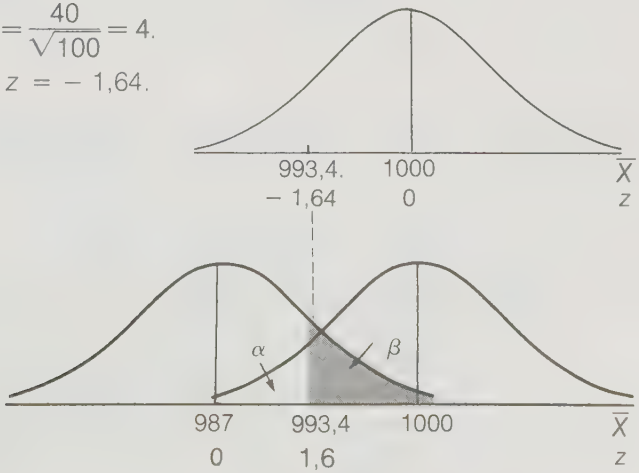
Exemple 4

Le manufacturier d'ampoules électriques affirme toujours que la vie moyenne de ses ampoules électriques est 1 000 heures avec un écart-type de 40 heures. Si $\alpha = 0{,}05$, trouver β pour l'hypothèse alternative $\mu = 987$ si la taille de l'échantillon est (a) 64 (b) 100.

(a) Voir exemple 3b précédent: $\beta = 0{,}16$ si $n = 64$.

(b) Si $n = 100$, $\sigma_{\bar{x}} = \dfrac{40}{\sqrt{100}} = 4$.

$\alpha = 0{,}05$ donc $z = -1{,}64$.



Comme $-1{,}64 = \dfrac{\overline{X} - 1000}{4}$, $\overline{X} = 993{,}4$. H_0: $\mu = 1\,000$ est rejetée si la moyenne de l'échantillon de taille 100 est inférieure à 993,4. Si H_1 dit que $\mu = 987$ alors $\beta = p\ (\overline{X} > 993{,}4$ étant donné que $\mu = 987)$
$= p\ (z\ \dfrac{993{,}4 - 987}{40/\sqrt{100}} = 1{,}6) = 0{,}5 - 0{,}445 = 0{,}06.$

Pour un $\alpha = (0{,}05)$ fixe, β devient plus petit si n augmente: si $n = 64$, $\beta = 0{,}16$; si $n = 100$, $\beta = 0{,}06$. Comment déterminer β si on a un test bilatéral?

Exemple 5

Un manufacturier de résistances électriques affirme que l'item P36483E a une résistance moyenne de 1 000 ohms avec un écart-type de 40 ohms. Un technicien en réparation de radios qui a besoin de cet item vérifie un échantillon aléatoire de 64 pris dans sa dernière livraison, avec un niveau de signification de 0,05.

(a) Quelles sont les limites d'acceptation de la livraison?

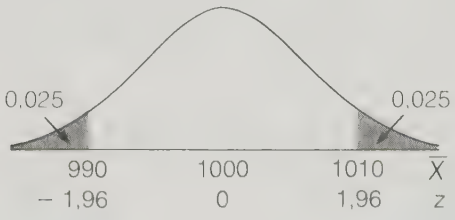
(b) Quelle est la probabilité d'une erreur de 2e espèce si la résistance moyenne de tous les items P36483E est de 980 ohms?

(c) Tracer la courbe de β en fonction de μ (courbe caractéristique de la vérification).

(a) H_0: $\mu = 1\,000$, H_1: $\mu \neq 1\,000$; $\sigma = 40$; $n = 64$; $\alpha = 0{,}05$, donc $z = \pm 1{,}96$

$\pm 1{,}96 = \dfrac{\overline{X} - 1000}{40/\sqrt{64}}$; $\overline{X} = 990$ ou 1010.

H_0 est acceptée si $\overline{X}$ est entre 990 et 1010.



(b) Si H_1 est $\mu = 980$ [ou 985 ou ... 1 020], quelle est la probabilité qu'un échantillon aléatoire de taille 64 aura une moyenne comprise entre 990 et 1 010?

(i) $\mu = 980$; $\sigma_{\overline{x}} = 40/\sqrt{64} = 5$; $p\ (990 < \overline{X} < 1\,010) = p\ (2 < z < 6) = 0{,}5 - 0{,}4772 = 0{,}02$;

(ii) $\mu = 985$; $\sigma_{\overline{X}} = 5$; $p\ (990 < \overline{X} < 1\,010) = p\ (1 < z < 5) = 0{,}5 - 0{,}341 = 0{,}16$;

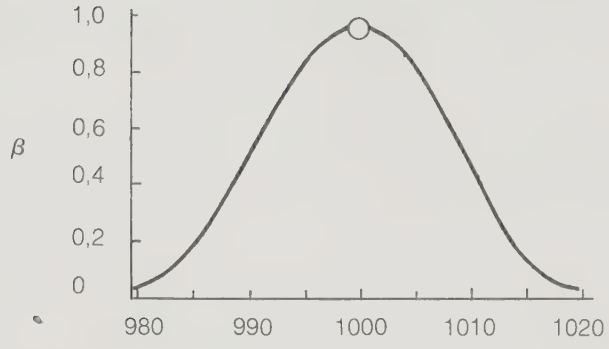
(iii) $\mu = 990$; $\sigma_{\overline{X}} = 5$; $(990 < \overline{X} < 1\,010) = p\ (0 < z < 4) = 0{,}50$;

(iv) $\mu = 995$; $\sigma_{\overline{X}} = 5$; $p\ (990 < \overline{X} < 1\,010) = p\ (-1 < z < 3) = 0{,}341 + 0{,}499 = 0{,}84$;

(v) $\mu = 999$; $p\ (990 < \overline{X} < 1\,010) = p\ (-1{,}8 < z < 2{,}2) = 0{,}464 +$

0,486 = 0,95
(vi) $\mu = 1\,005$; $p\ (990 < \bar{X} < 1\,010) = p\ (-3 < z < 1) = 0{,}341 + 0{,}499 = 0{,}84$

(c) Voici la courbe caractéristique du test:



H_1: $\mu =$	β
980	0,02
985	0,16
990	0,50
995	0,84
999	0,95
1005	0,84
1010	0,50
1015	0,16
1020	0,02

11.7 AVERTISSEMENTS

1. Selon notre code civil un accusé qui est acquitté par un juré ne peut être retraduit devant les tribunaux pour le même délit. Ce n'est pas toujours le cas avec une vérification d'hypothèse. Si un manufacturier affirme que la vie moyenne de ses lumières est de 1 000 heures, avec $\sigma = 40$ heures, une vérification sur un échantillon aléatoire de 64 lumières peut montrer une vie moyenne de 998 heures, résultat qui est en accord avec l'hypothèse à un niveau de signification de 0,05. Un tel résultat ne prouve pas que l'hypothèse nulle est vraie, mais seulement que l'hypothèse est probablement vraie.

2. Dans une vérification d'hypothèse, il faut que l'échantillon qui doit être vérifié soit réellement aléatoire, sinon les résultats peuvent être faussés et aucune conclusion n'est valide.

3. On suppose que la distribution d'échantillonnage de la statistique est connue. La distribution est normale si σ est connu ou encore que $n > 30$. Dans le prochain chapitre nous décrirons un autre type de distribution si σ est inconnu et n petit.

4. Se rappeler que H_0, H_1 et α sont déterminés avant que la vérification soit effectuée et que la forme de H_1 détermine si on fait un test unilatéral ou bilatéral.

11.8 VOCABULAIRE ET SYMBOLES

Hypothèse nulle	H_0	Test unilatéral	
Hypothèse rivale	H_1	Test bilatéral	
Niveau de signification	α	Courbe caractéristique du test	
Erreur de 1re espèce	α	Erreur de 2e espèce	β

11.9 EXERCICES

1. Une machine remplit des litres de lait; la quantité versée dans chaque

contenant doit être de 1 000 cm^3 avec un écart-type de 6 cm^3. Lors d'un examen journalier permettant de vérifier l'ajustement de la machine, 36 contenants sont choisis au hasard et on trouve qu'ils contiennent en moyenne 1 003 cm^3. La machine est-elle bien ajustée au niveau de signification de 0,05 (voir exemple 4, section 11.2)?

2. On sait qu'un chien d'un an gagne en moyenne 450 g par mois avec un écart-type de 180 g. Une diète spéciale riche en protéines est donnée à 50 chiens d'un an choisis au hasard et leur augmentation moyenne de poids pour un mois fut de 1 470 g avec un écart-type de 140 g. La diète spéciale favorise-t-elle l'augmentation de poids au niveau de signification de 0,01?

3. Un économiste qui travaille au ministère du travail sait que le mois-passé, le taux de chômage était de 7,1%. Ce mois-ci, il découvre que sur 5 000 travailleurs choisis au hasard, 350 sont sans emploi. Au niveau de signification de 0,05, le taux de chômage a-t-il varié?

4. Le registraire de l'Université Laval veut comparer les résultats des étudiants mariés avec ceux des célibataires. Il choisit au hasard 100 étudiants mariés et trouve que leur résultat moyen est 2,85 avec un écart-type de 0,4. En procédant de la même façon avec 100 célibataires, il trouve une moyenne de 2,73 avec un écart-type de 0,3. Au niveau de signification de 0,10, est-ce que les étudiants mariés ont des meilleurs résultats que les célibataires?

5. Un politicologue veut connaître la différence de proportions entre les libéraux qui sont convaincus et ceux qui le sont pas. (Les convaincus sont membres du parti, les autres pas.) Il choisit au hasard un échantillon de 1 000 votants dont les familles ont un revenu supérieur à $14 000 et 1 000 votants dont les familles ont un revenu inférieur à $8 000. Dans le premier groupe, il en trouve 25,0% qui sont membres du parti alors que dans le second 22,0% sont membres. Au niveau de signification de 0,05, est-ce que la proportion des membres du parti libéral est différente dans le premier groupe en regard du second?

6. Le directeur du département de français du collège Montréal trouve qu'un échantillon aléatoire de 100 étudiants a obtenu une moyenne de 470 lors d'un examen de français. A l'échelle provinciale, le même examen a donné une moyenne de 500 avec un écart-type de 100. Est-ce que les étudiants de Montréal ont un résultat qui diffère sensiblement du résultat provincial au niveau de signification de 0,01?

7. La scolarité moyenne des adultes au Canada est de 10,4 ans avec un écart-type de 2 ans. Le ministère de l'Éducation de l'Ontario fait l'étude d'un échantillon aléatoire de 200 adultes de sa province et trouve que la scolarité moyenne est de 11,3 ans avec un écart-type de 1,8 an. Au niveau de signification de 0,05, est-ce que la scolarité des Ontariens diffère de celle des Canadiens?

8. Est-ce que la proportion des étudiants qui demandent une bourse d'étude est la même pour les garçons que pour les filles? Le registraire d'une université choisit au hasard les dossiers de 60 garçons et 50 filles et trouve que 36 garçons et 35 filles ont fait une demande de bourse. Quelle sera sa conclusion au niveau de signification de 0,01?

9. Douze des 100 commerces choisis au hasard dans l'ouest furent cambriolés en 1976 alors que 24 des 150 commerces choisis au hasard dans l'est le furent. Au niveau de signification de 0,10, y a-t-il eu plus de cambriolages dans l'est?

10. Une analyse faite sur 100 cigarettes de marque Sport Plus choisies au hasard

a donné une moyenne de 17,2 mg de goudron avec un écart-type de 2 mg. (a) Ce résultat est-il en accord avec la publicité qui affirme que la cigarette Sport Plus contient 18 mg de goudron (niveau de signification de 0,05)? (b) Ce résultat permet-il d'affirmer que la cigarette Sport Plus contient au plus 17 mg de goudron ($\alpha = 0{,}05$)?

11. En 1970, 25% des appels pour incendies furent des fausses alertes. Cette année, sur 150 alertes choisies au hasard, 42 furent fausses. Est-ce que la proportion des fausses alertes a augmenté depuis 1970 ($\alpha = 0{,}01$)?

12. Une épidémie de dysenterie a obligé plusieurs employés de la compagnie ABC à s'absenter en janvier. Sur un échantillon aléatoire de 400 employés on a découvert que 150 employés qui avaient reçu un vaccin se sont absentés en moyenne 2,4 jours avec un écart-type de 0,3 alors que les 250 autres qui n'avaient pas reçu de vaccin se sont absentés en moyenne 6,5 jours avec un écart-type de 0,6 jour. Est-ce que le vaccin a diminué au moins de 2 jours le nombre moyen de jours perdus par cette maladie ($\alpha = 0{,}05$)?

13. On croit que le salaire moyen des professeurs des collèges publics est de $1 000 supérieur à celui des professeurs des collèges privés. Un échantillon aléatoire de 100 professeurs de collèges publics montre que ceux-ci reçoivent un salaire moyen de $15 000 avec un écart-type de $1 000 alors qu'un échantillon de 100 professeurs de collèges privés choisis au hasard montre que ceux-ci reçoivent un salaire moyen de $13 800 avec un écart-type de $900. Au niveau de signification de 0,01, est-ce que le salaire moyen des professeurs de collèges publics est au moins $1 000 supérieur à celui des professeurs de collèges privés?

14. Les super-marchés du Québec vendent en moyenne 460 l de lait par jour. Un échantillon de 36 jours choisis au hasard montre qu'un super-marché a vendu en moyenne 470 l de lait par jour avec un écart-type de 20 l. Ce super-marché vend-il plus de lait que la moyenne du Québec au niveau de signification de 0,05?

15. Un groupe de 40 étudiants de 1re collégiale et un groupe de 40 étudiants de 2e collégiale passent un examen de géographie avec les résultats suivants:

1re collégiale		2e collégiale	
résultats	*f*	résultats	*f*
90–99	8	90–99	7
80–89	12	80–89	8
70–79	13	70–79	15
60–69	7	60–69	10

Les deux groupes étant choisis au hasard, les étudiants de 2e année ont-ils moins bien réussi que ceux de 1re avec un niveau de signification de 0,01?

16. Un fabricant de produits pharmaceutiques affirme qu'il a un médicament qui soulage des symptômes du rhume pour une période de 10 heures dans 90% des cas. Dans un échantillon aléatoire de 400 personnes qui ont eu le rhume, 350 affirment avoir été soulagées pour une période de 10 heures. Au niveau de signification de 0,05, l'affirmation du manufacturier est-elle exacte?

17. On soupçonne un dé de ne pas être régulier. On le lance 1 200 fois et on obtient 230 fois la face 3; au niveau de signification de 0,05, nos soupçons sont-ils justifiés?

18. Un examen de lecture a été donné à des échantillons aléatoires de 200 garçons et 100 filles de 3e année. Les garçons ont obtenu un résultat moyen de 60 avec un écart-type de 10 alors que les filles ont obtenu 65 avec un écart-type de 8. Les garçons lisent-ils moins bien que les filles au niveau de signification de 0,01?

19. Des échantillons aléatoires de 400 axes produits par la machine A et 300 axes produits par la machine B donnent respectivement 10 et 3 axes défectueux. Au niveau de signification de 0,05,

(a) les deux machines diffèrent-elles dans la proportion de production des axes défectueux?

(b) est-ce que la machine A produit une plus grande proportion d'axes défectueux que la machine B?

20. En 1976, une famille québécoise de 4 personnes dépensait en moyenne $68 par semaine pour se nourrir avec un écart-type de $6. Au premier trimestre de 1977, un échantillon aléatoire de 50 familles dépensaient en moyenne $72 par semaine avec un écart-type de $5. Est-ce que la dépense hebdomadaire moyenne a augmenté au niveau de signification de 0,05?

● 21. En 1971, il y eut 54 700 accidents mortels de la route aux États-Unis, en 1972, 56 600. En Orégon en 1971, il y eut 695 accidents mortels et 734 en 1972. Au niveau de signification de 0,01, est-ce que le taux de mortalité en Orégon a varié par rapport à celui des États-Unis?

22. En 1960, un échantillon aléatoire de 1 000 familles canadiennes ayant au moins un enfant de moins de 18 ans révèle que 100 familles avaient 4 enfants ou plus. On reprend la même étude en 1970 et on trouve que, sur 1 000 familles, 98 ont 4 enfants ou plus. Au niveau de signification de 0,01, est-ce que la proportion des familles ayant 4 enfants ou plus au Canada a diminué pendant cette période?

23. Un train de passagers transporte en moyenne par jour 12 000 personnes avec un écart-type de 2 000 personnes. Des modifications furent apportées au train et pendant l'année qui suivit on a choisi 30 jours au hasard. On trouve une moyenne par jour de 12 500 passagers avec un écart-type de 1 000. Est-ce que le nombre moyen de passagers par jour a augmenté avec un niveau de signification de 0,02?

24. Sur une longue période de temps, le pourcentage de «A» obtenu en Mathématiques 103 au collège Laval fut de 20%. Lors de la dernière session d'automne sur 100 étudiants il y eut 24 «A». La proportion de A a-t-elle augmenté de façon significative ($\alpha = 0{,}05$)?

25. Une nouvelle cafétéria est aménagée à la polyvalente Rousseau. Le responsable croit que 1 800 étudiants peuvent prendre un repas entre 17:30 et 19:00 si le temps moyen que prend chaque étudiant pour manger est de 30 minutes ou moins. Il chronomètre 64 étudiants choisis au hasard et trouve en moyenne 32 minutes pour manger avec un écart-type de 6 minutes. Le responsable pourra-t-il conclure que la cafétéria pourra servir 1 800 étudiants dans le temps prévu ($\alpha = 0{,}02$)?

26. Un terrain est divisé en 60 lots de mêmes grandeur, fertilité, humidité, etc. Un nouveau fertilisant «RAPIDE» est appliqué sur 30 lots choisis au hasard, alors que les autres sont fertilisés avec du 5-10-5. On sème le même nombre de graines de tomates dans chaque lot. On récolte en moyenne 65 kg de tomates avec un écart-type de 5 kg des lots qui furent fertilisés avec du «RAPIDE» alors qu'avec du 5-10-5, on a récolté en moyenne 62 kg avec un écart-type de 4 kg.

(a) Y a-t-il une différence entre le poids moyen des tomates provenant des lots fertilisés au «RAPIDE» et ceux fertilisés au 5-10-5 ($\alpha = 0{,}05$)?

(b) Les lots fertilisés au «RAPIDE» produisent-ils un poids moyen plus grand au niveau de signification de 0,05?

27. À l'échelle québécoise, le résultat moyen des examens d'admission à l'Université est de 75 avec un écart-type de 8. Un échantillon aléatoire de 144 étudiants de l'Université de Sherbrooke donne un résultat moyen de 76 avec un écart-type de 8.

(a) Au niveau de signification de 0,05 les étudiants de Sherbrooke sont-ils supérieurs à la moyenne québécoise?

(b) Quelle est la probabilité d'accepter $H_0 : \mu = 75$ quand la moyenne des étudiants de Sherbrooke est actuellement (i) 78 (ii) 77 (iii) 76 (iv) 75 (v) 74?

(c) Tracer la courbe caractéristique du test.

28. Un adepte de la plongée sous-marine découvre un coffre rempli de pièces d'or et d'argent. Il choisit au hasard avec remise 10 pièces et adopte la règle de décision suivante: il accepte l'hypothèse que la moitié des pièces sont en or si son échantillon contient entre 3 et 7 pièces en or, et la rejette dans le cas contraire.

(a) Trouver la probabilité de rejeter l'hypothèse quand elle est vraie.

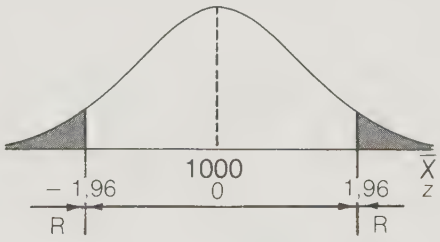
(b) Trouver la probabilité d'accepter l'hypothèse que la moitié des pièces sont en or quand la vraie proportion des pièces d'or est (i) 0,1 (ii) 0,2 (iii) 0,3 (iv) 0,4 (v) 0,6.

(c) Tracer la courbe de β en fonction de P.

RÉPONSES

1. $H_0 : \mu = 1\,000$; $H_1 : \mu \neq 1\,000$, $\alpha = 0{,}05$; H_0 est acceptée si z est entre $-1{,}96$ et $+1{,}96$.

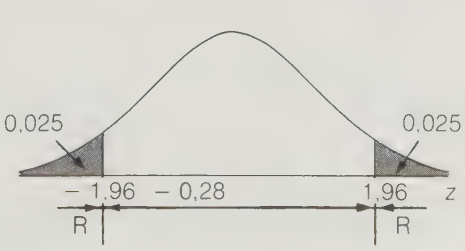
$z = \dfrac{1\,003 - 1\,000}{6/\sqrt{36}} = 3$; H_0 est rejetée; la machine doit être ajustée.



3. $H_0 : P = 0{,}071$; $H_1 : P \neq 0{,}071$; $\alpha = 0{,}05$; H_0 est rejetée si $z > 1{,}96$ ou si $z < -1{,}96$;

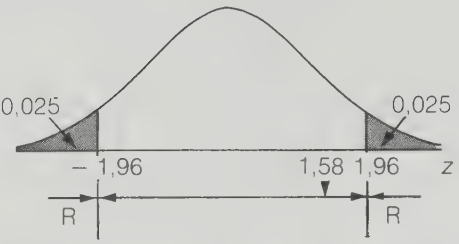
$z = \dfrac{0{,}070 - 0{,}071}{\sqrt{0{,}071 * 0{,}929/5000}} = -0{,}28.$

H_0 est acceptée; comparer avec exemple 2, section 11.5.



5. $H_0 : P_X - P_Y = 0$; $H_1 : P_X - P_Y \neq 0$; $\alpha = 0{,}05$; H_0 est rejetée si $z > 1{,}96$ ou si $z < -1{,}96$.

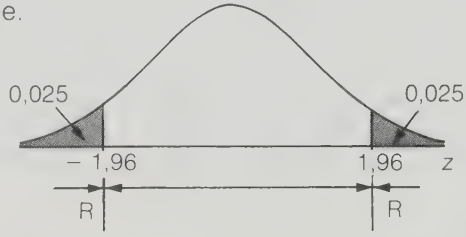
$$z = \frac{(0{,}25 - 0{,}22) - 0}{\sqrt{\frac{0{,}25 * 0{,}75}{1\,000} + \frac{0{,}22 * 0{,}78}{1\,000}}} = 1{,}58$$



Il n'y a pas de différence dans la proportion de libéraux.

7. $H_0 : \mu = 10{,}4$; $H_1 : \mu \neq 10{,}4$; $\alpha = 0{,}05$;

$z = \dfrac{11{,}3 - 10{,}4}{2/\sqrt{200}} = 6{,}36$; H_0 est rejetée.



9. O = Ouest E = Est $H_0 : P_0 = P_E$; $H_1 : P_0 < P_E$; $\alpha = 0.10$ donc $Z = -1{,}28$

$$P_0 = \frac{12}{100} = 0{,}12 \qquad P_E = \frac{24}{150} = 0{,}16$$

$$\sigma_{(P_0 - P_E)} = \sqrt{\frac{0{,}12 * 0{,}88}{100} + \frac{0{,}16 * 0{,}84}{150}} = 0{,}044$$

$z = \dfrac{(0{,}12 - 0{,}16) + 0}{0{,}044} = -0{,}91$ qui est dans la région d'acceptation; il n'y a pas plus de cambriolages dans l'Est que dans l'Ouest.

11. $H_0 : P = 0{,}25$; $H_1 : P > 0{,}25$; $p = 0{,}28$; $\sigma_p = \sqrt{\dfrac{0{,}25 * 0{,}75}{150}} = 0{,}0354$;

$z = \dfrac{0{,}28 - 0{,}25}{0{,}0354} = 0{,}85$

$< 2{,}33$; H_0 est acceptée; la proportion de fausses alarmes n'a pas augmenté de façon significative.

13. X = collèges publics, Y = collèges privés. $H_0 : \mu_X - \mu_Y = 1\,000$; $H_1 : \mu_X - \mu_Y > 1\,000$. $\alpha = 0{,}01$, $Z = +2{,}33$.

$$\sigma_{(\bar{X} - \bar{Y})} = \sqrt{\frac{1000^2}{100} + \frac{900^2}{100}} = 134{,}54$$

$$z = \frac{(15\,000 - 13\,800) - 1000}{134{,}54} = 1{,}49$$

H_0 est acceptée; les professeurs des collèges publics ne gagnent pas au moins \$1 000 de plus que ceux des collèges privés.

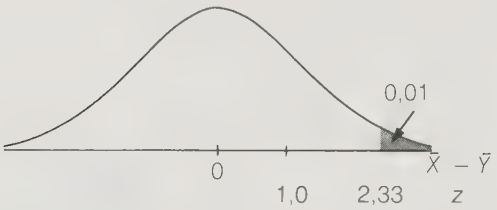
15. Moyenne de 1re collégiale $= \overline{X} = 79{,}8$, $s_X^2 = 102{,}5$.
Moyenne de 2e collégiale $= \overline{Y} = 77{,}5$, $s_Y^2 = 108{,}7$
$H_0 : \mu_X = \mu_Y$ ou $\mu_X - \mu_Y = 0$; $H_1 : \mu_X > \mu_Y$; $\alpha = 0{,}01$, test unilatéral.
H_0 est rejetée si $z > 2{,}33$.

$$n_X = 40;\ \sigma_{\overline{X}}^2 = \frac{102{,}5}{40} = 2{,}56$$

$$n_Y = 40,\ \sigma_{\overline{Y}}^2 = \frac{108{,}7}{40} = 2{,}72$$

$$\sigma_{(X-Y)} = \sqrt{2{,}56 + 2{,}72} = 2{,}30$$

$$z = \frac{(79{,}8 - 77{,}5) - 0}{2{,}30} = 1{,}00$$

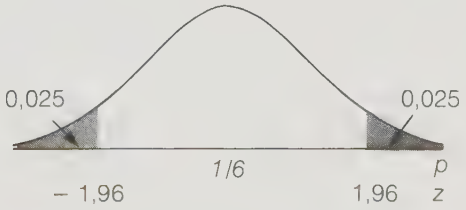


On accepte H_0. Les étudiants de 2e année n'ont pas moins bien réussi que ceux de 1re année.

17. $H_0 : P = 1/6$; $H_1 : P \neq 1/6$; $\alpha = 0{,}05$, test bilatéral. H_0 rejetée si $z > 1{,}96$ ou si $z < -1{,}96$.

$$p = 230/1200;\ \sigma_p = \sqrt{\frac{1/6 * 5/6}{1200}} = 0{,}011$$

$$z = \frac{230/1200 - 1/6}{.011} = 2{,}3$$



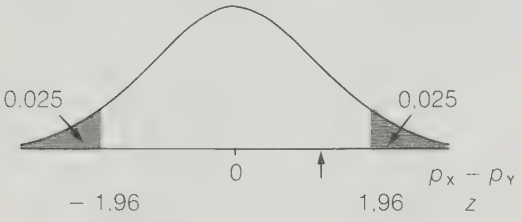
H_0 rejetée: les soupçons étaient fondés.

19. (a) $H_0 : P_X = P_Y$; $H_1 : P_X \neq P_Y$; $\alpha = 0{,}05$; test bilatéral, H_0 est rejetée si $z > 1{,}96$ ou si $z < -1{,}96$.

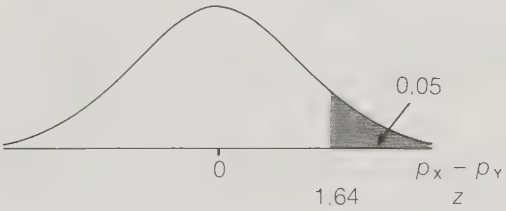
$$\sigma_{p_X}^2 = \frac{10/400 * 390/400}{400} = 0{,}000061; \sigma_{p_Y}^2 = \frac{3/300 * 297/300}{300} = 0{,}000033$$

$$\sigma_{(p_X - p_Y)} = \sqrt{0{,}000061 + 0{,}000033} = 0{,}0097; \quad P_X = 0{,}025, P_Y = 0{,}010$$

$$z = \frac{(0{,}025 - 0{,}010) - 0}{0{,}0097} = 1{,}55$$



Il n'y a pas de différence entre les deux machines au niveau de 0,05. Test unilatéral, H_0 est rejetée si $z > 1{,}64$; comparer avec la valeur de (a).



La machine A ne produit pas plus d'axes défectueux.

21. Ici on a un recensement et non des échantillons; l'inférence statistique n'est pas nécessaire. On peut dire qu'il y a eu une augmentation de 5,6% des accidents mortels en Orégon alors que pour l'ensemble des États-Unis elle fut de 3,5%.

23. $H_0 : \mu = 12\,000$; H_1 $\mu > 12\,000$; $\alpha = 0{,}02$, H_0 est acceptée si $z < 2{,}05$

$$z = \frac{12\,500 - 12\,000}{2000/\sqrt{30}} = 1{,}37$$

H_0 est acceptée; il n'y a pas eu une augmentation sensible.

25. $H_0 : \mu = 30$; $H_1 : \mu > 30$; $\alpha = 0{,}02$; la valeur critique de z est 2,05

$$z = \frac{32 - 30}{6/\sqrt{64}} = 2{,}67$$

H_0 est rejetée: il conclut qu'on ne peut servir 1 800 étudiants dans le temps prévu.

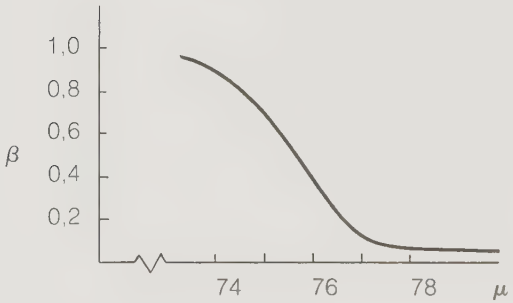
27. (a) $H_0 : \mu = 75$; $H_1 : \mu > 75$; $\sigma_x = 8/\sqrt{144} = 2/3$; $\alpha = 0{,}05$, si $z = 1{,}64$.

$$z = \frac{76 - 75}{2/3} = 3/2 = 1{,}5, < 1{,}64$$

H_0 est acceptée: les étudiants de Sherbrooke ne sont pas supérieurs.

(b) Si $\alpha = 0{,}05$ $H_0 : \mu = 75$ est acceptée si $z < 1{,}64$ ce qui implique que $\bar{X} < 76{,}1$

(i) $\beta = p(\bar{X} < 76{,}1 / \mu = 78) = p(z < -2{,}85) = 0{,}002$
(ii) $\beta = p(\bar{X} < 76{,}1 / \mu = 77) = p(z < -1{,}35) = 0{,}09$
(iii) $\beta = p(\bar{X} < 76{,}1 / \mu = 76) = p(z < 0{,}15) = 0{,}56$
(iv) $\beta = p(\bar{X} < 76{,}1 / \mu = 75) = p(z < 1{,}65) = 0{,}95$
(v) $\beta = p(\bar{X} < 76{,}1 / \mu = 74) = p(z < 3{,}15) = 1{,}00.$



Chapitre 12

Distribution *t* de Student

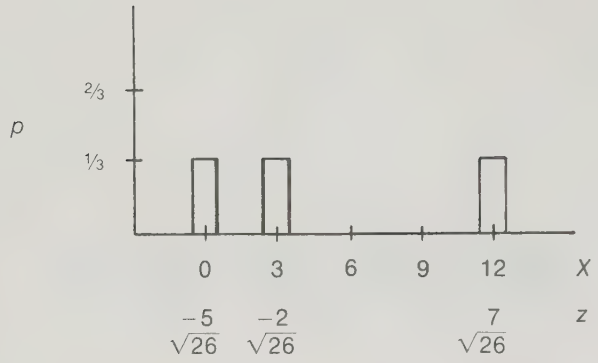
Nous savons comment faire une inférence statistique de la moyenne d'une population si la distribution de la moyenne échantillonnale est une distribution normale. Que pouvons-nous faire si σ *est inconnu et* n *est petit (la distribution n'est pas normale)? Nous répondrons à cette question dans ce chapitre.*

Nous aurons une famille de distributions t *de Student (une distribution pour chaque valeur de* n*), mais nous verrons également que si* n *augmente, la distribution* t *correspondante devient à toute fin pratique une distribution normale. Quand* n *est petit, la distribution* t *est utilisée de la même manière qu'une distribution normale.*

Nous verrons quand utiliser les cotes t *et quand utiliser les cotes z et comment utiliser la cote* t *pour une moyenne ou une différence de moyennes de populations. Finalement, nous verrons comment opérer avec des résultats jumelés provenant d'échantillons non-indépendants.*

12.1 LA NÉCESSITÉ DE LA DISTRIBUTION *t*

On doit être actuellement familier avec la distribution de la moyenne échantillonnale et savoir qu'elle est normale si (a) la population est normalement distribuée ou si (b) *n est suffisamment grand* ($n \geq 30$). Il est essentiel de se rappeler que si les moyennes des échantillons sont normalement distribuées alors leurs cotes *z* données par $\frac{\overline{X} - \mu_X}{\sigma_{\overline{X}}}$ le sont également. Si la distribution n'est pas normale alors la distribution correspondante des cotes *z* ne l'est pas. Un exemple, la distribution de probabilités des nombres 0; 3; 12 ayant la probabilité de 1/3 chacun et une moyenne de 5 avec une variance de 26 n'est pas normalement distribuée. Cette distribution a des cotes *z* de $-5/\sqrt{26}$; $-2/\sqrt{26}$; $7/\sqrt{26}$ et n'est pas distribuée normalement.



Quand il s'agit de trouver un intervalle de confiance pour la moyenne ou de vérifier une hypothèse nulle, quatre cas peuvent se présenter. On connaît déjà les trois premiers.

***1^er cas:* n est grand et σ est connu:** Les moyennes $\bar{X}$ sont normalement distribuées de même que leurs cotes $z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. En pratique n est grand si $n \geqslant 30$.

***2^e cas:* n est petit (< 30) et σ est connu:** Ici la distribution d'échantillonnage des moyennes est normale si la population est distribuée normalement (même si n est petit). (Si la population n'est pas normalement distribuée, on n'a pas le choix, il faut augmenter la taille de l'échantillon pour obtenir une inférence valide.)

Exemple 1

Des études antérieures montrent que la taille de tous les Canadiens âgés de 18 ans est normalement distribuée et que σ est toujours 6 cm mais que la taille moyenne augmente graduellement. En 1975, un échantillon aléatoire de 16 Canadiens de 18 ans a une taille moyenne de 175 cm. Trouver au niveau de 95%, un intervalle de confiance pour la taille de tous les Canadiens de 18 ans en 1975.

Au niveau de 95%, $z = \pm 1{,}96$

$$\pm 1{,}96 = \frac{175 - \mu}{6/\sqrt{16}}$$

$$\mu = 175 \pm 1{,}96 * 6/\sqrt{16} = 175 \pm 2{,}9.$$

L'intervalle de confiance pour μ est de 172,1 cm à 177,9 cm. On ne peut critiquer le choix de la distribution normale pour résoudre ce problème. On peut s'interroger sur deux points. Premièrement, il peut s'avérer difficile de choisir un échantillon réellement aléatoire pour une si vaste population. Deuxièmement, un petit échantillon donne un intervalle de confiance assez large, ce qui rend la comparaison avec les années antérieures assez difficile. Si un échantillon de taille $n = 100\,000$ avait été choisi, l'intervalle de confiance serait de 174,96 cm à 175,04 cm.

3e cas: *n* est grand et σ inconnu: Dans ce cas $\sigma_{\bar{X}}$ ne peut être obtenu par $\sigma/\sqrt{n}$, car σ est inconnu. On a vu à la section 9.8 que s^2 est un estimateur ponctuel de σ^2 et que les résultats $\dfrac{\bar{X}-\mu}{s/\sqrt{n}}$ sont normalement distribués si $n \geqslant 30$.

4e cas: *n* est petit et σ inconnu: Ici, s est connu ou peut être calculé. On prend tous les échantillons de taille n possibles, on calcule la moyenne de chaque échantillon, puis on calcule la cote «t» de chaque moyenne en utilisant la formule

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Si la population est normalement distribuée et seulement dans ce cas, la distribution de toutes les cotes t de tous les échantillons de taille n est appelée une distribution t; ce n'est pas une distribution normale. (William Gosset, un Irlandais, a publié cette distribution sous le pseudonyme de Student d'où le nom Student t ou plus simplement distribution t.) On peut déterminer la cote t de tout échantillon; l'ensemble de toutes les cotes t de tous les échantillons possède une distribution, mais celle-ci n'est pas nécessairement une distribution de Student.

Exemple 2

Deux échantillons de taille 100 sont pris d'une population où $\mu = 100$ et $\sigma = 10$. Les échantillons ont la même moyenne $\bar{X}_1 = \bar{X}_2 = 110$, mais ont des écarts-types différents: $s_1 = 9$ et $s_2 = 8$. Calculer les cotes z et t de chaque échantillon.

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}: \qquad z_1 = \frac{110-100}{10/\sqrt{100}} = 10, \qquad z_2 = \frac{110-100}{10/\sqrt{100}} = 10$$

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}: \qquad t_1 = \frac{110-100}{9/\sqrt{100}} = 11{,}1 \qquad t_2 = \frac{110-100}{8/\sqrt{100}} = 12{,}5$$

Il est à noter que $z_1 = z_2$ et que $t_1 \neq t_2$.

Cet exemple illustre le fait que si deux échantillons ont la même moyenne et proviennent de la même population, ils ont la même cote z tout en ayant des écarts-types différents, ce qui n'est pas le cas des cotes t. Les cotes t ne sont pas normalement distribuées; nous verrons dans ce qui suit le genre de courbe obtenue.

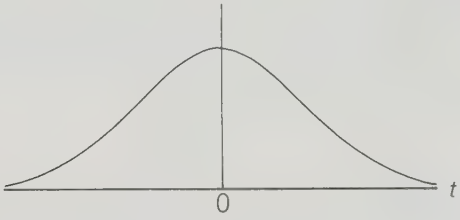
12.2 CARACTÉRISTIQUES DE LA DISTRIBUTION *t*

Les distributions t de Student ont les caractéristiques suivantes si la population est normalement distribuée:

1. Il y a autant de distributions t qu'il y a de valeurs possibles pour n. Il y a une seule courbe normale et une seule table permet de trouver la valeur de z qui nous intéresse. Il y a une famille de courbes t; il faudrait alors un grand nombre de tables chacune

correspondant à une valeur de n. On a réussi à condenser et à ramener le nombre à une seule table. Nous verrons à la section 12.5 comment l'utiliser.

2. Chaque courbe t est symétrique par rapport à 0, ce qui veut dire que la partie droite est identique à la partie gauche et que la moyenne de chaque distribution t est 0.



3. Le sommet de la courbe est obtenu pour $t = 0$.

4. Si n augmente, les courbes t deviennent de plus en plus près de la courbe normale. On peut le constater en examinant les graphiques superposés ci-dessous:



Si n est supérieur à 30, la courbe t est à toute fin pratique confondue avec la courbe normale et il suffit d'utiliser cette dernière. C'est cette superposition qui justifie l'utilisation de $s/\sqrt{n}$ comme estimation de $\sigma_{\bar{x}}$ quand n est grand et σ inconnu.

5. Chaque distribution t est une distribution de probabilité. Donc l'aire totale sous la courbe est 1 et la probabilité que t soit compris entre a et b est donnée par l'aire sous la courbe dans cet intervalle.

12.3 QUAND DOIT-ON UTILISER UNE DISTRIBUTION t?

Une distribution t est utilisée quand:

> (a) la population est normalement distribuée **et**
>
> (b) σ est inconnu (mais s est connu ou peut être calculé) **et**
>
> (c) $n < 30$.

Si la population n'est pas normale, on commet une erreur en utilisant la loi de Student. L'importance de cette erreur dépend de l'éloignement de la distribution de la population de la loi normale. Si la population est à peu près normale, l'erreur est petite. En pratique, on ne sait pas si la population suit une loi normale ou pas, ce qui augmente la difficulté.

Que ferons-nous alors? Nous pourrons utiliser un test qui n'exige pas que la population soit normale pour être valide (l'un de ces tests sera étudié au chapitre 17). On peut

également augmenter la taille de l'échantillon; plus on doute de la normalité de la population, plus on augmentera la taille de l'échantillon. Il faudra donc être certain que la population est normale si on veut utiliser de petits échantillons. C'est la difficulté des tests utilisant des petits échantillons, ils sont plus douteux qu'utiles.

Si σ est connu, on utilise une distribution normale, de même si $n \geqslant 30$ et que la population est normale.

12.4 DEGRÉ DE LIBERTÉ

Pour utiliser la distribution *t*, il faut connaître le degré de liberté. Imaginons la situation suivante: trois cartes sont marquées 0; 10; 20; si on pige 2 cartes, on peut prédire le résultat de la 3e; ainsi si on pige 0 et 20, on sait que la 3e sera 10. Dans ce cas on dira qu'il y a **deux degrés de liberté** (on peut piger au hasard 2 cartes et prédire à tous les coups la 3e).

Si 4 résultats ont une moyenne de 50, combien de ces résultats peuvent être choisis au hasard? Supposons que l'on choisisse 40; 30 et 70; peut-on choisir n'importe quel nombre comme 4e résultat et espérer obtenir une moyenne de 50? Évidemment non, le seul qui donnera une moyenne de 50 est 60, on obtient 40; 30; 70; **60** et $\overline{X} = 50$. On peut donc choisir au hasard 3 des 4 nombres. On a donc 3 **degrés de liberté.** Si *n* résultats ont une moyenne $\overline{X}$, $n - 1$ peuvent être choisis au hasard; on aura donc ***n* – 1 degrés de liberté.**

Pour trouver un intervalle de confiance ou vérifier une hypothèse concernant la moyenne, si une distribution *t* d'un échantillon de taille *n* est utilisée, le nombre de degrés de liberté sera $n - 1$. Soit D le nombre de degrés de liberté, on aura

$$D = n - 1$$

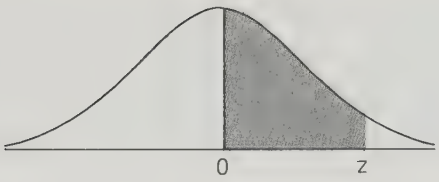
12.5 TABLE DE LA DISTRIBUTION *t*

Pour une distribution normale, une table (table 4, appendice C) suffit pour donner l'aire sous la courbe normale. Pour une distribution *t*, il nous faudrait autant de tables qu'il y a de degrés de liberté. On a réussi à concentrer en une seule table les principales valeurs d'une distribution *t*.

La table 5, appendice C, donne les valeurs de *t* pour une aire choisie sous la courbe *t*. Les différentes valeurs de D apparaissent dans la première colonne. La table est construite pour assurer une utilisation efficace des tests unilatéraux ou bilatéraux. Il est à noter que, dans le cadre du haut, on donne l'aire pour un test unilatéral ou bilatéral, alors que pour une courbe normale on donne l'aire de 0 à *z*.



Valeurs de *t* (Voir table 5, appendice C)



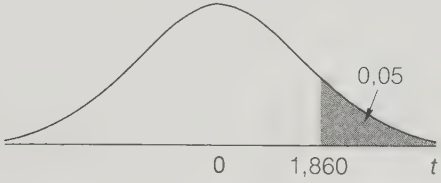
Aire sous la courbe normale
(Voir table 4, appendice C)

Pour trouver un intervalle de confiance avec une distribution *t*, on devra prendre un test bilatéral. Il est nécessaire de soustraire de 1 le niveau de confiance pour trouver l'aire des deux extrémités. L'intervalle de confiance est large pour des petites valeurs de *n* et, pour cette raison, on utilise souvent une distribution *t* pour vérifier une hypothèse.

Exemple 1

Si $D = 8$, trouver la valeur de *t* pour laquelle 5 pour cent de l'aire sous la courbe est à droite de celle-ci?

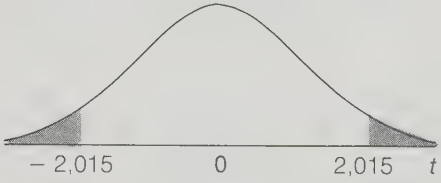
Dans la table 5 apprendice C, lire la ligne nommée «Unilatéral» jusqu'à la valeur de 0,05 puis descendre dans cette colonne jusqu'à la valeur de 8 dans la colonne *D*. On obtient $t = 1{,}860$.



Exemple 2

Si $D = 5$, quelle est la probabilité que la valeur de *t* soit supérieure à 2,015 ou inférieure à $-2{,}015$?

En suivant la ligne $D = 5$, on trouve la valeur la plus rapprochée qui est 2,02; dans cette colonne en haut en prenant la ligne bilatérale on obtient 0,10.



12.6 INTERVALLE DE CONFIANCE ET TEST D'HYPOTHÈSE CONCERNANT LA MOYENNE

Pour trouver un intervalle de confiance ou faire un test d'hypothèse sur la moyenne avec une distribution *t*, on procède comme précédemment, sauf qu'il faut tenir compte du degré de liberté et se servir de la table 5 au lieu de la table 4.

Exemple 1

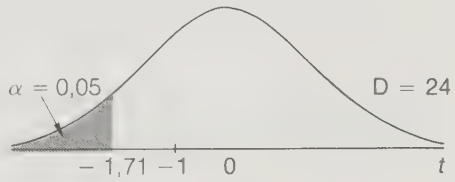
Un fabricant d'ampoules électriques affirme qu'elles ont une vie moyenne de 1 000 heures. Son affirmation est-elle justifiée au niveau de signification de 0,05, si un échantillon aléatoire de 25 ampoules a une vie moyenne de 994 heures avec un écart-type de 30 heures?

Supposer que la distribution de la vie des ampoules est normale.

$H_0 : \mu = 1\,000$; $H_1 : \mu < 1\,000$; test unilatéral; $\alpha \doteq 0{,}05$.

$D = 24$; H_0 est rejetée si $t < -1{,}71$

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{994 - 1\,000}{30/\sqrt{25}} = -1{,}00.$$



Son affirmation est justifiée: H_0 est acceptée.

Exemple 2

Trouver un intervalle de confiance pour μ au niveau de 98% si un échantillon aléatoire de taille 9 pris dans une population normale a une vie moyenne de 14 et un écart-type de 2.

$$\overline{X} = 14;\ s = 2;\ n = 9;\ D = 8$$

Au niveau de 98%, nous aurons à l'extérieur des limites de confiance une aire de $1 - 0{,}98 = 0{,}02$, $D = 8$ et $t = \pm 2{,}90$.

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

$$\pm 2{,}90 = \frac{14 - \mu}{2/\sqrt{9}}$$

$$\mu = 14 \pm 1{,}93$$

L'intervalle de confiance pour μ est 12,1 à 15,9.

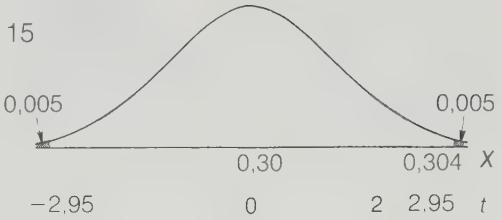
Exemple 3

Les spécifications d'un certain médicament indiquent que chaque comprimé doit contenir 30% de vitamine A. Seize comprimés sont choisis au hasard puis analysés; ils contiennent en moyenne 30,4% de vitamine A avec un écart-type de 0,8%. Est-ce que le médicament respecte les spécifications au niveau de signification de 0,01? Supposer que les erreurs sont normalement distribuées.

$H_0 : \mu = 0{,}30$; $H_1 : \mu \neq 0{,}30$; test bilatéral; $\alpha = 0{,}01$

$\bar{X} = 0{,}304$; $s = 0{,}008$; $n = 16$; $D = 15$

$$t = \frac{0{,}304 - 0{,}30}{0{,}008/\sqrt{16}} = 2{,}00$$



$D = 15$, $\alpha = 0{,}01$, H_0 est rejetée si $t > 2{,}95$ ou $t < -2{,}95$. H_0 est acceptée, les spécifications sont respectées au niveau de 0,01.

12.7 DIFFÉRENCES DE MOYENNES POUR DES ÉCHANTILLONS INDÉPENDANTS

Une distribution t peut être utilisée pour vérifier des hypothèses sur les différences de moyennes d'échantillons indépendants, **seulement si les deux populations sont normales.** Si ce n'est pas le cas, on doit augmenter la taille des échantillons (au moins 30) et la distribution d'échantillonnage de $\bar{X} - \bar{Y}$ est approximativement normale.

Deux cas peuvent se présenter: 1° les deux populations ont la même variance, 2° les variances ne sont pas égales. Comment savoir si les deux populations ont la même variance? Si on connaît la variance de chacune, on utilisera la variable z comme à la section 10.4. Au chapitre 14, nous verrons comment vérifier l'hypothèse que deux populations ont la même variance. Pour l'instant, supposons que l'on peut distinguer les deux cas.

1er cas: Les populations ont la même variance.

Les intervalles de confiance sont déterminés et les hypothèses sont vérifiées en utilisant un test t tout comme on le ferait avec un test z excepté que:

(a) $$\sigma_{\bar{x}\cdot\bar{y}} = \sqrt{\frac{(n_X - 1)\,s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}$$

Cette formule est démontrée à la section 12.8.

(b) $D = (n_X - 1) + (n_Y - 1) = n_X + n_Y - 2.$

(c) La table 5 est utilisée.

Exemple 1

Un échantillon aléatoire de 17 enfants de 3e année qui ont des difficultés de lecture a un Q.I. moyen de 98 avec un écart-type de 10; un échantillon aléatoire de 10 élèves sans difficulté de lecture a un Q.I. de 101 avec un écart-type de 9. Au niveau de signification de 0,05, y a-t-il une différence entre

les Q.I. moyens des mauvais lecteurs et des bons lecteurs? Supposer que les Q.I. des deux groupes sont normalement distribués et qu'ils ont la même variance.

$H_0: \mu_X - \mu_Y = 0$; $H_1: \mu_X - \mu_Y \neq 0$; test bilatéral, $\alpha = 0{,}05$

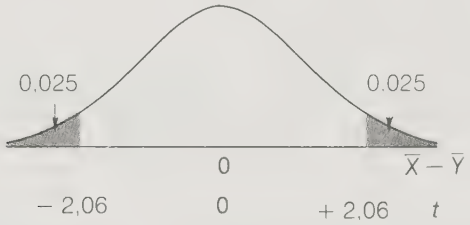
$$\overline{X} = 98,\ s_X = 10,\ n_X = 17;\ \overline{Y} = 101,\ s_Y = 9,\ n_Y = 10$$

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{16 * 10^2 + 9 * 9^2}{17 + 10 - 2}\left(\frac{1}{17} + \frac{1}{10}\right)} = \sqrt{\frac{2329}{25} * \frac{27}{170}} = 3{,}85$$

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma_{\overline{X}\ \overline{Y}}} = \frac{(98 - 101) - 0}{3{,}85} = -0{,}78$$

$D = 25$; $\alpha = 0{,}05$; test bilatéral, $t = \pm 2{,}060$.

Le résultat $-0{,}78$ est dans la région d'acceptation.



Il n'y a pas de différence entre les deux groupes.

Exemple 2

Deux procédés différents sont utilisés pour fabriquer du fil de coton. On doit comparer les deux procédés. Un échantillon aléatoire de 20 fils produits par le 1er procédé a un point de rupture de 12 g avec un écart-type de 1,5 g alors qu'un échantillon de 5 fils produits par le 2e procédé a un point de rupture de 10 g avec un écart-type de 2,0 g. Supposer que les deux populations sont normales et ont la même variance.

(a) Vérifier l'hypothèse qu'il n'y a pas de différence dans les points de rupture des deux procédés au niveau de signification de 0,05.

(b) Trouver avec une certitude de 95% la différence dans les points de rupture des fils produits par les deux procédés.

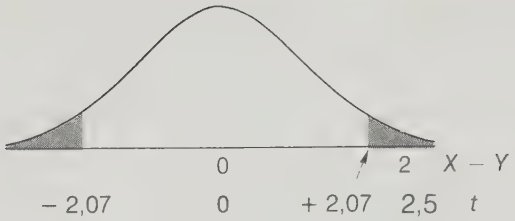
(a) $H_0: \mu_X - \mu_Y = 0$; $H_1: \mu_X - \mu_Y \neq 0$

$\overline{X} = 12$; $s_X = 1{,}5$; $n_X = 20$; $\overline{Y} = 10$; $s_Y = 2{,}0$; $n_Y = 5$

$D = 20 + 5 - 2 = 23$; $\alpha = 0{,}05$; donc $t = \pm 2{,}07$

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}\left(\dfrac{1}{n_X} + \dfrac{1}{n_Y}\right)}}$$

$$= \frac{(12 - 10) - 0}{\sqrt{\dfrac{19 * 1{,}5^2 + 4 * 2{,}0^2}{23}\left(\dfrac{1}{20} + \dfrac{1}{5}\right)}} = \frac{2{,}0}{0{,}80} = 2{,}5$$



H_0 est rejetée; il y a une différence.

(b) $$\pm 2{,}07 = \frac{(12-10)-(\mu_X-\mu_Y)}{0{,}80}$$

$$\mu_X - \mu_Y = 2 \pm 1{,}66 = 0{,}34 \text{ ou } 3{,}66$$

L'intervalle de confiance au niveau de 95% pour la différence des points de rupture est 0,3 à 3,7 g.

2e *cas:* Les populations n'ont pas la même variance.

Dans ce cas, $s_{\bar{X}}$ et $s_{\bar{Y}}$ sont utilisés pour estimer $\sigma_{\bar{X}}$ et $\sigma_{\bar{Y}}$ respectivement et

$$\sigma_{(\bar{X}-\bar{Y})} = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

Il est plus facile de trouver $\sigma_{(\bar{X}-\bar{Y})}$ que dans le 1er cas, mais la difficulté est de trouver le degré de liberté. Supposons que les échantillons ont des variances très voisines, mais que n_X est beaucoup plus petit que n_Y. Alors s_X^2/n_X est beaucoup plus grand que s_Y^2/n_Y et joue un rôle important dans le calcul de σ_{X-Y}. Ce serait une erreur que d'utiliser $n_X + n_Y - 2$ dans le calcul de D. La valeur du degré de liberté est donnée par la formule suivante:

$$D = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\left(\frac{s_X^2}{n_X}\right)^2 \frac{1}{n_X+1} + \left(\frac{s_Y^2}{n_Y}\right)^2 \frac{1}{n_Y+1}} - 2$$

où D est arrondi à l'entier le plus près.

Exemple 3

Reprendre l'exemple 2 où cette fois les populations n'ont pas la même variance.

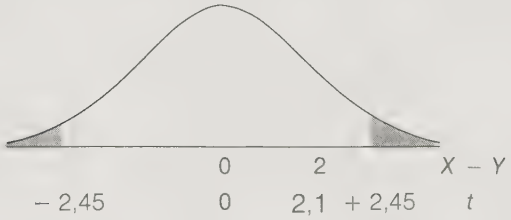
$H_0: \mu_X - \mu_Y = 0$; $H_1: \mu_X - \mu_Y \neq 0$.

$\bar{X} = 12$; $s_X = 1{,}5$; $n_X = 20$; $\bar{Y} = 10$; $s_Y = 2{,}0$; $n_Y = 5$

$$t = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}} = \frac{(12-10)-0}{\sqrt{\frac{1{,}5^2}{20}+\frac{2{,}0^2}{5}}}$$

$$= \frac{2,0}{0,96} = 2,1$$

$$D = \frac{\left(\frac{1,5^2}{20} + \frac{2,0^2}{5}\right)^2}{\left(\frac{1,5^2}{20}\right)^2 \frac{1}{21} + \left(\frac{2,0^2}{5}\right)^2 \frac{1}{6}} - 2 = \frac{0,83}{0,11} - 2 = 8 - 2 = 6$$



$\alpha = 0,05$, $D = 6$, test bilatéral, donc $t = \pm 2,45$
H_0 est acceptée (dans l'exemple 2, H_0 est rejetée).

★12.8 DÉMONSTRATION DE $\sigma_{(\bar{X} - \bar{Y})}$ QUAND LES POPULATIONS ONT DES VARIANCES ÉGALES

Si chaque résultat d'un ensemble a la même importance, la moyenne est non-pondérée. Si chaque résultat possède une importance propre, on associe à celle-ci un poids W en accord avec son importance relative. On obtient alors une moyenne pondérée. Pour trouver cette moyenne, chaque résultat est multiplié par son poids W, les produits sont additionnés puis le total divisé par la somme des W. On obtient

$$\bar{X}_W = \frac{\Sigma(WX)}{\Sigma W}$$

Exemple 1

André a conservé une moyenne au bâton de 0,300 dans les six premières semaines de la saison et de 0,400 dans les 4 semaines suivantes. Quelle est sa moyenne au bâton pour la saison? Il a été au bâton 70 fois dans les six premières semaines et 30 fois dans les 4 suivantes.

On peut répondre à la question de deux manières.

1. Il a frappé $0,300 * 70 = 21$ coups sûrs dans les 6 premières semaines et $0,400 * 30 = 12$ coups sûrs dans les 4 semaines suivantes. Au total, il a frappé $21 + 12 = 33$ coups sûrs en $70 + 30 = 100$ présences au bâton, ce qui lui donne une moyenne pour la saison de 0,330.

2. En utilisant la moyenne pondérée, on obtient:

$$\frac{70 * 0,300 + 30 * 0,400}{70 + 30} = 0,330.$$

On a déjà utilisé la moyenne pondérée quand on a calculé la moyenne de données regroupées où le poids était la fréquence de chaque classe.

classe	f
1–3	9
4–6	18

$$\mu = \frac{2(9) + 5(18)}{9 + 18} = 4$$

La variance σ^2 de la population est estimée par la moyenne pondérée de s^2_X et s^2_Y en prenant pour poids leur degré de liberté. Ainsi,

$$\sigma^2 \approx \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{(n_X - 1) + (n_Y - 1)}.$$

Si les deux populations ont la même variance ($\sigma_X^2 = \sigma_Y^2 = \sigma^2$)

$$\sigma_{(\bar{X}-\bar{Y})}^2 = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y} = \sigma^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right) = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right).$$

En prenant la racine carrée, on obtient la formule (a) de la section 12.7.

12.9 DIFFÉRENCE DE MOYENNES DE DEUX ÉCHANTILLONS DÉPENDANTS

Quelquefois, nous n'avons pas d'échantillons indépendants, par exemple, on veut comparer les résultats obtenus par un échantillon avant et après avoir subi un certain traitement ou encore on ne veut pas que certains facteurs (âge, religion, salaire...) viennent déranger les résultats. Dans les deux cas, on veut avoir le même échantillon pour assurer la stabilité des facteurs externes. Un test pour la différence de moyennes ne peut être utilisé puisque les échantillons ne sont pas indépendants. Les résultats de chaque individu avant et après sont compilés et un test t peut être fait pour la différence de résultats de chaque paire. Soit d cette différence, $\bar{d}$ la moyenne, s_d l'écart-type des différences de résultats et μ_d la moyenne des différences de la population. S'il y a n paires de résultats alors $D = n - 1$.

Exemple 1

Un échantillon aléatoire de 10 étudiants passent un examen d'arithmétique et obtiennent les résultats (X) ci-dessous. Ils passent un peu plus tard dans la session, après une revue, le même examen et obtiennent les résultats (Y). Au niveau de signification de 0,05, est-ce que les résultats du second test sont meilleurs?

étudiant	1er Examen (X)	2e Examen (Y)	$Y - X = d$	$d - \bar{d}$	$(d - \bar{d})^2$
1	80	84	+4	2,8	7,84
2	50	56	6	4,8	23,04
3	78	81	3	1,8	3,24
4	90	92	2	0,8	0,64
5	75	76	1	−0,2	0,04
6	70	65	−5	−6,2	38,44
7	62	72	10	8,8	77,44
8	87	84	−3	−4,2	17,64
9	95	93	−2	−3,2	10,24
10	68	64	−4	−5,2	27,04
			12		205,60

$H_0 : \mu_d = 0$

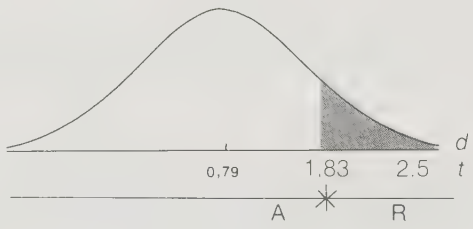
$H_1 : \mu_d > 0$

moyenne des différences $= \bar{d} = 12/10 = 1{,}2$.

écart-types des différences $= s_d = \sqrt{\frac{205{,}6/9}{9}} = 4{,}78$

Un test t est appliqué à 10 différences.
$\mu_d = 0$ implique que la différence des résultats est nulle pour toute la population.
H_1 indique qu'on doit faire un test unilatéral (à droite).
$n = 10$, il y a donc 9 degrés de liberté, $\alpha = 0{,}05$.
$D = 9$, la valeur critique de t est 1,83.

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{1{,}2 - 0}{4{,}78/\sqrt{10}} = 0{,}79$$



H_0 est acceptée; il n'y a pas de différence entre les résultats des deux examens.

Exemple 2

De 240 étudiants inscrits au cours de statistiques, 8 paires d'étudiants sont choisies en tenant compte de leurs antécédents mathématiques (les élèves d'une même paire sont de même force, mais les paires ne sont pas nécessairement de même force). Le professeur utilise deux méthodologies pour ce cours: (X), 30 élèves rencontrent le professeur 3 fois par semaine; (Y), les 210 autres élèves ont 3 heures de cours plus une récitation hebdomadaire. Pour chacune des 8 paires d'étudiants, on choisit au hasard qui fera partie du

groupe (X) et qui sera du groupe (Y). Les résultats des 8 paires d'étudiants à l'examen final qui est le même pour les deux groupes sont donnés dans le tableau ci-dessous. Au niveau de signification de 0,01, y a-t-il une différence entre les moyennes des deux groupes?

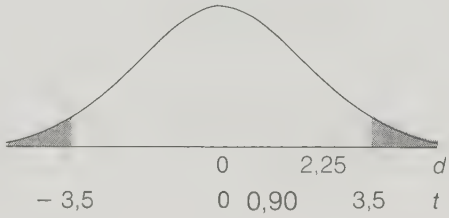
Paire	X	Y	$X - Y = d$	$d - 2{,}25$	$(d - 2{,}25)^2$
1	90	82	+ 8	5,75	33,0625
2	85	95	−10	−12,25	150,0625
3	75	79	− 4	−6,25	39,0625
4	78	81	− 3	−5,25	27,5625
5	95	88	+ 7	4,75	22,5625
6	95	91	+ 4	1,75	3,0625
7	60	50	+10	7,75	60,0625
8	87	81	+ 6	3,75	14,0625
			18		349,50

$H_0: \mu_d = 0$; $H_1: \mu_d \neq 0$

$$\bar{d} = \frac{18}{8} = 2{,}25, \; s_d = \sqrt{349{,}5/7} = 7{,}07$$

$n = 8$, $D = 7$, $\alpha = .01$; test bilatéral; les valeurs critiques de t sont $\pm 3{,}5$.

$$t = \frac{2{,}25 - 0}{7{,}07/\sqrt{8}} = 0{,}90$$



H_0 est acceptée; il n'y a pas de différence significative entre les deux groupes.

12.10 VOCABULAIRE ET SYMBOLES

Distribution t
Degré de liberté D
Moyenne pondérée
s_d
Différence de paires
Distribution t de Student
$\sigma_{(\bar{X} - \bar{Y})}$
μ_d
d

12.11 EXERCICES

1. Quelle est la probabilité que la moyenne d'un échantillon aléatoire de taille 17 pris dans une population normale ait une valeur t inférieure à 2,583?

2. Dans une distribution t avec 4 degrés de liberté, 10% des résultats t sont supérieurs à quelle valeur?

3. Neuf travailleurs choisis au hasard dans une usine de 250 travailleurs ont un salaire moyen de \$125 par semaine avec un écart-type de \$12. Estimer le salaire moyen de tous les travailleurs au niveau de confiance de 95%. (Supposer que les salaires sont distribués normalement.)

4. Un métallurgiste fait 8 mesures du point de fusion d'un alliage. Ses lectures furent 450°; 448°; 452°; 448°; 451°; 450°; 451°; 450°. Quel est l'intervalle de confiance du point de fusion de l'alliage au niveau de 90%? (Supposer que les erreurs de lectures sont normalement distribuées.)

5. Cinq malades souffrant d'hypertension sont traités par un nouveau médicament qui a diminué leur pression de 4; 25; 13; 18 et 20 unités respectivement. Au niveau de signification de 0,05, est-ce que le nouveau médicament diminue la pression d'au moins 20 unités? Poser que les variations de pression des malades sont normalement distribuées.

6. Une école de lecture dynamique affirme que les adultes peuvent lire 400 mots à la minute plus vite après avoir complété le cours. Un échantillon aléatoire de 12 adultes montre qu'ils ont augmenté leur rapidité de lecture de 350 mots à la minute avec un écart-type de 50 mots à la minute. Est-ce que l'affirmation de l'école est justifiée au niveau de signification de 0,05?

●7. Un nouveau fusible fabriqué par une compagnie doit avoir selon les spécifications un temps de fusion de 5 secondes avec un écart-type de 1 s. Un échantillon aléatoire de 16 fusibles a un temps moyen de fusion de 5,8 s, avec un écart-type de 0,95 s. Est-ce que la production des fusibles d'où provient l'échantillon est conforme aux spécifications au niveau de signification de 0,01? Supposer que les temps de fusion sont normalement distribués.

●8. Les nombres suivants sont pris en ordre dans une table de nombres aléatoires: 3; 9; 5; 2; 7; 1; 7; 6; 2; 4. Vérifier si les nombres proviennent réellement d'une table de nombres aléatoires; $\alpha = 0{,}05$ (la moyenne des nombres d'une table de nombres aléatoires est 4,5).

9. Un inspecteur du département des poids et mesures vérifie le poids de 10 contenants de sucre de 1 kg. Il trouve que leur poids moyen est de 996 g avec un écart-type de 6 g. Le poids des contenus de sucre se distribue normalement.

(a) Y a-t-il une différence entre le poids réel et le poids indiqué ($\alpha = 0{,}05$)?
(b) Sont-ils plus légers qu'ils doivent l'être de façon significative (au niveau de signification de 0,05)?

10. Un propriétaire veut déterminer le pH de son sol; il doit ajouter de la chaux si le pH est inférieur à 6,5. Il prend 16 échantillons de sol choisis au hasard et trouve que le pH moyen est de 6,3. Doit-il ajouter de la chaux à son sol ($\alpha = 0{,}10$)?

●11. Dix typographes d'une imprimerie commettent en moyenne 2,3 erreurs par page. On choisit au hasard 10 pages d'un apprenti et on trouve qu'il fait en moyenne 2,6 erreurs avec un écart-type de 0,5. Est-ce que l'apprenti commet de façon significative plus d'erreurs que les typographes d'expérience?

12. L'absorption d'eau de 8 linges damassés est mesurée quand ils sont neufs et remesurée après qu'ils furent lavés 3 fois. Voici les résultats obtenus.

Linges damassés	1	2	3	4	5	6	7	8
Absorption (neufs) (cm^3)	34,0	33,0	35,0	32,0	35,0	34,0	36,0	36,0
Absorption (après lavages) (cm^3)	37,0	35,0	35,0	34,0	39,0	36,0	40,0	40,0

Après lavages est-ce que les linges ont une meilleure absorption ($\alpha = 0{,}10$)?

13. À la session d'automne, 5 étudiants choisis au hasard dans le cours de statistiques ont obtenu une moyenne de 75 avec un écart-type de 5. À la session d'hiver, un échantillon aléatoire de 10 étudiants au cours de statistiques a une moyenne de 77 avec un écart-type de 5. Au niveau de signification de 0,05, est-ce que les étudiants de la session d'hiver ont obtenu de meilleurs résultats? Les résultats sont normalement distribués aux deux sessions et ont la même variance.

14. Huit blouses de laboratoire furent lavées avec un détergent bleu et six avec un détergent blanc. Après le lavage, un test de blancheur (échelle 1 à 10; 10 étant le plus blanc) est appliqué à chacune des blouses. La moyenne pour les 8 blouses lavées au détergent bleu est de 7,5 avec un écart-type de 0,3 alors que les 6 blouses lavées au détergent blanc ont une moyenne de 6,8 avec un écart-type de 0,4. Supposer que les valeurs de toutes les blouses sont normalement distribuées pour les deux sortes de détergent, mais que les variances ne sont pas égales. Y a-t-il une différence entre les deux détergents au niveau de signification de 0,05?

15. Un groupe de professeurs du collégial veut enquêter sur le travail d'été de leurs confrères. Parmi ceux qui donnent des cours d'été, 20 sont choisis au hasard. On trouve qu'ils gagnent en moyenne \$600 par mois, avec un écart-type de \$100. Parmi ceux qui travaillent dans l'immobilier, 10 sont choisis au hasard et gagnent en moyenne \$700 par mois avec un écart-type de \$50. Au niveau de signification de 0,02, y a-t-il une différence dans les salaires des deux groupes? Les salaires des deux groupes sont normalement distribués, mais n'ont pas nécessairement le même écart-type.

16. Deux types de fusils sont vérifiés à l'arsenal de Valcartier. Douze du 1er type et dix du second sont choisis au hasard puis vérifiés. Ceux du 1er type ont une erreur moyenne de 3,0 cm avec un écart-type de 0,40 cm. Ceux du second type ont une erreur moyenne de 2,1 cm avec un écart-type de 0,30 cm. Supposer que les erreurs des deux types de fusils sont normalement distribuées avec la même variance. Au niveau de signification de 0,05 est-ce que l'erreur moyenne du second type est au moins de 1 cm inférieur à celle du 1er type?

17. Douze étudiants choisis au hasard ont été pesés à leur entrée au collège en septembre et furent pesés à nouveau le 1er novembre. Voici les résultats obtenus.

Étudiant	1	2	3	4	5	6	7	8	9	10	11	12
Sept.	68	80	72	68	90	74	90	94	88	82	90	76
1er nov.	73	80	65	68	95	69	95	84	91	87	95	80

Est-ce que les étudiants en général ont gagné du poids pendant ces 2 mois au

niveau de signification de 0,05? Supposer que les poids sont normalement distribués et ont le même écart-type.

18. Un économiste veut étudier la différence de salaire entre les travailleurs de deux usines appartenant à la même compagnie. Il choisit un échantillon aléatoire de 16 travailleurs dans chaque usine et forme 16 paires de travailleurs en tenant compte de la hiérarchie d'embauche dans chaque usine. Il trouve les salaires horaires suivants:

Paire:	1	2	3	4	5	6	7	8
X:	$2,50	3,40	3,20	2,60	3,00	2,75	4,00	3,75
Y:	$2,80	3,70	3,00	3,00	3,10	2,85	4,10	3,85
Paire:	9	10	11	12	13	14	15	16
X:	$5,20	4,50	3,00	4,05	4,20	4,15	4,30	4,60
Y:	$5,30	4,60	3,20	4,15	4,30	4,35	4,00	4,70

En supposant que les différences de salaire sont normalement distribuées, vérifier si les salaires sont différents dans les deux usines au niveau de signification de 0,01.

19. Deux machines différentes sont utilisées pour fabriquer des enjoliveurs de roues d'automobile. Le nombre produit par heure pour chaque machine est noté pour des heures choisies au hasard sur une période de plusieurs mois. La première machine sur un échantillon de 10 heures a une production de 100 enjoliveurs à l'heure avec un écart-type de 6,0. La seconde machine, également pour 10 heures, a une production moyenne de 94 avec un écart-type de 5,0. Est-ce que la première machine produit plus de 5 enjoliveurs de plus à l'heure que la seconde au niveau de signification de 0,05? Poser que la production des deux machines est normalement distribuée et que l'écart-type est le même.

20. Un psychologue a découvert que 10 orphelins qui regardent des photos (de même grandeur) d'amis ou d'enfants inconnus, ont tendance à juger leurs amis 5 cm plus grands que la réalité en moyenne avec un écart-type de 0,05 cm. Huit enfants habitant avec leurs parents passent le même test et trouvent que leurs amis sont 3 cm plus grands avec un écart-type de 0,04 cm. Trouver un intervalle de confiance au niveau de 95% pour la différence des grandeurs estimées par les orphelins et les enfants vivant avec leurs parents.

RÉPONSES

1. $D = 16$; $p\ (t > 2{,}583) = 0{,}01$; $p\ (t < 2{,}583) = 0{,}99$.

3. $\overline{X} = 125, s = 12, n = 9, D = 8$. Au niveau de confiance de 95% avec $D = 8$, $t = \pm 2{,}31$

$$\pm 2{,}31 = \frac{125 - \mu}{12/\sqrt{9}} \qquad \mu = 125 \pm 9{,}2 = 115{,}8 \text{ ou } 134{,}2$$

l'intervalle de confiance au niveau de 95% est \$115,80 à \$134,20.

5. $\bar{X} = 16$; $s = 8{,}0$; $n = 5$; $D = 4$; $\alpha = 0{,}05$; $H_0 : \mu = 20$; $H_1 : \mu < 20$; test unilatéral.
H_0 est rejetée si $t < -2{,}13$; $t = \dfrac{16 - 20}{8/\sqrt{5}} = -1{,}1$; H_0 est acceptée.

7. ($\sigma = 1$) est connu, donc on utilise la cote z et non t.
$H_0 : \mu = 5$; $H_1 : \mu \neq 5$; test bilatéral; $\alpha = 0{,}01$; H_0 est rejetée si $z > 2{,}58$ ou $z < -2{,}58$

$$\bar{X} = 5{,}8; n = 16; \mu = \frac{5{,}8 - 5}{1/\sqrt{16}} = 3{,}2$$

H_0 est rejetée; les spécifications ne sont pas respectées.

9. (a) $H_0 : \mu = 1\,000$ g; $H_1 : \mu \neq 1\,000$ g; $\bar{X} = 996$; $s = 6$; $n = 10$;

$t = \dfrac{996 - 1\,000}{6/\sqrt{10}} = -2{,}11$; $\alpha = 0{,}05$; $D = 9$; test bilatéral. Les valeurs critiques de t sont $\pm 2{,}26$. H_0 est acceptée.

(b) $H_0 : \mu = 1\,000$; $H_1 : \mu < 1\,000$; $t = -2{,}11$ (comme en a) test unilatéral. $D = 9$; $\alpha = 0{,}05$. La valeur critique de t est $-1{,}83$ donc H_0 est rejetée.

11. Les erreurs de l'apprenti ne sont probablement pas distribuées normalement, car l'échantillon est trop petit.

13. $\bar{X} = 75$, $s_X = 5$, $n_X = 5$; $\bar{Y} = 77$, $s_Y = 5$, $n_Y = 10$.
$H_0 : \mu_X - \mu_Y = 0$; $H_1 : \mu_X - \mu_Y < 0$. test unilatéral $\alpha = 0{,}05$; $D = 13$; H_0 est rejetée si $t < -1{,}77$

$$t = \frac{(75 - 77) - 0}{\sqrt{\dfrac{4 * 5^2 + 9 * 5^2}{13}\left(\dfrac{1}{5} + \dfrac{1}{10}\right)}} = -0{,}73; H_0 \text{ est acceptée.}$$

15. $H_0 : \mu_X - \mu_Y = 0$, $H_1 : \mu_X - \mu_Y \neq 0$. $\bar{X} = 600$, $n_X = 20$, $s_X = 100$; $\bar{Y} = 700$, $n_Y = 10$, $s_Y = 50$.

$$t = \frac{(600 - 700) - 0}{\sqrt{\dfrac{100^2}{20} + \dfrac{50^2}{10}}} = -3{,}65$$

$$D = \frac{\left(\dfrac{100^2}{20} + \dfrac{50^2}{10}\right)^2}{\left(\dfrac{100^2}{20}\right)^2 \dfrac{1}{21} + \left(\dfrac{50^2}{10}\right)^2 \dfrac{1}{11}} - 2 = 30, \alpha = 0{,}02$$

Donc les valeurs critiques de t sont $\pm 2{,}46$; H_0 est rejetée; il y a une différence significative entre les salaires des deux groupes.

17. $H_0 : \mu_d = 0$; $H_1 : \mu_d > 0$. d: 5; 0; -7; 0; 5; -5; 5; -10; 3; 5; 5; 4.

$$\bar{d} = 10/12 = 0{,}833\ ,\ s_d = \sqrt{\frac{324 - 10^2/12}{11}} = 5{,}36$$

$\alpha = 0{,}05$; test unilatéral; $n = 12$; $D = 11$. La valeur critique de t est 1,80.

$$t = \frac{0{,}833 - 0}{5{,}36/\sqrt{12}} = 0{,}54.$$

H_0 est acceptée, il n'y a pas de gain de poids.

19. H_0 : $\mu_X - \mu_Y = 5$; H_1 : $\mu_X - \mu_Y > 5$. $D = 10 + 10 - 2 = 18$, $\alpha = 0{,}05$, valeur critique de $t = 1{,}73$.

$$\sigma_{(\bar{X}-\bar{Y})} = \sqrt{\frac{9 * 6^2 + 9 * 5^2}{18} * \left(\frac{1}{10} + \frac{1}{10}\right)} = 2{,}47; t = \frac{(100 - 94) - 5}{2{,}47} = 0{,}405$$

H_0 est acceptée.

Chapitre 13

χ^2 (khi — carré)

Dans les quatre derniers chapitres, nous avons appris à trouver des intervalles de confiance ou à faire des tests d'hypothèse sur la moyenne et les proportions d'une population. L'inférence sur la moyenne utilisait des données numériques (c'est un non-sens de parler de moyenne avec des données à caractères spécifiques); on peut faire de l'inférence statistique sur des proportions en utilisant deux données à caractères spécifiques. Dans ce chapitre, nous apprendrons à faire de l'inférence statistique en utilisant plus de deux données à caractères spécifiques.

Une nouvelle famille de distributions de probabilité, χ^2, est utilisée; tout comme pour la distribution t, *il y a autant de membres que de nombres de degré de liberté. Deux types de problèmes seront discutés: l'ajustement de courbes qui consiste à vérifier si une distribution de fréquences observées se comporte comme une distribution théorique (binômiale, normale) et finalement un problème qui consiste à vérifier si deux variables sont indépendantes ou non.*

13.1 INTRODUCTION

La lettre grecque χ (khi) est utilisée et comme on l'utilise **seulement** sous la forme χ^2, on dira khi-deux ou khi-carré.

Nous avons déjà rencontré différentes familles de distributions de probabilités: la binômiale, la normale et la distribution *t* de Student. Nous allons étudier maintenant une nouvelle distribution, celle du χ^2. Comme pour la distribution *t*, elle possède autant de courbes qu'il y a de degrés de liberté. Avant d'en faire l'étude, examinons un problème qui justifie son utilisation.

Exemple 1

Une étude faite en 1950 montre que parmi tous les hommes de 30 ans, 20% avaient complété leurs études collégiales, 50% avaient complété leur secondaire, mais pas le collégial, et 30% n'avaient pas complété leur secondaire. Aujourd'hui avons-nous la même distribution? Une étude portant sur 1000 hommes âgés de 30 ans choisis au hasard donne les résultats suivants:

250 ont complété leur collégial, 520 leur secondaire, mais pas leur collégial, et 230 n'ont pas complété leur secondaire. Faire un test au niveau de signification de 0,01.

H_0 : Il n'y a pas de changement entre 1950 et aujourd'hui (remarquer que H_0 est une hypothèse portant sur la distribution de la population et non sur la moyenne ou la proportion).

H_1 : La population actuelle (les hommes de 30 ans) a une distribution différente de celle de 1950.

Soit O, la fréquence observée actuellement et E, la fréquence théorique. S'il n'y a pas de changement dans les distributions entre 1950 et maintenant, les informations peuvent être présentées comme suit:

	O	E
Études collégiales	250	200
Études secondaires	520	500
Secondaire non complété	230	300
	1000	1000

En 1950, 20% avaient complété leurs études collégiales donc si on prend un échantillon de 1000 hommes, théoriquement, 200 ont dû compléter leur collégial. Toujours théoriquement, 50% représente 500 hommes et 30%, 300 hommes. Il est important que les colonnes O et E aient la même taille, c'est pour cette raison qu'on transforme les pourcentages de E en données numériques.

Une décision doit être prise; quel critère permettra de rejeter ou d'accepter H_0? Est-ce que $\Sigma(O-E)$ permet d'établir la différence des deux distributions? $\Sigma(O-E) = \Sigma O - \Sigma E = n - n = 0$. on peut le constater dans la colonne $O - E$ ci-dessous:

	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Études collégiales	250	200	50	2500	12,5
Études secondaires	520	500	20	400	0,8
Secondaire non complété	230	300	−70	4900	16,3
	1000	1000	0	7800	29,6

On se rappellera que lors de l'étude de la variance $\frac{\Sigma(X-\bar{X})}{n} = 0$ et pour cette raison il nous a fallu prendre $\frac{\Sigma(X-\bar{X})^2}{n}$. Pour la même raison $\Sigma(O-E)^2$ permet de voir la différence entre les deux distributions si $\Sigma(O-E)^2 = 0$, alors les deux distributions sont parfaitement ajustées. Plus $\Sigma(O-E)^2$ est grande, plus il y a disparité entre les distributions. En pratique, on utilise $\Sigma(O-E)^2/E$ de façon à diminuer l'importance d'une grande valeur dans la colonne $(O-E)^2$ en tenant compte de sa fréquence théorique. Ainsi une valeur de 2500 dans la colonne $(O-E)^2$ devient 12,5 si E est 200, mais devient 5,0 si $E = 500$.

Nous définissons: $\chi^2_p = \sum \frac{(O - E)^2}{E}$

L'indice p est utilisé en l'honneur de Karl Pearson, qui développa cette formule. Cet indice indique également que χ^2_p est une statistique. Dans la prochaine section, nous étudierons une nouvelle distribution appelée χ^2. La distribution de tous les χ^2_p de tous les échantillons de taille n d'une population est approximativement une distribution du χ^2. Dans la plupart des ouvrages, l'indice p n'est pas utilisé quand il s'agit d'une statistique. Ici, on conserve cet indice pour bien indiquer la différence entre la statistique et la distribution.

Dans l'exemple précédent, on a obtenu $\Sigma(O-E)^2/E$ et $\chi^2_p = 29{,}6$. Si on examine la table du χ^2 (table 6 appendice C) on remarque que la colonne de gauche est D tout comme pour la distribution t. D représente le degré de liberté. Pour déterminer D, on examine la colonne E de notre exemple; il y a 3 valeurs et la somme est 1000, il suffit alors d'en connaître 2 pour trouver la 3e, donc $D = 3 - 1 = 2$.

	Probabilités (aire sous la courbe χ^2 à droite de la valeur donnée de χ^2)			
α	0,10	0,05	0,02	0,01
D	Valeurs de χ^2			
1	2,71	3,84	5,41	6,63
→ 2	4,61	5,99	7,02	9,21
3	6,75	7,82	9,84	11,3

$\alpha = 0{,}01$; $D = 2$; en suivant la ligne $D = 2$ jusqu'à la colonne $\alpha = 0{,}01$, on trouve $\chi^2 = 9{,}21$. Ceci nous donne la valeur critique de décision pour l'hypothèse nulle; H_0 est acceptée si $\chi^2_p \leq \chi^2$. Ici, $\chi^2_p = 29{,}6$ et $\chi^2 = 9{,}21$ donc $\chi^2_p > \chi^2$; H_0 est rejetée et H_1 acceptée. Il y a une différence significative entre l'instruction des hommes de 30 ans d'aujourd'hui et ceux de 1950. Examinons de plus près la table du χ^2.

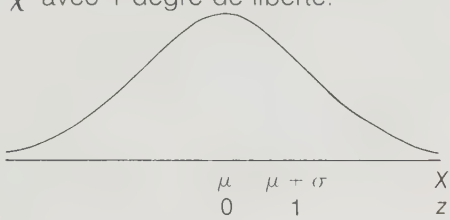
3.2 DISTRIBUTION DU χ^2

D'une distribution normale de résultats X ayant une moyenne μ et un écart-type σ, considérons tous les échantillons de taille 1 ($n = 1$). Pour chaque échantillon déterminons

$$z^2 = \frac{(X - \mu)^2}{\sigma^2}.$$

Le nombre de z^2 obtenus est égal au nombre de résultats X de la population. Tous z^2 forment une distribution, la distribution du χ^2 avec 1 degré de liberté.

$\chi^2 = z^2$ quand $D = 1$.



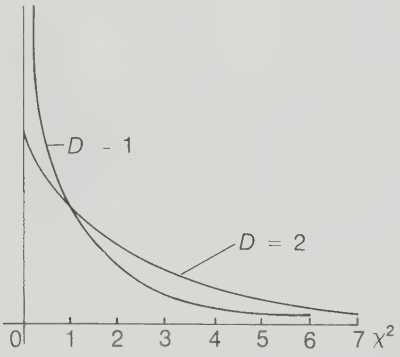
À quoi ressemble la courbe du χ^2 quand $D = 1$?

1. Les valeurs de z s'étendent de $-\infty$ à ∞ z^2 est toujours positif donc les valeurs varient de 0 à ∞ (il en est de même pour χ^2).
2. 68% de l'aire sous la courbe z est compris entre -1 et 1. Donc 68% de l'aire sous la courbe χ^2 est compris entre 0 et 1 quand $D = 1$.
3. 95,4% de l'aire sous la courbe z est compris entre -2 et 2. Donc 95,4% sous la courbe χ^2 est compris entre 0 et 4 quand $D = 1$.

Pour χ^2 avec deux degrés de liberté, il faut prendre tous les échantillons de taille 2 de la population. Si un échantillon comprend les résultats X_1 et X_2, cet échantillon donne un résultat χ^2 de:

$$\frac{(X_1 - \mu)^2}{\sigma^2} + \frac{(X_2 - \mu)^2}{\sigma^2} = z_1^2 + z_2^2.$$

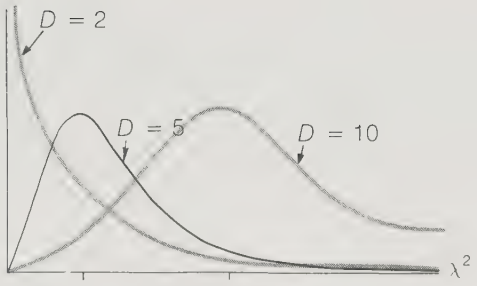
On effectue ce calcul pour tous les échantillons de taille 2 et on trouve la distribution du χ^2 pour $D = 2$. Ici il est difficile de trouver l'aire sous la courbe en examinant les probabilités. Si $z_1 < 1$ et $z_2 < 1$ alors $z_1^2 + z_2^2 < 2$, mais $p\ (\chi^2 < 2) \neq p\ (z_1 < 1$ et $z_2 < 1)$; en effet il y a d'autres possibilités; $z_1 \simeq 0$ et $z_2 \leq \sqrt{2}$ et χ^2 sera plus petit que 2.



De façon analogue la distribution du χ^2 pour $D = 3$ est la distribution de $z_1^2 + z_2^2 + z_3^2$ ou z_1, z_2 et z_3 sont les cotes z de chaque élément X d'un échantillon de taille 3.

Les distributions du χ^2 ont en général les caractéristiques suivantes:

1. chaque distribution du χ^2 s'étend de 0 à l'infini;
2. chaque distribution du χ^2 a une seule queue à droite, nous ne ferons alors que des **tests unilatéraux**
3. au fur et à mesure que D augmente, les courbes épousent de plus en plus la forme de cloche et s'approchent de la normale.



courbes du χ^2 pour différentes valeurs de D

Quelle est la relation entre χ^2 qu'on vient de définir et $\chi^2_p = \sum \frac{(O-E)^2}{E}$ utilisé dans l'exemple 1?

Le χ^2_p est une statistique (l'échantillon est utilisé dans le calcul) dont la distribution est approximativement χ^2. La valeur critique de χ^2 pour l'acceptation ou le rejet de H_0 est déterminée en consultant la table 6 pour α et un D donnés puis une décision est prise selon la région dans laquelle χ^2_p se trouve. «La distribution de χ^2_p est **approximativement χ^2**» si les conditions suivantes sont respectées: au moins 3/4 des résultats de la colonne E doivent être 5 ou plus et tous les résultats doivent être plus grands que 1. Sinon il faut augmenter la taille de l'échantillon.

13.3 TEST D'EFFICACITÉ D'UN AJUSTEMENT

La comparaison des distributions des hommes de 30 ans par niveaux d'instruction, au début de ce chapitre, est un exemple **d'efficacité d'un ajustement.** Dans cette utilisation de la distribution du χ^2, les fréquences observées dans la distribution de l'échantillon sont utilisées pour vérifier l'hypothèse que la population d'où l'échantillon provient se comporte comme une certaine population connue.

H_0 prend la forme de «la population (A) d'où provient l'échantillon a la même distribution que la population (B)»; la distribution B est connue.

H_1 s'énonce «les populations A et B n'ont pas la même distribution». Un échantillon de taille n est pris dans la population A puis les fréquences ou résultats des différentes catégories sont observés (colonne O). Ensuite, les fréquences théoriques ou expérimentales pour les mêmes catégories sont calculées pour la population B (colonne E). Finalement, $\sum \frac{(O-E)^2}{E}$ est calculée.

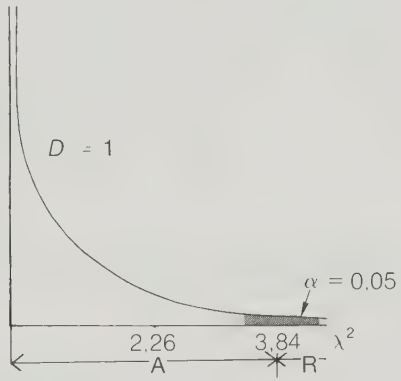
Pour tous les tests du χ^2, le nombre de degrés de liberté D est égal au nombre d'entrées de la colonne E moins le nombre de statistiques d'échantillonnage utilisées pour déterminer E. Dans les exemples de cette section, une seule statistique d'échantillonnage est utilisée, à savoir le nombre total de résultats dans l'échantillon donc **D = (nombre d'entrées de E) − 1.**

La valeur critique de χ^2 est déterminée par la table 6, appendice C, pour une valeur de α et de D. Cette valeur critique est ensuite comparée avec la statistique calculée $\chi^2_p = \sum \frac{(O-E)^2}{E}$ puis, une décision est prise.

Exemple 1

Un examen de 400 bébés choisis au hasard montre que 185 sont des filles. Est-ce que la naissance d'un garçon ou d'une fille est également probable en se basant sur cette expérience ($\alpha = 0{,}05$)?

$H_0: P = 1/2$; $H_1: P \neq 1/2$.



	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Filles	185	200	−15	225	1,13
Garçons	215	200	15	225	1,13
	400	400			$2{,}26 = \chi^2_P$

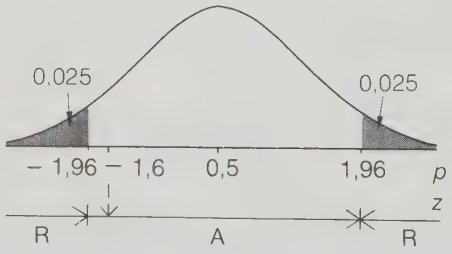
Il y a deux entrées dans la colonne E; une statistique d'échantillonnage (n) est utilisée pour compléter la colonne E, donc $D = 2 - 1 = 1$. La valeur critique de χ^2 pour $\alpha = 0{,}05$ et $D = 1$ est 3,84. H_0 est acceptée car $2{,}26 < 3{,}84$; la probabilité de naissance d'une fille est 1/2.

Ce problème peut se faire également en utilisant la méthode du chapitre 11.

$H_0: P = 1/2$; $H_1: P \neq 1/2$; $\alpha = 0{,}05$: les valeurs critiques de z sont $\pm 1{,}96$

$$p = 185/400 = 0{,}46, n = 400, \sigma_p = \sqrt{PQ/n} = \sqrt{0{,}5*0{,}5/400} = 0{,}025$$

$$z = \frac{0{,}46 - 0{,}50}{0{,}025} = -1{,}6$$



H_0 est acceptée. Comme on peut le voir, il y a souvent plusieurs façons de faire un problème. Il est à remarquer qu'on a utilisé un test bilatéral dans le cas de z et un test unilatéral pour χ^2.

Exemple 2

Un ingénieur industriel bâtit un modèle permettant d'étudier le nombre de

rames de papier (20 mains de 500 feuilles) utilisées par un grand bureau de dactylographie. Son modèle comprend 5 événements: 1; 2; 3; 4; 5 correspondant au nombre de mains de papier utilisées. Les probabilités de ces événements sont 0,10; 0,30; 0,40; 0,10; 0,10; respectivement. Pour vérifier son modèle, le nombre de rames utilisées est enregistré pendant 300 jours. Il trouve qu'on utilise 1 rame en 20 jours, 2 en 60 jours, 3 en 110 jours, 4 en 50 jours, et 5 en 60 jours. Son modèle est-il exact au niveau de confiance de 0,10?

X	O	$p(X)$	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
1	20	0,1	30	−10	100	3,3
2	60	0,3	90	−30	900	10,0
3	110	0,4	120	−10	100	0,8
4	50	0,1	30	20	400	13,3
5	60	0,1	30	30	900	30,0
	300		300			$57,4 = \chi^2_p$

La colonne E est calculée en multipliant p (X) par 300. Ainsi, la probabilité que 1 rame soit utilisée est 0,1 donc en 300 jours on prévoit utiliser 1 rame en $0,1 * 300 = 30$ jours.

Il y a 5 entrées dans la colonne E; $D = 5 - 1 = 4$ pour $\alpha = 0,10$ et $D = 4$, $\chi^2 = 7,78$. H_0 est rejetée; l'ingénieur devrait construire un autre modèle.

Dans plusieurs expériences le rejet de H_0 nous permet d'affirmer qu'il y a une différence significative entre les distributions. En vérifiant l'efficacité d'un ajustement avec χ^2_p, on espère que l'hypothèse nulle est vraie. Si on est un expérimentateur prudent, on voudra que la probabilité β d'accepter H_0 quand elle est fausse soit petite. Pour cette raison, plusieurs statisticiens utilisent $\alpha = 0,20$ ou $\alpha = 0,25$, car β diminue si α augmente pour un n donné.

13.4 RELATION ET INDÉPENDANCE DES VARIABLES; TABLEAUX DE CONTINGENCES

Dans la section précédente, on a vu comment χ^2_p permet de vérifier l'efficacité d'un ajustement. La distribution d'une population est-elle semblable à une distribution théorique?

χ^2_p peut être utilisé aussi pour vérifier si deux variables sont en relation ou indépendantes. Est-ce que la proportion des hommes qui votent libéral est plus grande que celle des femmes ou y a-t-il une relation entre le sexe et le parti politique? Y a-t-il une relation entre les années de scolarité post-collégiales et les impôts à l'âge de 40 ans? Un test du χ^2 aide à répondre à ce genre de problèmes.

Exemple 1

La table suivante donne la relation entre le nombre d'accidents en une année et l'âge des conducteurs dans un échantillon de 500 conducteurs âgés de 18 à 50 ans. Vérifier au niveau de signification de 0,01 l'hypothèse que le nombre d'accidents est indépendant de l'âge du conducteur.

Fréquences observées

Nombre d'accidents	Âge du conducteur 18—25	26—40	41—50	
0	75	120	105	300
1	50	60	40	150
2	25	20	5	50
	150	200	150	500

Il y a 75 conducteurs âgés de 18 à 25 ans qui n'ont pas eu d'accident, 120 entre 26 et 40 ans et ainsi de suite.

Un tel tableau est appelé «**tableau de contingences**». C'est un tableau de 3 × 3 car il y a 3 lignes et 3 colonnes. Remarquer que les totaux ne contribuent pas à la dimension du tableau. Le tableau suivant a une dimension de 2 × 4 (2 lignes, 4 colonnes).

L'hypothèse nulle pour ce genre de problème affirme toujours que les classifications sont indépendantes (ici, qu'il n'y a pas de relation entre l'âge des conducteurs et le nombre d'accidents). H_1 : les variables sont dépendantes, $\alpha = 0{,}01$.

Quelles sont les fréquences théoriques? Concentrons-nous sur la case supérieure gauche en se rappelant que si A et B sont des événements indépendants, alors p (A et B) $= p$ (A) $* p$ (B). Si un conducteur est choisi au hasard dans le groupe de 500, la probabilité qu'il ait entre 18 et 25 ans est de 150/500 et la probabilité qu'il n'ait pas d'accident est de 300/500. Si H_0 est vraie (l'âge et le nombre d'accidents sont indépendants), alors p (âgé de 18 à 25 ans et pas d'accident) $= p$ (âgé de 18 à 25 ans) p (pas d'accident) $= \frac{150}{500} * \frac{300}{500}$. S'il y a 500 conducteurs, théoriquement, la 1re case devrà contenir $\frac{150}{500} * \frac{300}{500} * 500 = 90$ conducteurs âgés de 18 à 25 ans et qui n'ont pas eu d'accident. Ce résultat peut être obtenu en multipliant le total de la colonne (150) par le total de la ligne (300) divisé par le grand total (500). En procédant de la même manière, on peut trouver les fréquences théoriques des autres cas.

Solution

p (26 – 40 et 0 accident) $= \frac{200}{500} * \frac{300}{500}$, et le nombre de conducteurs dans

cette catégorie sera $\frac{200}{500} * \frac{300}{500} * 500 = 120$; [total de la colonne (200) * total de la ligne (300) ÷ le grand total (500)].

On peut résumer les calculs dans le tableau suivant:

Fréquences théoriques si H_0 est vraie

Âge des conducteurs

Nombre d'accidents	18—25	26—40	41—50	total
0	$\frac{150 * 300}{500} = 90$	$\frac{200 * 300}{500} = 120$	$\frac{150 * 300}{500} = 90$	300
1	$\frac{150 * 150}{500} = 45$	$\frac{200 * 150}{500} = 60$	$\frac{150 * 150}{500} = 45$	150
2	$\frac{150 * 50}{500} = 15$	$\frac{200 * 50}{500} = 20$	$\frac{150 * 50}{500} = 15$	50
	150	200	150	500 grand total

Noter que les fréquences marginales (totales) de ce tableau sont les mêmes que dans le tableau des fréquences observées. Pour trouver la fréquence théorique de chaque case, **on multiplie les fréquences marginales de la ligne et de la colonne, puis on divise par le grand total.**

On peut maintenant trouver la valeur de χ^2_p comme précédemment:

$$\chi^2_P = \sum \frac{(O-E)^2}{E}$$

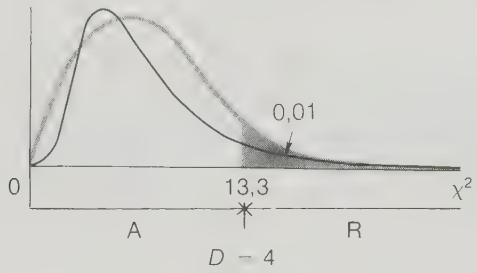
Nombre d'accidents	Âge des conducteurs	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
0	18–25	75	90	−15	225	2,5
0	26–40	120	120	0	0	0
0	41–50	105	90	15	225	2,5
1	18–25	50	45	5	25	0,6
1	26–40	60	60	0	0	0
1	41–50	40	45	− 5	25	0,6
2	18–25	25	15	10	100	6,7
2	26–40	20	20	0	0	0
2	41–50	5	15	−10	100	6,7
				0		19,6 $= \chi^2_P$

Quel est le degré de liberté D? Combien devons-nous connaître de cases pour pouvoir compléter le tableau, les fréquences marginales étant connues? Examinons le tableau suivant:

√	√	X	140
√	X	√	252
X	X	X	108
100	300	100	

Si deux cases de la première ligne sont connues (√), alors la fréquence de la 3e case peut être déterminée (**X**). De même si 2 cases de la deuxième ligne sont connues, alors la 3e case peut être déterminée. Finalement, les cases de la 3e ligne peuvent toutes être déterminées, car on a les fréquences marginales de chaque colonne qui sont (100; 300; 100). Il suffit donc de connaître 4 cases (jamais 3 cases dans une même ligne ou colonne) pour compléter le tableau; donc $D = 4$. La valeur critique de χ^2 (table 6, appendice C) pour $D = 4$ et $\alpha = 0{,}01$ est 13,3; comme la valeur de $\chi^2_p = 19{,}6$ est supérieure à 13,3, H_0 est rejetée.

Il y a une relation entre le nombre d'accidents et l'âge du conducteur.



Degré de liberté d'un tableau de contingences

Si les fréquences marginales d'un tableau 2 × 2 sont données, il suffit de connaître une seule case pour déterminer les 3 autres. Dans l'exemple suivant, s'il y a un 9 dans la 1re case, on obtient:

(1) √	(2) X	10
(3) X	(4) X	40
20	30	50

9	1	10
11	29	40
20	30	50

Un tableau 2 × 2 a donc un degré de liberté $D = 1$. Soit le tableau 3 × 4 suivant, combien de cases devons-nous connaître pour compléter le tableau?

(1)	(2)	(3)	(4)	25
(5)	(6)	(7)	(8)	50
(9)	(10)	(11)	(12)	25
10	20	30	40	100

Si les cases (1), (2) et (3) sont connues, la 4e est déterminée puisque la somme est 25. Si les cases (5), (6) et (7) sont connues alors la case (8) est déterminée. Toutes les cases de la dernière ligne de même que celles de la colonne de droite peuvent être déterminées puisque toutes les fréquences marginales sont connues. Dans cet exemple, le nombre de degrés de liberté est donné par $D = (3 - 1) * (4 - 1) = 6$. **En général, le nombre de degrés de liberté d'un tableau de $n * m$ est donné par:**

$$D = (n - 1)(m - 1)$$

Exemple 2

Y a-t-il une relation entre le fait de porter des lunettes et la scolarité? Un échantillon de 50 finissants du collégial montre qu'il y en a 15 qui portent des lunettes. Soixante-dix des 200 finissants du secondaire portent des lunettes et seulement 28 adultes sur 100 qui ont fait un secondaire professionnel portent des lunettes. Faire un test au niveau de signification de 0,05.

Les fréquences observées donnent le tableau 2 × 3 suivant:

Fréquences observées

	Scolarité			
	Collège	Secondaire	Secondaire Professionnel	
Lunettes	15	70	28	113
Sans lunettes	35	130	72	237
	50	200	100	350

Pour trouver les fréquences théoriques, s'il n'y a pas de relation entre la scolarité et le fait de porter des lunettes, il suffit de conserver les fréquences marginales puis, pour trouver la fréquence d'une case, de multiplier les fréquences marginales de la ligne et de la colonne correspondantes puis de diviser par le grand total.

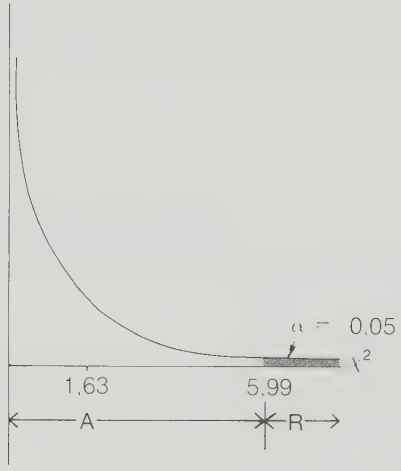
Fréquences théoriques

	Collège	Scolarité Secondaire	Secondaire Professionnel	
Lunettes	$\frac{50 * 113}{350}$ 16,1	$\frac{200 * 113}{350}$ 64,6	$\frac{100 * 113}{350}$ 32,3	113
Sans lunettes	$\frac{50 * 237}{350}$ 33,9	$\frac{200 * 237}{350}$ 135,4	$\frac{100 * 237}{350}$ 67,7	237
	50	200	100	350

Calculons χ^2_p

lunettes	Scolarité	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Oui	Collège	15	16,1	− 1,1	1,21	0,08
Oui	Secondaire	70	64,6	5,4	29,2	0,45
Oui	Sec. Pro.	28	32,3	− 4,3	18,5	0,57
Non	Collège	35	33,9	1,1	1,21	0,04
Non	Secondaire	130	135,4	− 5,4	29,2	0,22
Non	Sec. Pro.	72	67,7	4,3	18,5	0,27
				0		1,63 $= \chi^2_p$

Il y a $D = (2 - 1) * (3 - 1) = 2$ degrés de liberté $\alpha = 0{,}05$, $\chi^2 = 5{,}99$ comme $\chi^2_p < \chi^2$. H_0 est acceptée; il n'y a pas de relation entre le fait de porter des lunettes et le degré de scolarisation.



Exemple 3

Le psychologue d'une polyvalente trouve qu'il y a des étudiants qui éprouvent des difficultés à cause de l'usage de l'alcool, de colle ou de marijuana. Il définit avec soin le terme «consommateur de drogue» puis recueille les renseignements suivants d'un échantillon aléatoire de 100 étudiants consommateurs de drogue.

Type de drogue	Degré Secondaire 1	2 et 3	4 et 5	
Alcool	16	12	12	40
Colle	8	15	7	30
Marijuana	6	8	16	30
	30	35	35	100

Au niveau de signification de 0,01, y a-t-il une relation entre le degré du secondaire et le type de drogue consommée?

H_0 : il n'y a pas de relation.

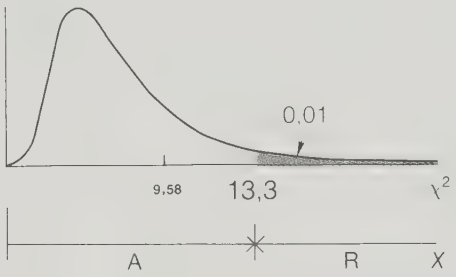
H_1 : il y a une relation.

Fréquences théoriques

	Degré Secondaire 1	2 et 3	4 et 5
Alcool	$\frac{30 * 40}{100} = 12$	$\frac{35 * 40}{100} = 14$	$\frac{35 * 40}{100} = 14$
Colle	$\frac{30 * 30}{100} = 9$	$\frac{35 * 30}{100} = 10{,}5$	$\frac{35 * 30}{100} = 10{,}5$
Marijuana	$\frac{30 * 30}{100} = 9$	$\frac{35 * 30}{100} = 10{,}5$	$\frac{35 * 30}{100} = 10{,}5$

Type de drogue	Degré	O	E	$O - E$	$(O - E)^2$	$(O - E)^2/E$
Alcool	1	16	12	4	16	1,33
Alcool	2 et 3	12	14	− 2	4	0,29
Alcool	4 et 5	12	14	− 2	4	0,29
Colle	1	8	9	− 1	1	0,11
Colle	2 et 3	15	10,5	4,5	20,2	1,92
Colle	4 et 5	7	10,5	− 3,5	12,2	1,16
Marijuana	1	6	9	− 3	9	1,00
Marijuana	2 et 3	8	10,5	− 2,5	6,25	0,60
Marijuana	4 et 5	16	10,5	5,5	30,25	2,88
				0		9,58 = χ^2_p

$D = (3 - 1) * (3 - 1) = 2 * 2 = 4$, $\alpha = 0{,}01$ donc $\chi^2 = 13{,}3$. H_0 est acceptée; il n'y a pas de relation entre le degré et le type de drogue consommée.



13.5 CORRECTION DE YATES

Quand $D = 1$, la distribution de $\chi^2_P = \sum \frac{(O-E)^2}{E}$ est approximativement celle de χ^2. Un statisticien nommé Yates a découvert que la distribution $\chi^2_p = \Sigma \frac{(|O-E| - 0{,}5)^2}{E}$ est une meilleure approximation de χ^2 quand $D = 1$.

Pourquoi cette correction? Pour les mêmes raisons que celles énoncées précédemment à la section 7.11. C'est une correction de continuité: en effet χ^2_p est une distribution discrète (fréquence de cases) et χ^2 est continue. Si D est plus grand que 1 la correction n'est pas nécessaire.

Exemple 1

Est-ce que les variables du tableau suivant sont indépendantes ($\alpha = 0{,}05$)?

	C	D	
A	40	105	145
B	60	95	155
	100	200	300

O	E	$(O-E)$	$\lvert O-E\rvert-0{,}5$	$(\lvert O-E\rvert-0{,}5)^2$	$(\lvert O-E\rvert-0{,}5)^2/E$
40	48,3	8,3	7,8	60,84	1,26
105	96,7	8,3	7,8	60,84	0,63
60	51,7	8,3	7,8	60,84	1,18
95	103,3	8,3	7,8	60,84	0,59
					3,66 $= \chi^2_p$

La valeur critique de χ^2 pour $D = 1$ et $\alpha = 0{,}05$ est 3,84, donc H_0 est acceptée. La correction de Yates est utilisée **seulement** quand $D = 1$ et peut être utilisée pour les tests impliquant des tableaux de contingences ou des ajustements de distributions.

13.6 VOCABULAIRE ET SYMBOLES

Distributions du khi-carré χ^2, χ^2_p
Efficacité d'ajustement
Cases
Fréquences théoriques E
Fréquences observées O
Degré de liberté D
Tableau de contingences
Correction de Yates

13.7 EXERCICES

1. Dans une distribution du χ^2 où $D = 10$, quelle est la probabilité qu'un résultat soit plus grand que 18,30?

2. Dans une distribution du χ^2 avec 8 degrés de liberté (a) 1% des résultats sont supérieurs à quelle valeur? (b) 90% sont inférieurs à quelle valeur?

3. Combien de degrés de liberté possède une distribution du χ^2 si la probabilité est 0,05 qu'un résultat de la distribution soit plus grand que 26,3?

4. Un dé est lancé 600 fois; voici les résultats obtenus.

Face	1	2	3	4	5	6
Fréquence	95	110	90	85	112	108

Le dé est-il régulier au niveau de $\alpha = 0{,}05$?

5. Une enquête auprès de 160 familles ayant 4 enfants montre que 6 familles n'ont pas de fille, 34 ont une fille, 56 en ont deux, 46 en ont 3, et 18 en ont 4. Se basant sur cette enquête peut-on affirmer que la probabilité de naissance d'un garçon ou d'une fille est la même ($\alpha = 0{,}05$)?

6. Une boîte de menthes digestives contient 40 menthes vertes, 60 blanches, 50 jaunes, 50 roses et 40 oranges. À l'emballage, deux distributrices sont utilisées. La distributrice *A* opère avec les pourcentages suivants: 20% de menthes vertes, 25% de blanches, 20% de jaunes, 15% de roses, 20% d'oranges. La distributrice *B* opère avec les pourcentages respectifs suivants: 15; 30; 20; 20; 15. Calculer χ^2_p pour chaque distributrice puis déterminer si la boîte du début a été remplie par la distributrice *A* ou par la distributrice *B*.

7. Un croisement entre des tomates rouges à grandes feuilles et des tomates roses à feuilles naines donne les résultats suivants:

Rouges à grandes feuilles	280
Roses à grandes feuilles	100
Rouges à feuilles naines	80
Roses à feuilles naines	20

Est-ce que ces résultats vérifient l'hypothèse que les proportions sont 9; 3; 3; 1 (α = 0,10)?

8. Un examen de 1 000 pare-chocs montre que 300 pare-chocs avant sont bosselés à gauche et 250 pare-chocs avant sont bosselés à droite. Au niveau de signification de 0,01 est-ce qu'il y a autant de pare-chocs bosselés en avant à droite, qu'en avant à gauche?

9. Un manufacturier de télévisions achète ses transistors de quatre fournisseurs. Chaque mois 1 000 transistors de chaque fournisseur sont choisis au hasard puis testés. En janvier, 46; 53; 42; 59 transistors défectueux furent trouvés dans les échantillons des fournisseurs *A*, *B*, *C*, *D* respectivement. Est-ce que le nombre de transistors défectueux est le même pour chacun des fournisseurs (α = 0,05)?

●10. Dans un examen de Vrai ou Faux, un professeur ajoute 5 questions qui, croit-il, ne peuvent être répondues que par devinette. Voici les résultats obtenus à ces 5 questions.

Nombre de bonnes réponses	Nombre d'étudiants
0	1
1	10
2	20
3	25
4	15
5	4

Son appréhension est-elle justifiée au niveau de 0,10?

11. Le professeur du problème précédent reprend la même expérience avec une classe plus nombreuse; cette fois le nombre d'étudiants qui a répondu correctement à 0; 1; 2; 3; 4; 5 questions est 2; 20; 40; 50; 30; 8 respectivement. Son appréhension est-elle justifiée au niveau de 0,01?

12. Une association de chimistes fait parvenir un questionnaire à ses membres; voici les résultats obtenus:

	Questionnaires		
Diplôme	retournés	non retournés	total
B.S.	50	40	90
B.Chem.E.	100	90	190
M.S.	50	40	90
M.Chem.E.	100	90	190
Ph.D	200	240	440

Est-ce que la proportion de questionnaires retournés est indépendante du diplôme (α = 0,10)?

13. Dans un échantillon aléatoire de 100 immigrants à Montréal, 50 ont des revenus inférieurs à \$10 000; 40 gagnent entre \$10 000 et \$15 000; et 10 gagnent plus de \$15 000 par année. D'un échantillon de 100 Montréalais, 30 gagnent moins de \$10 000; 40 entre \$10 000 et \$15 000; 30 plus de \$15 000 par année. Au niveau de signification de 0,01, est-ce que le revenu est en relation avec la citoyenneté?

14. Se basant sur les données suivantes, on veut vérifier s'il y a une relation entre le parti politique et l'attitude face à un projet de loi ($\alpha = 0{,}05$).

	en faveur	contre	indécis
Libéraux	100	150	150
Unionistes	150	100	150

15. Trois marques de fixatif à cheveux sont vérifiés. Parmi les blondes qui les utilisent 20 préfèrent la marque *A*, 30 la marque *B* et 10 la marque *C*. Parmi les brunettes 30; 30 et 20 préfèrent les marques *A*; *B*; *C* respectivement. Parmi les rousses, 10 préfèrent la marque *A*, 5 la marque *B* et 5 la marque *C*. Au niveau de signification de 0,025, y a-t-il une relation entre la marque du fixatif et la couleur des cheveux?

16. Des bâtons de rouge à lèvres sont fabriqués par trois procédés différents (*A, B* et *C*). Les bâtons défectueux sont classés de la façon suivante: couleur, consistance, bâton ébréché. Au niveau de signification de 0,05, est-ce que les données suivantes permettent de dire si les défauts sont indépendants du procédé utilisé?

		Défauts		
		Couleur	Consistance	Bâton ébréché
Procédé	*A*	40	25	20
	B	10	30	20
	C	10	25	20

17. La production de fruits en Ontario pour la période de 1970 à 1973 exprimée en milliers de caisses est la suivante:

	Pêches	Cerises	Pommes	Fraises
1970	50	10	50	10
1971	55	8	50	12
1972	35	20	50	14
1973	60	12	50	14

Au niveau de signification de 0,05, est-ce que la production de la sorte de fruits est indépendante de l'année?

18. Un professeur affirme que les résultats obtenus à son cours sont normalement distribués. Il donne «A» à ceux qui ont un résultat 1σ au-dessus de la moyenne, «B» à ceux qui ont entre $0{,}5\sigma$ et 1σ au-dessus de la moyenne, «C» à ceux qui ont entre 1σ sous la moyenne et $0{,}5\sigma$ au-dessus, «D» à ceux qui ont entre 1σ et 2σ sous la moyenne, et «F» à ceux qui ont moins de 2σ sous la moyenne. Les 100 étudiants de sa classe ont obtenu les résultats suivants: 10A, 15B, 63C, 7D, et 5F. Au niveau $\alpha = 0{,}05$ son affirmation est-elle justifiée?

19. Quelques personnes (dégustateurs) trouvent une tisane très amère lors d'une dégustation alors que d'autres (non-dégustateurs) la trouvent sans goût ou à peu près; on pose l'hypothèse que le rapport de dégustateurs et non-dégustateurs est de 2 à 1. Dans un test de 300 personnes choisies au hasard, 87 sont trouvées non-dégustateurs et 213 sont dégustateurs. L'hypothèse est-elle vérifiée par ce test ($\alpha = 0{,}10$)?

20. Un fabricant de sacs de couchage affirme que 90% de ceux qui utilisent ses sacs seront confortables jusqu'à une température de – 10° C. Parmi 100 campeurs utilisant ces sacs en automne, 15 se sont déclarés inconfortables à une température de – 10° C. L'affirmation du fabricant est-elle justifiée au niveau de 0,10?

21. Un généticien étudie l'hérédité d'une certaine maladie. Il recueille les résultats suivants d'un échantillon aléatoire de 180 enfants.

		L'enfant a la maladie	
		Oui	Non
La mère a la maladie	oui	10	70
	non	8	92

Au niveau ($\alpha = 0{,}01$) est-ce que les enfants ont plus de chances d'avoir la maladie si leur mère a la maladie?

22. Un échantillon aléatoire de 50 personnes écoutant de la musique Pop sont interrogées pour connaître leur réaction à ce genre de musique. Voici les résultats obtenus:

		Aime	N'aime pas
Âge	25 ans ou moins	20	8
	Plus de 25 ans	10	12

Au niveau de ($\alpha = 0{,}05$) y a-t-il une relation entre l'âge des auditeurs et leur goût pour ce genre de musique?

23. Un sociologue demande à un échantillon aléatoire de 60 étudiants de répondre à 40 questions en utilisant les nombres 1 (fortement d'accord) à 5 (fortement en désaccord). Se basant sur les réponses, il classe les étudiants en deux groupes, ceux qui ont des préjugés raciaux et ceux qui n'en ont pas. Il forme également 30 autres classes doubles (le père a complété des études collégiales ou pas, homme ou femme, pratique religieuse ou pas, etc.) selon les résultats obtenus à 30 autres questions. Il fait 30 tests du χ^2 au niveau de signification de 0,05 pour déterminer si les préjugés raciaux sont en relation avec chacune des 30 autres catégories.

(a) En supposant vraie H_0 : les préjugés raciaux sont indépendants du sexe de l'étudiant; quelle est la probabilité qu'elle soit rejetée par un test du χ^2? Quelle est la probabilité qu'elle soit acceptée?

(b) Poser que chacune des hypothèses nulles qu'il pose soit vraie. Trouver la probabilité qu'elles soient toutes acceptées par un test du χ^2 si les tests sont indépendants.

(c) Poser que les 30 hypothèses nulles sont vraies et que les tests sont indépendants. Quelle est la probabilité qu'au moins une hypothèse soit rejetée?

24. Cent femmes et deux cents hommes âgés de 20 à 40 ans sont interrogés. Quinze femmes et 50 hommes sont en faveur de la peine de mort. Utiliser un test du χ^2 au niveau $\alpha = 0{,}05$ pour vérifier si les femmes et les hommes ont la même opinion sur la peine de mort.

25. Dans un échantillon de 40 téléspectateurs à Montréal, 15% regardent le canal 2 à 20:00 et 85% regardent les autres canaux. Ceci est-il en accord avec l'hypothèse que 20% des téléspectateurs regardent le canal 2 à 20:00 ($\alpha = 0{,}05$)?

RÉPONSES

1. 0,05.

3. $D = 16$.

5. Colonne E: 10; 40; 60; 40, 10; $\chi^2_p = 1,6 + 0,9 + 0,3 + 0,9 + 6,4 = 10,1$; $D = 4$; $\alpha = 0,05$; $\chi^2 = 9,49$. H_0 est rejetée; le nombre de filles dans les familles ne vérifie pas l'hypothèse que $P = 1/2$.

7. Colonne E: 270; 90; 90; 30; $\chi^2_p = 0,37 + 1,11 + 1,11 + 3,33 = 5,92$; $D = 3$; $\chi^2 = 6,25 > 5,92$. H_0 est acceptée.

9. $\chi^2_p = 0,32 + 0,18 + 1,28 + 1,62 = 3,40$; $D = 3$; $\chi^2 = 7,81$. H_0 est acceptée.

11. Colonne E: 4,7; 23,4; 46,9; 46,9; 23,4; 4,7. $\chi^2_p = 7,45$; $D = 5$; $\alpha = 0,01$; $\chi^2 = 15,1$. H_0 est acceptée. Comparer les réponses des problèmes 10 et 11 et noter ce que devient χ^2_p quand chaque fréquence de la colonne E est doublée.

13. $\chi^2_p = 2,5 + 0 + 5,0 + 2,5 + 0 + 5,0 = 15,0$; $D = (2 - 1) * (3 - 1) = 2$; $\chi^2 = 9,21$; H_0 est rejetée. Il y a une relation entre la citoyenneté et le revenu.

15. Colonne E: 22,5; 24,4; 13,1; 30,0; 32,5; 17,5; 7,5; 8,1; 4,4; $\chi^2_p = 4,9$; $D = (3 - 1) * (3 - 1) = 4$; $\chi^2 = 11,1$; H_0 est acceptée; la marque du fixatif est indépendante de la couleur des cheveux.

17. Colonne E (Pêches, Cerises, Pommes, Fraises): 48; 50; 47,6; 54,4; 12; 12,5; 11,9; 13,6; 48; 50; 47,6; 54,4; 12; 12,5; 11,9; 13,6; $\chi^2_p = 13,5$; $D = 9$; $\chi^2 = 16,9$. H_0 est acceptée; la production est indépendante de l'année.

19. (Correction de Yates) $\chi^2_p = 0,78 + 1,56 = 2,34$; $D = 1$; $\chi^2 = 2,71$. L'hypothèse est vérifiée.

21. Colonne E: 8,72; 10; 90; $\chi^2_p = 0,28 + 0,03 + 0,23 + 0,02 = 0,56$; $D = 1$; $\chi^2 = 6,63$; H_0 est acceptée, la maladie de l'enfant est indépendante de la mère.

23. (a) La probabilité de rejet $= 0,05$ (α)
Probabilité d'acceptation $= 1 - 0,05 = 0,95$.
(b) $(0,95)^{30} = 0,21$.
(c) $1 - (0,95)^{30} = 1 - 0,21 = 0,79$.

25. Colonne O: 6,36; Colonne E: 8,32 $(|O - E| - 0,5)^2 / E$: 0,28; 0,38; $\chi^2_p = 0,66$; $D = 1$; $\alpha = 0,05$; $\chi^2 = 3,84$; les résultats de l'enquête vérifient l'hypothèse.

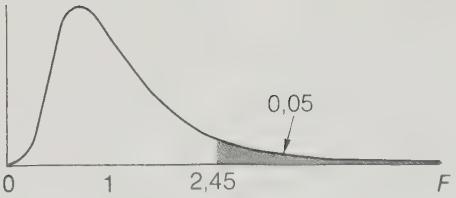
Chapitre 14

Analyse de variance

Au chapitre 11, nous avons vérifié des hypothèses concernant les différences des moyennes de deux populations. Un économiste peut utiliser cette méthode pour comparer les salaires de deux industries; un psychologue peut comparer les résultats obtenus pour un groupe expérimental avec ceux d'un groupe témoin. Il est souvent souhaitable de comparer plusieurs moyennes au lieu de deux; l'économiste veut comparer les salaires de plusieurs industries; le psychologue veut comparer plusieurs groupes au groupe témoin. La méthode qui permet de comparer plus de deux moyennes est appelée **analyse de variance.** *La connaissance de la distribution* F *est un préalable pour comprendre cette méthode. Après l'étude de cette distribution, nous l'utiliserons pour comparer les variances de deux populations, puis pour comparer plusieurs moyennes. Finalement si l'hypothèse que toutes les moyennes sont semblables est rejetée, nous verrons comment trouver quelles paires de moyennes sont sensiblement différentes.*

14.1 LA DISTRIBUTION *F*

Supposons que l'on a **deux** populations normalement distribuées et ayant la même variance. Elles peuvent avoir des moyennes différentes. Prenons dans la 1re population un échantillon de taille 9 dont la variance $s_X^2 = 5{,}0$ et dans la 2e population, un échantillon de taille 21 et dont la variance est $s_Y^2 = 2{,}0$. Le rapport $F = s_X^2/s_Y^2$ est calculé pour les deux échantillons, on a 5,0/2,0 = 2,5. Les échantillons sont remis et de nouveaux échantillons de tailles 9 et 21 sont choisis dans chaque population. Cette fois, $s_X^2 = 4{,}0$ et $s_Y^2 = 8{,}0$ alors $F = 4/8 = 0{,}5$. Ce procédé est poursuivi pour tous les échantillons possibles de tailles 9 et 21 respectivement dans chacune des populations et ce de façon indépendante (avec remise). Les résultats *F* forment une nouvelle distribution. Comme s_X^2 et s_Y^2 sont toujours positifs alors le résultat *F* est toujours positif. Puisque les deux populations ont la même variance, on s'attend à ce que beaucoup d'échantillons aient la même variance et les résultats *F* correspondants près de 1. Quelques rapports peuvent être très grands (si la variance du 1er échantillon est grande et celle du second petite). Cinq pour cent des résultats *F* seront au-dessus de 2,45.



Si on change les tailles des échantillons, par exemple de 9 et 21 qu'elles étaient à 28 et 7, on obtient alors une nouvelle distribution F différente de la première. En général, il y a une double famille de distributions F; chaque membre de la famille dépend du choix des tailles des échantillons des deux populations. Il faut se rappeler qu'une distribution F peut être utilisée **seulement si les deux** populations sont **normales,** même si les échantillons sont grands.

Toutes les distributions F ont des points communs:

1. Un résultat F est le quotient des variances de deux échantillons:

$$F = \frac{s_X^2}{s_Y^2}$$

où s_X^2 est une estimation ponctuelle de la variance d'une population basée sur la taille n_X de l'échantillon. Il en est de même pour s_Y^2.

2. Chaque distribution F est une distribution de probabilités. L'aire totale sous la courbe est 1 et la probabilité qu'un résultat F soit compris entre a et b est donnée par l'aire sous la courbe entre a et b.
3. L'étendue des valeurs F est 0 à $+\infty$.
4. Les tables de F ne sont pas données en fonction de n_X et n_Y, mais en fonction des degrés de liberté

$$D_X = n_X - 1; D_Y = n_Y - 1.$$

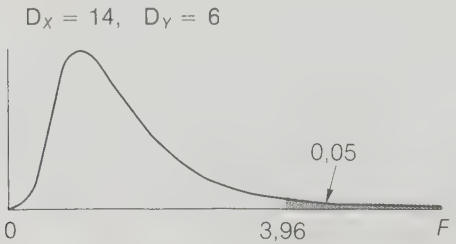
La forme de la distribution F dépend des valeurs de D_X et de D_Y. Toutes sont étalées à droite, mais deviennent symétriques si les degrés de liberté deviennent grands.

14.2 UTILISATION DE LA TABLE *F*

Puisqu'il y a tellement de distributions F, les valeurs données pour chacune sont très limitées. La table 7 appendice C donne les valeurs de F pour des valeurs d'aire de 0,05; 0,025 et 0,01 de l'extrémité droite. Un test d'hypothèse bilatéral pourra être fait seulement si $\alpha = 0{,}10$; 0,05 ou 0,02.

Exemple 1

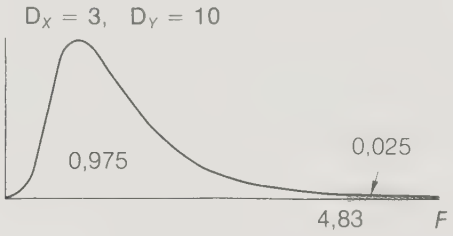
$n_X = 15$, $n_Y = 7$; 5% des résultats F sont supérieurs à quelle valeur?



$D_X = 14$, $D_Y = 6$; dans la table 7 (a) appendice C, trouver sur la 1re ligne la valeur de $D_X = 14$ puis dans la 1re colonne la valeur de $D_Y = 6$. À l'intersection de la 6e ligne et de la 14e colonne, on trouve le nombre 3,96. $p\ (F > 3{,}96) = 0{,}05$; 5% des résultats F sont supérieurs à 3,96 si les échantillons ont des tailles de 15 et 7 respectivement et sont choisis au hasard dans deux populations normales ayant la même variance. On peut résumer ceci par la notation suivante: si $\alpha = 0{,}05$ alors $F\ (14; 6) = 3{,}96$.

Exemple 2

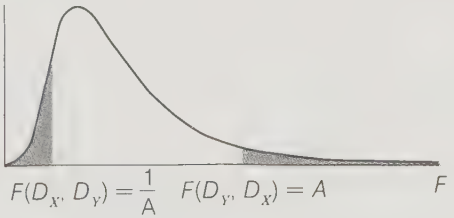
Si $n_X = 4$, $n_Y = 11$; 97, 5% des résultats F sont inférieurs à quelle valeur?



(Voir table 7 (b) appendice C.) $D_X = 3$, $D_Y = 10$; 97,5% des résultats F sont inférieurs à 4,83.

Les tests bilatéraux ou unilatéraux utilisant la partie gauche de la distribution F sont possibles. La table 7 donne seulement les valeurs de F pour des pourcentages dans la partie droite de la courbe. Nous verrons, dans les exemples 3 et 4 de la section suivante, qu'on peut éviter le calcul de gauche dans un test bilatéral. Supposons que l'on veuille quand même trouver cette valeur critique de gauche. Voici la règle à suivre:

Pour trouver la valeur critique de gauche de F pour D_X et D_Y, on trouve la valeur critique de droite de F pour D_Y et D_X (l'ordre inverse des degrés de liberté) et on prend son inverse.

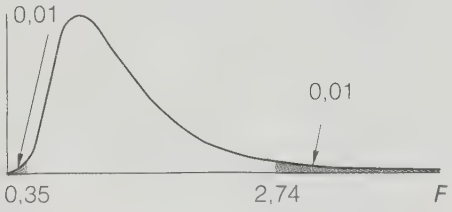


Exemple 3

Si $n_X = 7$ et $n_Y = 15$; 5% des résultats F sont inférieurs à quelle valeur? $D_X = 6$, $D_Y = 14$, $F\ (14; 6) = 3{,}96$. Voir exemple 1, $F\ (6; 14) = 1/3{,}96$ à gauche; 5% des résultats F sont inférieurs à $1/3{,}96 = 0{,}25$.

Exemple 4

$H_0 : \sigma_X^2 = \sigma_Y^2$; $H_1 : \sigma_X^2 \neq \sigma_Y^2$, $\alpha = 0{,}02$. Quelles sont les valeurs critiques de F qui déterminent la zone d'acceptation de H_0 si $n_X = 21$ et $n_Y = 25$?



Un test bilatéral est utilisé et l'aire de chaque zone de rejet est 0,01. $D_X = 20$, $D_Y = 24$; pour $\alpha = 0{,}01$, $F(20; 24) = 2{,}74$. C'est la valeur critique supérieure. Comme $F\ (24; 20) = 2{,}86$ pour $\alpha = 0{,}01$, la valeur inférieure de F est $1/2{,}86 = 0{,}35$.

14.3 UTILISATION DE LA DISTRIBUTION *F* DANS LES TESTS DE VARIANCE

Exemple 1

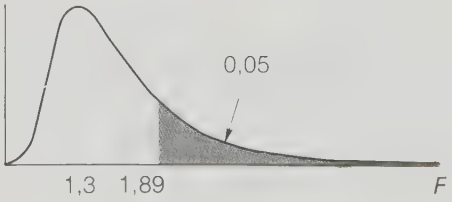
Deux distributrices sont utilisées pour remplir des boîtes de croustilles. Un échantillon aléatoire de 25 boîtes provenant de la 1re distributrice a un poids moyen de 250 g avec une variance de 8 g. Un échantillon aléatoire de 31 boîtes provenant de la 2e distributrice a un poids moyen de 255 g avec une variance de 6 g. Au niveau de signification de 0,05, est-ce que la première distributrice a plus de variabilité que la seconde? (Les poids sont normalement distribués.)

$H_0 : \sigma_X^2 = \sigma_Y^2$; $H_1 : \sigma_X^2 > \sigma_Y^2$; $\alpha = 0{,}05$ test unilatéral.

$$n_X = 25 \qquad D_X = 24 \qquad s_X^2 = 8$$
$$n_y = 31 \qquad D_Y = 30 \qquad s_Y^2 = 6$$

Les poids moyens sont donnés, mais ne sont pas nécessaires.

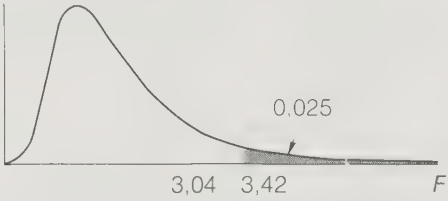
$$F = \frac{8}{6} = 1{,}33$$



La valeur critique de F est $F(24; 30) = 1{,}89$ pour $\alpha = 0{,}05$; comme 1,3 est inférieur à 1,89, H_0 est acceptée. Les deux distributrices ont la même variabilité (variance).

Exemple 2

Un psychologue pense que les résultats des enfants à un examen d'arithmétique seront plus variables à cause du stress; les enfants brillants réussiront mieux que les autres qui paniquent sous le stress. Il trouve que 19 enfants stressés ont un écart-type de 15,0 alors qu'un groupe témoin de 11 enfants non-stressés ont un écart-type de 8,6. Au niveau de signification de 0,025, est-ce que les enfants stressés ont des résultats qui ont une plus grande variabilité que ceux qui ne sont pas soumis à un stress?



$H_0: \sigma_X^2 = \sigma_Y^2$; $H_1: \sigma_X^2 > \sigma_Y^2$.

$$n_X = 19,\ D_X = 18,\ s_X^2 = 15^2 = 225$$
$$n_Y = 11,\ D_Y = 10,\ s_Y^2 = 8.6^2 = 74$$
$$F = 225/74 = 3,04$$

Pour $\alpha = 0,025$, la valeur critique de $F\ (18;10) > F\ (20;10) = 3,42$ donc H_0 est acceptée; les enfants soumis au stress ont la même variance que les autres.

Exemple 3

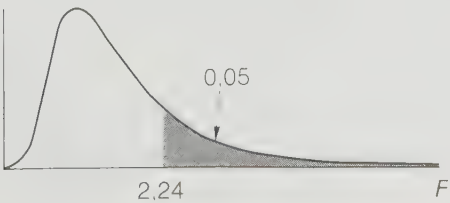
Est-ce que les salaires des professeurs invités ont une plus faible variance que celle des professeurs agrégés de l'université de Sherbrooke? Un échantillon aléatoire de 17 professeurs invités a un écart-type de $1 000 alors que 25 professeurs agrégés choisis au hasard ont un écart-type de $2 500. Vérifier au niveau de signification de 0,05.

1re solution:

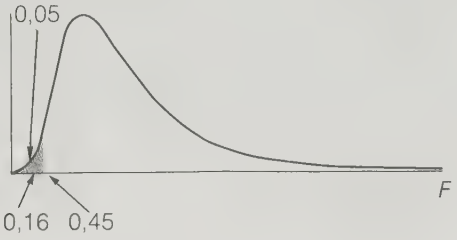
$H_0: \sigma_X^2 = \sigma_Y^2$; $H_1: \sigma_X^2 < \sigma_Y^2$; test unilatéral à gauche

$$s_X^2 = 1000^2,\ s_Y^2 = 2\,500^2;\ F = s_X^2/s_Y^2 = 0,16$$

$D_X = 16$, $D_Y = 24$. $F(24;16) = 2,24$ si $\alpha = 0,05$



Noter que $F(24; 16) = F(D_Y; D_X)$ pour un test à gauche la valeur critique est $1/2{,}24 = 0{,}45$. Comme $0{,}16 < 0{,}45$, H_0 est rejetée; il y a moins de variabilité dans les salaires des professeurs invités.



2 ieme *solution:*

La nécessité de trouver $F(D_Y; D_X)$ et de prendre l'inverse peut être évitée si on donne à la population dont l'échantillon a la plus grande variance l'indice X. Dans ce problème,

$$H_0: \sigma_X^2 = \sigma_Y^2;\ H_1: \sigma_X^2 > \sigma_Y^2;\ \text{test unilatéral à droite}$$

$$s_X^2 = 2500^2,\ s_Y^2 = 1000^2,\ F = s_X^2/s_Y^2 = 6{,}25$$

$$D_X = 24,\ D_Y = 16.\ F(24,16) = 2{,}24 \text{ si } \alpha = 0{,}05$$

Donc H_0 est rejetée.

Si le même stratagème qui consiste à nommer s_x^2 la plus grande variance est utilisée pour un test bilatéral, la plus petite valeur critique de F n'est pas nécessaire. En effet, celle-ci est sûrement inférieure à 1 puisqu'elle est l'inverse d'une des valeurs de la table qui sont toutes plus grandes que 1 alors que la valeur calculée de F sera plus grande que 1.

Note: Dans l'exemple 3, les calculs furent bien faits, les échantillons choisis au hasard et il se peut que la conclusion soit non valide. On présume que les salaires sont **distribués normalement**. Or, il se peut qu'en pratique à l'université de Sherbrooke, les salaires des professeurs soient distribués uniformément (le salaire est fonction des années d'expérience) et dans ce cas le test F sur les variances est non applicable.

Exemple 4

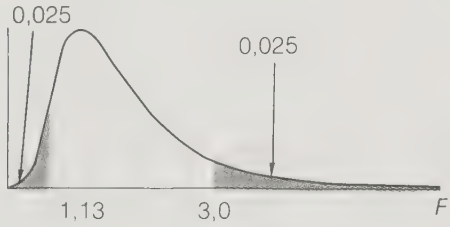
Un psychologue se demande si les enfants soumis à un stress présentent des résultats ayant une variance différente de ceux qui ne sont pas soumis à un stress. Les enfants brillants peuvent bien réussir, mais les enfants moyens peuvent paniquer et ainsi augmenter la variance. D'un autre côté, le stress n'a pas d'effet sur les enfants familiers avec le matériel, mais permet à ceux qui ont moins de connaissances ou d'habiletés de mieux réussir ce qui se traduit par une variance plus faible. Il trouve qu'un échantillon aléatoire de 15 enfants qui sert de groupe témoin a un résultat moyen de 77 avec un écart-type de 3,1, alors que 13 enfants soumis à un stress ont obtenu un résultat moyen de 82

avec un écart-type de 3,3. Au niveau de 0,05, est-ce que les deux groupes ont des variances différentes?

H_0: $\sigma_X^2 = \sigma_Y^2$; H_1: $\sigma_X^2 \neq \sigma_Y^2$; test bilatéral

$$s_X^2 = 3{,}3^2 = 10{,}89, \quad n_X = 13$$
$$s_Y^2 = 3{,}1^2 = 9{,}61, \quad n_Y = 15$$
$$F = s_X^2/s_Y^2 = 1{,}13$$

Noter que X représente les enfants soumis à un stress.



$D_X = 12$, $D_Y = 14$; $F(12; 14) = 3{,}0$ quand $\alpha = 0{,}05$. Il n'est pas nécessaire de trouver la valeur critique inférieure puisqu'elle est certainement inférieure à 1 alors que la valeur calculée de $F(1;13)$ est plus grande que 1. H_0 est acceptée; les variances des deux groupes sont identiques.

Noter: Les courbes F sont toutes différentes et dépendent de D_X et D_Y. L'illustration utilisée représente la forme générale d'une de ces courbes et permet une bonne compréhension.

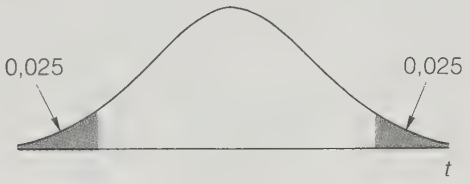
14.4 INTRODUCTION À L'ANALYSE DE VARIANCE

Nous avons appris à utiliser une distribution F pour répondre à la question «est-ce que deux populations ont la même variance?». Si **oui**, alors le rapport F de deux échantillons pris dans les populations est près de 1. Si ce rapport n'est pas près de 1, ceci indique que les populations ont probablement des variances différentes.

Maintenant, nous utiliserons la distribution F pour d'autres types d'hypothèses. Est-ce que trois populations ont la même moyenne? Supposons que des graines de tomates peuvent être traitées avec 3 types d'engrais différents ou ne pas être traitées et les récoltes de 4 méthodes sont comparées. Quatre échantillons aléatoires sont choisis, les trois premiers avec engrais et l'autre sans engrais. Le champ est divisé en 40 lots de même dimension qui sont partagés au hasard entre les 4 groupes de semences. Le rendement moyen par lot pour chaque méthode est déterminé. Voici les résultats obtenus.

Engrais 1	Engrais 2	Engrais 3	Engrais 4
$\bar{X}_1 = 96$	$\bar{X}_2 = 99$	$\bar{X}_3 = 92$	$\bar{X}_4 = 101$

Est-ce que le traitement affecte la récolte? Peut-on accepter l'hypothèse $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$?



Pour chaque paire de moyennes, on peut faire un test «t». Comme il y a quatre différentes moyennes à comparer, il y aura ${}_4C_2 = 6$ tests «t» à effectuer. Il faut se rappeler que si $\alpha = 0{,}05$, la probabilité de rejet de $H_0: \mu_X = \mu_Y$ est de 0,05. Si k tests indépendants sont faits, la probabilité de rejet d'une des hypothèses nulles est $1 - (0{,}95)^k$. Si 6 test indépendants sont faits cette probabilité est 0,26 même si $\alpha = 0{,}05$ pour chaque test. Si on fait un test sur une paire à la fois (6 tests), ceux-ci ne sont pas indépendants, alors la probabilité calculée ci-haut n'est pas exacte et est seulement une approximation. Ceci doit être suffisant pour nous convaincre que les tests «t» ne sont pas appropriés quand on a plus de 2 moyennes à comparer. On doit utiliser une autre technique qui s'appelle **l'analyse de variance.**

Supposons que 3 échantillons de taille 3 sont pris d'une même population et qu'il n'y a pas de variance dans la population. Les résultats peuvent être:

Échantillon 1	Échantillon 2	Échantillon 3
5	5	5
5	5	5
5	5	5

Si les échantillons proviennent de populations différentes, mais sans variance, les résultats peuvent être:

Échantillon 1	Échantillon 2	Échantillon 3
4	6	5
4	6	5
4	6	5

Mais inévitablement, chaque population a une variance et alors les résultats seront comme suit.

Échantillon 1	Échantillon 2	Échantillon 3
3	5	3
4	7	7
5	6	5
$\overline{X}_1 = 4$	$\overline{X}_2 = 6$	$\overline{X}_3 = 5$

Il est à noter que les résultats précédents peuvent provenir d'une même population. Comment le savoir?

Supposons qu'on a 3 groupes de 3 étudiants qui apprennent une technique statistique avec trois méthodes différentes, puis on vérifie leurs connaissances avec un problème. Voici les résultats obtenus:

1er groupe	2e groupe	3e groupe
3	5	3
4	7	7
5	6	5
$\overline{X}_1 = 4$	$\overline{X}_2 = 6$	$\overline{X}_3 = 5$

Au niveau de signification de 0,05, est-ce que les moyennes des populations sont différentes?

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: Il existe une différence entre quelques moyennes.

Si les populations desquelles les trois échantillons furent pris sont identiques, si elles ont la même moyenne et le même écart-type et si elles sont normales, alors les trois groupes peuvent être considérés comme provenant d'une même population. Les moyennes (4; 6; 5) ne sont pas identiques, mais on peut s'imaginer qu'elles peuvent être identiques. Est-ce que la variation entre les moyennes est due à la fluctuation de l'échantillonnage ou parce que les populations ne sont pas identiques?

Nous devrons faire deux estimations différentes de la variance de la population σ^2 en partant de l'hypothèse que les trois populations sont identiques et que les trois échantillons proviennent d'une seule population. La première estimation dépend de la variation des moyennes des échantillons par rapport à la moyenne obtenue en considérant tous les résultats; cette estimation dépend des différences entre les échantillons. La seconde estimation dépend de la différence des résultats d'un échantillon, de la moyenne de l'échantillon; cette estimation dépend des différences à l'intérieur de chaque échantillon et ne permet pas de comparer les échantillons entre eux. La première estimation dépend des fluctuations de l'échantillonnage et de la différence réelle entre les échantillons; si les échantillons viennent de populations ayant des moyennes différentes (si H_0 est fausse) cette estimation sera plus grande que la seconde de façon significative, puisque cette dernière dépend de la fluctuation de l'échantillon seulement. Le quotient des deux variances (leur résultat F) sera alors plus grand que 1. Si H_0 est vraie et que les trois échantillons proviennent à toute fin pratique de la même population, alors les deux estimations diffèrent seulement à cause des fluctuations d'échantillonnage et leur quotient (résultat F) sera approximativement 1. Un test «F» unilatéral à droite sera utilisé avec $\alpha = 0{,}05$.La notation utilisée sera $\boldsymbol{F\ \alpha\ (D_X; D_Y)}$. **(Il ne faut pas utiliser cette notation pour un test bilatéral.)**

La première estimation est basée sur la variance des moyennes d'échantillons et de la moyenne totale. On a $\overline{X}_1 = 4$; $\overline{X}_2 = 6$; $\overline{X}_3 = 5$. La moyenne totale $\overline{X}_T = (3 + 4 + 5 + 5 + 7 + 6 + 3 + 7 + 5)/9 = 5$. Comme on veut avoir un estimateur non-biaisé, on divisera par $(n - 1)$ ici $(3 - 1) = 2$:

$$s_{\overline{X}}^2 = \frac{(4-5)^2 + (6-5)^2 + (5-5)^2}{2} = 1.$$

Cette valeur donne une estimation de $\sigma_{\bar{X}}^2 = \sigma^2/n$ donc $\sigma^2 \approx n\, s_{\bar{X}}^2 = 3 * 1 = 3$ ($n = 3$ car chaque échantillon est de taille 3). Appelons cette valeur de σ^2 «$\hat{\sigma}^2$ **entre**» (lire σ carré chapeau entre) pour se rappeler que c'est un estimateur basé sur les quantités des échantillons.

$$\hat{\sigma}^2 \text{ entre} = 3$$

On aurait pu calculer directement la valeur de $\hat{\sigma}^2$ entre =

$$\frac{3(4-5)^2 + 3(6-5)^2 + 3(5-5)^2}{2} = 3.$$

Quand les échantillons ont des tailles différentes par exemple, $n_1 = 9$; $\bar{X}_1 = 4$ $n_2 = 9$; $\bar{X}_2 = 6$; $n_3 = 7$; $\bar{X}_3 = 5$; $\bar{X}_t = 5$; alors

$$\hat{\sigma}^2 \text{ entre} = \frac{9(4-5)^2 + 9(6-5)^2 + 7(5-5)^2}{2} = 9.$$

Nous avons besoin maintenant d'une autre estimation de σ^2. Nous avons vu précédemment que s^2 est un estimateur non biaisé de σ^2. Si les trois échantillons viennent de la même population (si H_0 est vraie), on a trois façons d'estimer σ^2: on peut utiliser s_1^2 ou s_2^2 ou s_3^2. Nous utiliserons à la place une moyenne pondérée des trois échantillons en utilisant comme poids le degré de liberté de chaque échantillon.

	1er groupe	2e groupe	3e groupe	Total
ΣX	12	18	15	45
ΣX^2	50	110	83	243
n	3	3	3	
D	2	2	2	
s^2	1	1	$4 = \dfrac{\Sigma X^2 - \dfrac{(\Sigma X)^2}{n}}{n-1}$	

$$\sigma^2 \approx \frac{2(1) + 2(1) + 2(4)}{2 + 2 + 2} = 2$$

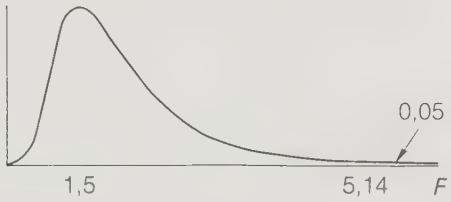
Cette estimation de σ^2 dépend de la variance à l'intérieur de chaque échantillon et sera appelé $\hat{\sigma}^2$ **intérieur.**

$$\hat{\sigma}^2 \text{ intérieur} = 2$$

$$F = \frac{\hat{\sigma}^2 \text{ entre}}{\hat{\sigma}^2 \text{ intérieur}} = \frac{3}{2} = 1{,}5$$

Quels sont les degrés de liberté? Pour $\hat{\sigma}^2$ entre, le degré de liberté (D_X) est le nombre d'échantillons moins 1 : ici $D_X = 3 - 1 = 2$. Pour $\hat{\sigma}^2$ intérieur le degré de liberté (D_Y) est le nombre total de résultats (dans tous les échantillons) moins le nombre d'échantillons: ici

$D_Y = 9 - 3 = 6$. Pour $\alpha = 0{,}05$, la valeur critique de $F(2; 6)$ est 5,14 pour un test unilatéral à droite. La valeur calculée de F est 1,5 et est dans la région d'acceptation de H_0 donc $\mu_1 = \mu_2 = \mu_3$. Les trois populations d'où sont pris les échantillons ont la même moyenne et de ce fait sont des populations identiques puisqu'au départ on avait admis qu'elles avaient la même variance et étaient normalement distribuées.



Reprenons le même problème, mais cette fois, en synthétisant les étapes.

Soit k = le nombre d'échantillons
n_T = le nombre total de résultats de tous les échantillons.

Des justifications des étapes suivantes sont données à la section 14.6.

1. Écrire en colonne les résultats de chaque échantillon.
2. Additionner les résultats du premier échantillon, mettre au carré cette somme puis diviser par le nombre de résultats dans l'échantillon.
3. Refaire l'étape 2 avec chaque échantillon.
4. Additionner les résultats obtenus en 2 et 3, appeler cette somme A.
5. Additionner tous les résultats de tous les échantillons, appeler cette somme B.
6. Élever au carré chaque résultat d'un échantillon puis additionner. Appeler cette somme C, répéter pour tous les échantillons.
7. Trouver le nombre total de résultats de tous les échantillons n_T.
8. Calculer $\hat{\sigma}^2$ entre $= \dfrac{A - B^2/n_T}{k - 1}$ et $\hat{\sigma}^2$ intérieur $= \dfrac{C - A}{n_T - k}$
9. $F = \dfrac{\hat{\sigma}^2 \text{ entre}}{\hat{\sigma}^2 \text{ intérieur}}$ est comparé avec la valeur critique de $F_\alpha(k - 1, n_T - k)$

Voici la disposition des calculs.

	groupe 1	groupe 2	groupe 3	
	3	5	3	
	4	6	5	
	5	7	7	Total
Somme:	12	18	15	$45 = B$
Somme au carré:	144	324	225	
Taille de l'échantillon:	3	3	3	$9 = n_T$

	groupe 1	groupe 2	groupe 3	Total
Somme au carré : Taille	48	108	75	$231 = A$
Somme des carrés:	$9 + 16 + 25 =$ 50	$25 + 36 + 49 =$ 110	$9 + 25 + 49 =$ 83	$243 = C$

	Numérateur	Degrés de liberté (dénominateur)	Quotient
$\hat{\sigma}^2$ entre	$A - B^2/n_T = 231 - 45^2/9 = 6$	$k - 1 = 2$	$6/2 = 3$
$\hat{\sigma}^2$ intérieur	$C - A = 243 - 231 = 12$	$n_T - k = 6$	$12/6 = 2$

$F = 3/2 = 1{,}5$; $F_{0,05}(2; 6) = 5{,}14 > 1{,}5$ donc H_0 est acceptée.

14.5 MÉTHODE COURTE: SOUSTRAIRE UNE CONSTANTE *A* DE CHAQUE DONNÉE D'UN ÉCHANTILLON

Quelquefois, les données sont passablement grandes; on aura avantage à procéder comme suit.

Exemple 1

Un économiste veut vérifier au niveau de signification de 0,01 l'hypothèse que les menuisiers, les électriciens et les maçons gagnent le même salaire hebdomadaire à Montréal. Il suppose que les salaires des trois groupes sont normalement distribués et ont la même variance. Il recueille les résultats suivants:

Menuisiers	Maçons	Électriciens
240	230	290
200	270	270
220	250	250
160		270
190		

$H_0 : \mu_1 = \mu_2 = \mu_3$; H_1 : il n'est pas vrai que $\mu_1 = \mu_2 = \mu_3$.

	groupe 1	groupe 2	groupe 3	total
ΣX	1 010	750	1 080	$2\,840 = B$
$(\Sigma X)^2$	1 020 100	562 500	1 166 400	
n	5	3	4	$12 = n_T$
D	4	2	3	9
$(\Sigma X)^2/n$	204 020	187 500	291 600	$683\,120 = A$
$\Sigma(X^2)$	207 700	188 300	292 400	$688\,400 = C$

	Numérateur	Degrés de liberté (dénominateur)	Quotient
$\hat{\sigma}^2$ entre	$683\,120 - 2\,840^2/12 = 10\,987$	$3 - 1 = 2$	5 493,5
$\hat{\sigma}^2$ intérieur	$688\,400 - 683\,120 = 5\,280$	9	587

$$F = \frac{5\,493{,}5}{587} = 9{,}4.\ F_{0{,}01}\,(2;\ 9) = 8{,}02.\ H_0 \text{ est rejetée.}$$

Si une constante est soustraite de chaque donnée d'un échantillon, la variance ne change pas et les calculs peuvent être simplifiés. Dans l'exemple 1, si on soustrait 250 de chaque donnée, on obtiendra le même résultat F.

Exemple 2

Reprendre l'exemple 1, en soustrayant 250 de chaque donnée.

groupe 1	groupe 2	groupe 3
−10	−20	40
−50	20	20
−30	0	0
−90		20
−60		

				Total
Somme:	− 240	0	80	$-160 = B$
Somme au carré:	57 600	0	6 400	
Taille:	5	3	4	$12 = n_T$
Somme au carré / Taille	11 520	0	1 600	$13\,120 = A$
Somme des carrés:	15 200	800	2 400	$18\,400 = C$

	Numérateur	Degrés de liberté (dénominateur)	Quotient
$\hat{\sigma}^2$ entre	$13\,120 - (-160)^2/12 = 10\,987$	2	5 493,5
$\hat{\sigma}^2$ intérieur	$18\,400 - 13\,120 = 5\,280$	9	587

$F = \frac{5\,493{,}5}{587} = 9{,}4$ qui est la même valeur obtenue que précédemment. H_0 est rejetée: il y a une différence significative entre les salaires moyens des trois populations. Ce problème sera discuté à la section 14.7.

14.6★ D'OÙ VIENNENT LES FORMULES?

Examinons de plus près l'exemple des étudiants de la section précédent. Les indices 1; 2 et 3 identifient les trois groupes. À cette notation, ajoutons que:

$$A = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}$$

$$B = \Sigma X_1 + \Sigma X_2 + \Sigma X_3$$

$$C = \Sigma(X_1^2) + \Sigma(X_2^2) + \Sigma(X_3^2)$$

On a trouvé précédemment que $\hat{\sigma}^2$ **entre** est la variance des moyennes des échantillons par rapport à la moyenne $\overline{X}_T$ de tous les résultats; ainsi:

$$\hat{\sigma}^2 \text{ entre} = \frac{n_1(\overline{X}_1 - \overline{X}_T)^2 + n_2(\overline{X}_2 - \overline{X}_T)^2 + n_3(\overline{X}_3 - \overline{X}_T)^2}{k-1}$$

(dans cet exemple $K = 3$). Examinons le développement du 1er terme du numérateur.

$$n_1(\overline{X}_1 - \overline{X}_T)^2 = n_1\overline{X}^2 - 2n_1\overline{X}_1\overline{X}_T + n_1\overline{X}_T^2 \qquad \text{en élevant au carré } (\overline{X}_1 - \overline{Y}_T)^2$$

$$= n_1\left(\frac{\Sigma X_1}{n_1}\right)^2 - 2n_1\left(\frac{\Sigma X_1}{n_1}\right)\overline{X}_T + n_1\overline{X}_T^2 \qquad \text{puisque } \overline{X}_1 = \frac{\Sigma X_1}{n_1}$$

$$= \frac{(\Sigma X_1)^2}{n_1} - 2(\Sigma X_1)\overline{X}_T + n_1\overline{X}_T^2$$

Les deux autres termes $n_2(\overline{X}_2 - \overline{X}_T)^2$ et $n_3(\overline{X}_3 - \overline{X}_T)^2$ peuvent se développer de la même manière. En additionnant ces développements au numérateur, on obtient:

$$\left[\frac{(\Sigma X_1)^2}{n_1} - 2(\Sigma X_1)\overline{X}_T + n_1\overline{X}_T^2\right]$$

$$+\left[\frac{(\Sigma X_2)^2}{n_2} - 2(\Sigma X_2)\overline{X}_T + n_2\overline{X}_T^2\right]$$

$$+\left[\frac{(\Sigma X_3)^2}{n_3} - 2(\Sigma X_3)\overline{X}_T + n_3\overline{X}_T^2\right]$$

$$= \underbrace{\left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right]}_{=A} - \underbrace{2(\Sigma X_1 + \Sigma X_2 + \Sigma X_3)\overline{X}_T + (n_1 + n_2 + n_3)\overline{X}_T^2}_{=B^2/n_T?}$$

Montrons que le dernier terme de la somme $= -B^2/n_T$, on sait que $n_1 + n_2 + n_3 = n_T$ et que $\overline{X}_T = \dfrac{\Sigma X_1 + \Sigma X_2 + \Sigma X_3}{n_1 + n_2 + n_3} = \dfrac{(\text{somme de tous les } x)}{n_T} = \dfrac{B}{n_T}$.

$$-2(\Sigma X_1 + \Sigma X_2 + \Sigma X_3)\overline{X}_T + (n_1 + n_2 + n_3)\overline{X}_T^2 = -2B\left(\frac{B}{n_T}\right) + (n_T)\left(\frac{B}{n_T}\right)^2$$

$$= \frac{-2B^2}{n_T} + \frac{B^2}{n_T} = \frac{-B^2}{n_T}$$

Finalement $\hat{\sigma}^2 \text{ entre} = \dfrac{A - B^2/n_T}{k-1}$.

Examinons maintenant $\hat{\sigma}^2$ intérieur. C'est la variance pondérée moyenne où les poids sont les degrés de liberté de chaque échantillon.

$$\hat{\sigma}^2 \text{ intérieur} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}$$

Développons le 1er terme du numérateur.

$$(n_1 - 1)s_1^2 = (n_1 - 1)\left[\frac{\Sigma X_1^2 - (\Sigma X_1)^2/n_1}{n_1 - 1}\right] = \Sigma X_1^2 - (\Sigma X_1)^2/n_1$$

Les autres termes peuvent se développer de la même façon.

$$\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1} + \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2} + \Sigma X_3^2 - \frac{(\Sigma X_3)^2}{n_3}.$$

en redisposant les termes:

$$(\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2) - \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right] = C - A.$$

Au dénominateur on a: $(n_1 - 1) + (n_2 - 1) + (n_3 - 1) = n_T - k$,

Finalement $\hat{\sigma}^2$ intérieur $= \dfrac{C - A}{n_T - k}$.

14.7 QUELLES MOYENNES SONT DIFFÉRENTES?

Supposons que H_0: $\mu_1 = \mu_2 \ldots\ldots = \mu_K$ a été rejetée après avoir fait un test F. Il serait intéressant de connaître quelles paires de moyennes sont significativement différentes. Un certain nombre de statisticiens ont trouvé un test qui peut s'appliquer après qu'un test d'analyse de variances a montré une différence significative entre les moyennes. Le test décrit ci-dessous a été développé par Scheffé.† Il peut être utilisé pour des échantillons de tailles différentes et même si les populations ne sont pas normales et n'ont pas les mêmes variances. Ce test est valide pour **toutes paires d'échantillons** et diffère considérablement d'un test t répété.

Test de Scheffé

Le test de Scheffé comprend les étapes suivantes:

1. Calculer $S = \sqrt{(k - 1)\, F\, (k - 1, n_T - k)\, (\hat{\sigma}^2 \text{ intérieur})}$.
2. Si les échantillons sont de même taille n vérifier si $|\overline{X}_i - \overline{X}_j|$ est inférieur ou supérieur à $S\sqrt{2/n}$.
   Si les échantillons sont de tailles différentes vérifier si $|\overline{X}_i - \overline{X}_j|$ est inférieur ou supérieur à $S\sqrt{\dfrac{1}{n_i} + \dfrac{1}{n_j}}$

†Scheffé, H., *The Analysis of Variance*, John Wiley, New York, 1959.

Ce test est fait pour chaque paire de moyennes d'échantillons $\overline{X}_i$, $\overline{X}_j$.

3. Si $|\overline{X}_i - \overline{X}_j|$ est supérieur, $H_0: \mu_i = \mu_j$ est rejetée, on conclut alors que cette paire de moyennes a contribué au rejet de l'hypothèse dans le test de l'analyse de variances. Si $|\overline{X}_i - \overline{X}_j|$ est inférieur, alors on ne peut rejeter l'hypothèse que $\mu_i = \mu_j$.

Note: α dans le test de Scheffé est souvent le même que celui qui a servi dans le test de la variance, mais il peut ne pas l'être.

Exemple 1

Un botaniste veut vérifier 5 variétés de maïs sur 5 lopins de terre de mêmes dimensions et de même fertilité. La récolte moyenne de maïs fut: 3,4; 7,4; 10; 10; 11,8 boisseaux par lot. Au niveau de signification de 0,05, $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ est rejetée avec $A = 2\,027{,}8$; $C = 2\,217$. Quelles paires de moyennes diffèrent au niveau $\alpha = 0{,}05$? $n_t = 25$, $k = 5$, $\hat{\sigma}^2$ intérieur $= \dfrac{2\,217 - 2\,027{,}8}{25 - 5} = 9{,}46$, $F_{0,05}\,(4;20) = 2{,}87$; $S\sqrt{2/n} = \sqrt{4 * 2{,}87 * 9{,}46} * \sqrt{2/5} = \sqrt{108{,}6} * \sqrt{0{,}4} = 6{,}6$.

La seule paire de moyennes qui diffère par plus de 6,6 est la paire 1;5 donc $\mu_1 = \mu_5$ doit être rejetée. Nous sommes incapables, toutefois, de rejeter l'énoncé qui affirme que les autres paires de moyennes sont égales.

Exemple 2

Considérons les menuisiers, les maçons et les électriciens de l'exemple 1 de la section 14.5. Le test de l'analyse de variance a montré que $H_0: \mu_1 = \mu_2 = \mu_3$ est rejetée. Si $\alpha = 0{,}01$, appliquer le test de Scheffé pour savoir quelles paires de moyennes diffèrent.

$k - 1 = 3 - 1 = 2$; $F_{0,01}\,(2;9) = 8{,}02$; $\hat{\sigma}^2$ intérieur $= 587$; $S = \sqrt{2 * 8{,}02 * 587} = 97{,}0$.

On doit examiner chaque paire de moyennes.
Utilisons le tableau suivant:

		Différences de moyennes			$\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$		$S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$	
		$\overline{X}_2 = 250$	$\overline{X}_3 = 270$	n_i	n_2	n_3		
Menuisiers	$\overline{X}_1 = 202$	−48	−68	5	0,730	0,671	70,8	65,1
Maçons	$\overline{X}_2 = 250$		−20	3		0,764		74,1
Électriciens	$\overline{X}_3 = 270$			4				

L'énoncé «$\mu_1 = \mu_3$» est rejeté car $|-68| > 65{,}1$. On ne peut rejeter les énoncés «$\mu_1 = \mu_2$» car $|-48| < 70{,}8$ et «$\mu_2 = \mu_3$» car $|-20| < 74{,}1$.

14.8 VOCABULAIRE ET SYMBOLES

Résultat F
$F(D_X ; D_Y)$
$F \alpha (D_X ; D_Y)$
Distribution F
Analyse de variance
$\hat{\sigma}^2$ entre
$\hat{\sigma}^2$ intérieur
k
n_t
A, B, C
Test de Scheffé S

14.9 EXERCICES

1. Trouver la probabilité que F (8; 10) soit inférieure à 5,06.

2. Trouver la probabilité que F (7; 12) soit inférieure à 0,280.

3. Dix câbles de nylon sont choisis au hasard; leur force de rupture moyenne est 800 kg avec un écart-type de 50 kg; 12 câbles d'orlon ont une force de rupture moyenne de 900 kg avec un écart-type de 28 kg. Au niveau de signification de 0,05, vérifier les hypothèses suivantes.

(a) Il n'y a pas de différence dans la variance des forces de rupture des deux types de câbles.

(b) Le câble de nylon a plus de variabilité dans sa force de rupture que le câble d'orlon.

4. A un examen final, 30 étudiants qui ont suivi le cours de statistiques à la T.V. ont obtenu un résultat moyen de 78 avec une variance de 25 alors que 21 étudiants qui ont eu un cours magistral ont obtenu une moyenne de 80 avec une variance de 30. Au niveau de signification de 0,01, est-ce que les étudiants qui ont suivi le cours à la T.V. ont moins de variabilité?

5. Vingt-cinq bouteilles de Cola-Coca choisies au hasard ont une moyenne de 12,0 calories avec un écart-type de 0,60. 41 bouteilles de Popi-Cola ont une moyenne de 20,0 calories avec un écart-type de 0,80. Au niveau de signification de 0,02, est-ce que la variance des calories du Cola-Coca diffère de celle du Popi-Cola?

6. Se référant à l'exercice 14 de la section 12.11, est-ce que les deux sortes de détergents ont la même variance au niveau de signification de 0,05?

7. Deux échantillons de taille 8 et 5 respectivement sont choisis parmi les enfants de 1re année à l'école Christ-Roy puis on leur fait passer un test de lecture. Le premier échantillon formé d'enfants qui ont fait une année de pré-maternelle a obtenu une moyenne de 104 avec un écart-type de 7. Le deuxième échantillon formé d'enfants qui n'ont pas fait de pré-maternelle a obtenu une moyenne de 95 avec un écart-type de 4. En supposant que les résultats du test de lecture sont normalement distribués dans les deux groupes.

(a) Est-ce que les résultats des enfants qui ont fait une année de pré-maternelle ont la même variance que ceux qui ne l'ont pas fait ($\alpha = 0{,}02$)?

(b) Y a-t-il une différence entre les résultats des deux groupes d'élèves ($\alpha = 0{,}05$)?

8. Au niveau de signification de 0,05, est-ce que les populations d'où sont prélevés les échantillons suivants ont la même moyenne? Supposer que les populations sont normalement distribuées et ont la même variance.

Échantillon 1	Échantillon 2	Échantillon 3	Échantillon 4
0	−1	1	3
1	2	4	5
2	5	7	8
3			10

Pour les problèmes 9 à 12 utiliser le test de l'analyse de variances pour vérifier si les échantillons proviennent de populations ayant la même moyenne ($\alpha = 0{,}05$). Écrire H_0 et H_1 puis décider si H_0 est acceptée ou rejetée.

9.

Échantillon 1	Échantillon 2	Échantillon 3	Échantillon 4
0	4	4	8
−2	0	6	6
2	2	2	4

10.

Échantillon 1	Échantillon 2	Échantillon 3
2	3	−1
4	5	0
5	6	1
		3

11.

Échantillon 1	Échantillon 2
10	8
12	10
14	12
16	14
18	16
20	

(Suggestion: Soustraire 12 à chaque résultat).

12.

Échantillon 1	Échantillon 2
10	6
12	8
14	10
16	12
18	14
20	

13. Un botaniste vérifie 5 variétés de maïs dans 5 lopins de terre de mêmes dimensions et de même fertilité. Voici les récoltes obtenues pour chaque variété dans chaque lopin:

Variété 1	Variété 2	Variété 3	Variété 4	Variété 5
4	7	10	16	10
3	8	14	14	13
6	9	12	10	12
2	8	9	7	10
2	5	5	3	14

Au niveau de signification de 0,05, est-ce que les variétés produisent la même récolte?

14. On enseigne la statistique avec trois méthodes différentes (a) la T.V. (b) bobines audio (c) cours magistral. Un échantillon aléatoire de 5 étudiants pour chaque méthode a donné les résultats suivants à l'examen final:

(a)	(b)	(c)
75	80	90
78	75	95
80	70	60
60	75	80
95	80	80

Y a-t-il une différence significative dans les résultats obtenus entre chacune des méthodes ($\alpha = 0{,}05$). (Conseil: Soustraire 80 de chaque résultat.)

15. Les résultats de la table suivante montrent la performance relative au cours d'allemand par des échantillons aléatoires d'étudiants qui ont suivi le cours dans un laboratoire de langues et ceux qui ont utilisé un procédé informatisé.

Laboratoire de langues	Informatique
1	2
2	3
3	4
3	4
3	4
4	5
5	6
6	6
27	6
	6
	46

(a) Utiliser un test t pour vérifier $H_0 : \mu_X = \mu_Y$ au niveau de signification de 0,05, trouver en premier si les populations ont la même variance.

(b) Utiliser l'analyse de variances pour vérifier $H_0 : \mu_1 = \mu_2$ au niveau de signification de 0,05.

16. La force de tension de trois marques de fil de nylon est comparée en mesurant trois bobines de chaque marque choisies au hasard. Supposer que les forces sont normalement distribuées et que la variance est la même pour chaque marque.

Marque A	Marque B	Marque C
98	105	110
100	103	98
102	101	106

17. Un agronome mesure l'efficacité de trois insecticides différents. Il vérifie des échantillons aléatoires de chaque formule sur les colonies de fourmis et recueille les résultats suivants:

Pourcentages d'extermination
Formules

#1	#2	#3
40,2	50,4	71,2
35,1	60,2	60,3
50,3	53,8	75,4
	40,2	80,3

Au niveau de signification de 0,05, y a-t-il une différence entre le rendement moyen des trois insecticides? Poser que les distributions sont normales et ont la même variance.

18. Le même agronome reprend la même expérience, mais cette fois avec trois autres insecticides. Voici les résultats obtenus:

Pourcentages d'extermination
Formules

#1	#2	#3
40,2	50,4	71,2
39,6	51,0	72,0
43,0	52,0	73,0
	50,6	70,0

Même question qu'au numéro 17.

19. La table suivante donne les résultats de 5 étudiants à 4 examens de statistiques.

		Étudiants				
		1	2	3	4	5
Examens	1	85	70	90	95	70
	2	90	80	95	90	90
	3	70	75	80	85	70
	4	80	65	75	80	75

Vérifier au niveau de signification de 0,05 les hypothèses que (a) les étudiants ont la même moyenne et (b) les examens sont de même difficulté. (Conseil: Soustraire 80 de chaque résultat.)

20. La production de quatre travailleurs sur le même type de machinerie a été mesurée sur 4 périodes de temps choisies au hasard. Au niveau de signification de 0,05, les travailleurs ont-ils la même productivité?

		Travailleurs			
		A	B	C	D
Périodes	1	22	26	20	28
	2	20	28	16	26
	3	21	24	18	30
	4	24	22	20	26

(Conseil: Soustraire 25 de chaque résultat.)

21. Se référer au problème 9. Avec $\alpha = 0{,}05$, $H_0: \mu_1 = \mu_2 = \mu_3$ est rejetée. Utiliser un test de Scheffé avec $\alpha = 0{,}05$ pour trouver quelles paires de moyennes sont différentes de façon significative ($\hat{\sigma}^2$ intérieur $= 4$).

22. Se référer au problème 10. Utiliser un test de Scheffé avec $\alpha = 0{,}01$ pour trouver quelles paires de moyennes sont différentes de façon significative ($\hat{\sigma}^2$ intérieur $= 2{,}59$).

23. Se référer au problème 12. Utiliser un test de Scheffé avec $\alpha = 0{,}05$ pour trouver les paires de moyennes qui sont différentes de façon significative ($\hat{\sigma}^2$ intérieur $= 12{,}22$).

24. Au problème 20, $H_0: \mu_1 = \mu_2 = \mu_3$ est rejetée quand $\alpha = 0{,}05$. Trouver quelles paires de moyennes sont différentes (a) au niveau de signification de 0,05 (b) au niveau de 0,01.

25. Un test d'analyse de variances est fait sur les données suivantes:

X_1	X_2	X_3	X_4
0	2	4	6
1	3	5	7
1	3	5	7
2	4	6	8
2	4	6	8

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ est rejetée au niveau de signification de 0,05. Utiliser un test de Scheffé pour trouver quelles paires de moyennes sont différentes au niveau de signification de 0,05 ($A = 452{,}8$; $C = 464$).

26. Voir problème 13. Les 5 variétés de maïs ne donnent pas la même récolte en utilisant un test d'analyse de variances au niveau de signification de 0,05. Faire un test avec $\alpha = 0{,}05$ pour déterminer quelles paires de moyennes sont différentes ($\hat{\sigma}^2$ intérieur $= 9{,}46$).

27. La quantité de cola distribuée par 3 distributrices automatiques est mesurée comme suit:

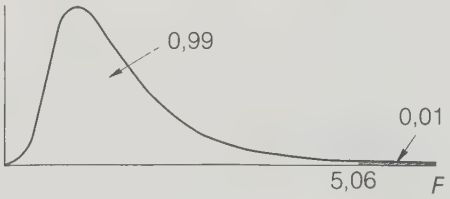
Machine A: 208; 214; 212; 213 ml
Machine B: 198; 200; 199; 201; 200 ml
Machine C: 216; 212; 213 ml

Au niveau de signification de 0,01 est-ce que les distributrices donnent la même quantité de cola?

28. Se référer au problème 27. Utiliser un test de Scheffé avec $\alpha = 0{,}05$ pour trouver quelles paires de moyennes sont différentes ($\hat{\sigma}^2$ intérieur = 3,85).

RÉPONSES

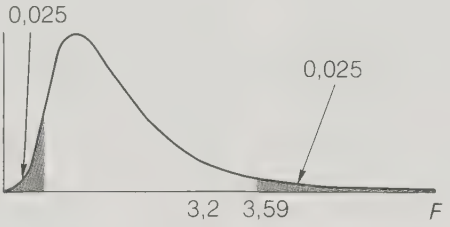
1. $F(8; 10) = 5{,}06$ quand $\alpha = 0{,}01$. Donc la probabilité que $F(8; 10)$ soit inférieure à 5,06 est 0,99.



3. $H_0: \sigma_X^2 = \sigma_Y^2$; $H_1: \sigma_X^2 \neq \sigma_Y^2$ (test bilatéral).

$n_X = 10$, $s_X^2 = 50^2$, $n_Y = 12$, $s_Y^2 = 28^2$.

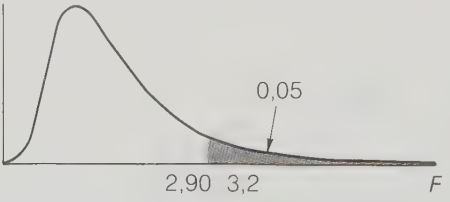
$F = s_X^2/s_Y^2 = 3{,}2$. $D_X = 9$, $D_Y = 11$; $F(9; 11) = 3{,}59$.



H_0 est acceptée: il n'y a pas de différence entre les variances.

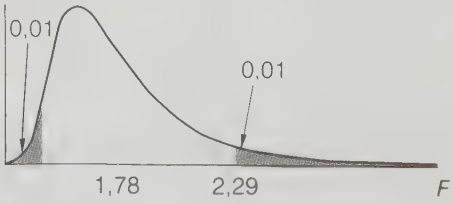
(b) $H_0: \sigma_X^2 = \sigma_Y^2$; $H_1: \sigma_X^2 > \sigma_Y^2$ (test unilatéral à droite).

La valeur calculée de F est trouvée en (a). $F(9;11) = 2{,}90$ quand $\alpha = 0{,}05$.



H_0 est rejetée.

5. $H_0: \sigma_X^2 = \sigma_Y^2$; $H_1: \sigma_X^2 \neq \sigma_Y^2$; test bilatéral. $n_X = 41$, $s_X = 0{,}80$; $n_Y = 25$, $s_Y = 0{,}60$; $F = 0{,}80^2/0{,}60^2 = 1{,}78$. $\alpha = 0{,}02$; $F(40; 24) = 2{,}49$.



H_0 est acceptée: Cola-Coca et Popi-Cola ont la même variance.

7. (a) $s_X = 7$, $n_X = 8$, $s_Y = 4$, $n_Y = 5$, $F = \frac{7^2}{4^2} = 3{,}06$; si $\alpha = 0{,}02$, test bilatéral, $F\,(7; 4) = 15{,}00$. Les variances sont identiques.

(b) $\sigma_{(\bar{X}-\bar{Y})} = \sqrt{\frac{7 * 7^2 + 4 * 4^2}{8 + 5 - 2}\left(\frac{1}{8} + \frac{1}{5}\right)} = 3{,}47$; $t = \frac{(104 - 95) - 0}{3{,}47} = 2{,}59$; $\alpha = 0{,}05$; $D = 11$. Les valeurs critiques de t sont $\pm 2{,}20$. H_0 est rejetée; il y a une différence entre les enfants qui ont fait une pré-maternelle et ceux qui ne l'ont pas fait.

9. $F = \frac{(168 \quad 36^2/12)/3}{(200 - 168)/8} = \frac{20}{4} = 5{,}00$; $F_{0,05}\,(3; 8) = 4{,}07$. H_0 est rejetée; les moyennes ne sont pas les mêmes.

11. $F = \frac{(54 - 18^2/11)/1}{(164 - 54)/9} = \frac{24{,}55}{12{,}22} = 2{,}01$; $F_{0,05}\,(1; 9) = 5{,}12$. H_0 est acceptée.

13. Soustraire 10 de chaque résultat. $F = \frac{213{,}0/4}{189{,}2/20} = \frac{53{,}3}{9{,}46} = 5{,}6$; $F_{0,05}\,(4; 20) = 2{,}87$. H_0 est rejetée; les récoltes sont différentes.

15. (a) H_0: $\mu_X - \mu_Y = 0$; H_1: $\mu_X - \mu_Y \neq 0$; $\alpha = 0{,}05$.
$\bar{X} = 27/8 = 3{,}38$; $s_X^2 = (109 - 27^2/8)/7 = 2{,}55$.
$\bar{Y} = 46/10 = 4{,}60$; $s_Y^2 = (230 - 46^2/10)/9 = 2{,}04$.

$s_X^2/s_Y^2 = 2{,}55/2{,}04 = 1{,}25$; avec $\alpha = 0{,}05$, test bilatéral, $F\,(7; 9) = 4{,}20$.
Les populations ont la même variance.

$$t = \frac{(3{,}38 - 4{,}6) - 0}{\sqrt{\frac{7 * 2{,}55 + 9 * 2{,}04}{(8 - 1) + (10 - 1)} * \left(\frac{1}{8} + \frac{1}{10}\right)}} = \frac{-1{,}22}{0{,}71} = -1{,}72$$

$D = (8 - 1) + (10 - 1) = 16$; $\alpha = 0{,}05$; la valeur critique de $t = 2{,}12$. H_0 est acceptée; il n'y a pas de différence significative entre les deux méthodes d'enseignement.

(b) $F = \frac{(302{,}7 - 73^2/18)/1}{(339 - 302{,}7)/16} = \frac{6{,}64}{2{,}27} = 2{,}93$; $F_{0,05}(1;16) = 4{,}49$. H_0 est acceptée.

Il est à noter que la valeur calculée et la valeur critique de F sont, en arrondissant, les carrés des valeurs calculées et critiques de t. Pour H_0: $\mu_1 = \mu_2$ et H_1: $\mu_1 \neq \mu_2$, le test t et le test d'analyse de variance sont équivalents.

17. (Soustraire 50 de chaque résultat.) $F = \frac{(2\,104{,}70 - 67{,}4^2/11)/2}{(2\,651{,}6 - 2\,104{,}7)/8} = \frac{845{,}9}{68{,}4} = 12{,}4$.
$F_{0,05}\,(2; 8) = 4{,}46$. H_0 est rejetée; les trois insecticides n'ont pas le même rendement.

19. (a) $F = \frac{(612{,}50 - 10^2/20)/4}{(1600 - 612{,}50)/15} = 2{,}31$; $F_{0,05}(4; 15) = 3{,}06$. Les moyennes des étudiants sont identiques.

(b) $F = \dfrac{(630{,}25 - 10^2/20)/3}{(1600 - 630)/16} = 3{,}44$ $F_{0{,}05}(3; 16) = 3{,}24$. Les moyennes des examens sont identiques.

21. $\overline{X}_1 = 0; \overline{X}_2 = 2, \overline{X}_3 = 4; \overline{X}_4 = 6$; $S\sqrt{2/n} = \sqrt{3 \cdot 4{,}07 \cdot 4}\sqrt{2/3} = 5{,}71$; $\mu_1 \neq \mu_4$, puisque $|6 - 0| > 5{,}71$. On ne peut conclure qu'il y a d'autres paires qui sont différentes.

23. $F_{0{,}05}(1; 9) = 5{,}12$; $S\sqrt{(1/n_1) + (1/n_2)} = \sqrt{1 \cdot 5{,}12 \cdot 12{,}22}\sqrt{(1/6) + (1/5)} = 4{,}79$. $\overline{X}_1 = 5$, $\overline{X}_2 = 0$, $\mu_1 \neq \mu_2$ puisque $|5 - 0| > 4{,}79$.

25. $\hat{\sigma}^2$ intérieur $= (464 - 452{,}8)/16 = 0{,}70$; $F_{0{,}05}(3; 16) = 3{,}24$; $S\sqrt{2/n} = \sqrt{5 \cdot 3{,}24 \cdot 0{,}70}\sqrt{2/5} = 2{,}13$. $\overline{X}_1 = 1{,}20$; $\overline{X}_2 = 3{,}20$; $\overline{X}_3 = 5{,}20$; $\overline{X}_4 = 7{,}2$. $\mu_1 \neq \mu_3$, $\mu_1 \neq \mu_4$, $\mu_2 \neq \mu_4$; les autres paires de moyennes ne sont pas différentes.

27. (Soustraire 200 de chaque résultat.) $F = \dfrac{(1113{,}38 - 86^2/12)/2}{(1148 - 1113{,}38)/9} = \dfrac{248{,}52}{3{,}85} = 64{,}55$.

$F_{0{,}01}(2; 9) = 8{,}02$. Les distributrices ne donnent pas des consommations équivalentes.

Chapitre 15

Corrélation linéaire et régression

Plusieurs problèmes de statistiques consistent à établir la relation qui existe entre deux variables: poids et grandeur des étudiants, l'âge et le salaire des plombiers, le Q.I. et le résultat d'un test d'aptitudes manuelles, pour ne citer que quelques exemples; on doit être en mesure de répondre aux deux questions suivantes: existe-t-il une relation entre les variables? Si oui, la connaissance des valeurs d'une variable permet-elle de prédire les valeurs de l'autre? (Y a-t-il une relation entre les résultats obtenus en secondaire V et ceux obtenus au collégial? Si oui, est-ce que le régistraire du collège peut prévoir la réussite d'un étudiant en examinant ses résultats obtenus en secondaire V?)

Dans ce chapitre, nous verrons comment tracer des paires de données et trouver l'équation de la droite qui représente le mieux les points du graphique, puis trouver un coefficient qui permet de dire si la droite représente bien l'ensemble des points du graphique. Il ne sera pas nécessaire que la population d'où proviennent les échantillons possède des caractéristiques spéciales. Si on doit prédire les valeurs de la seconde variable à partir des valeurs de la première ou si on doit mesurer la proximité des valeurs des paires de données d'une population par rapport à une droite, nous devrons poser certaines hypothèses concernant la population. Ces hypothèses et inférences seront étudiées au chapitre 16.

15.1 INTRODUCTION

Supposons que l'on connaît, pour un certain nombre de personnes, la scolarité et le revenu annuel. Dans chaque cas, les observations se présentent sous forme de paires: pour la première personne, sa scolarité (X) et son revenu (Y) sont connus, la paire (X_1; Y_1) représente la première personne; la seconde personne est représentée par la paire d'observations (X_2; Y_2) et ainsi de suite. Nous ne sommes pas intéressés à avoir une liste des scolarités et une autre pour les revenus, mais nous voulons des observations qui se présentent en paire. La paire (X; Y) est appelée une «paire ordonnée» puisque la paire

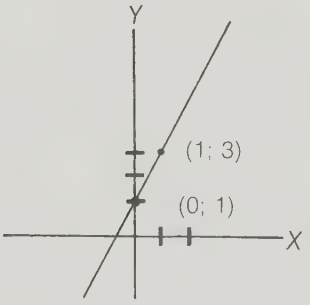
(12; 10 000) veut dire 12 ans de scolarité et un revenu de $10 000. Il est évident qu'on ne peut inverser l'ordre de cette paire. Attention à la notation: nous sommes familiers avec le système de coordonnées x, y. Ici nous utiliserons X et Y puisque ces lettres représentent les résultats d'un échantillon. De même l'équation de la droite qui s'écrit $y = mx + b$ où m est la pente et b l'ordonnée à l'origine est remplacée par $Y = a + bX$ où cette fois, b est la pente et «a» l'ordonnée à l'origine.

Une certaine connaissance de la fonction linéaire est nécessaire. Si vous pouvez répondre aux questions suivantes alors passez directement à la section 15.5; sinon étudiez attentivement les sections suivantes.

1. Pouvez-vous tracer la droite $Y = 1 + 2X$? (Si non, voir section 15.2.)
2. Quelle est la pente de la droite $Y = 4 + 2X$? Quelle est son ordonnée à l'origine?
3. Quelle est l'équation de la droite passant par P (1; 4) et Q (2; 3) (section 15.4)?

RÉPONSES

1.

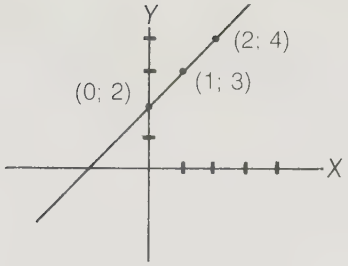


2. Pente = 2, ordonnée à l'origine = 4.
3. $Y = 5 - X$.

15.2 GRAPHIQUE DE LA DROITE

Une équation de la forme $Y = a + bX$ est représentée par une ligne droite (il est aussi vrai que toute droite qui n'est pas verticale a pour équation générale $Y = a + bX$). En remplaçant X par 2 et Y par 4, l'équation $Y = 2 + X$ devient une identité $4 = 2 + 2$; on dit alors que le point P dont les coordonnées sont (2; 4) est sur la droite $Y = 2 + X$. Le point Q (0; 2) est sur la droite, car $2 = 2 + 0$. Puisque deux points suffisent à tracer une droite, on peut alors tracer le graphique de $Y = 2 + X$. On peut vérifier par un troisième point, par exemple: R (1; 3) dont les coordonnées satisfont l'équation et voir si ce point tombe sur la droite déterminée par P et Q. Il est souvent pratique d'écrire un tableau des valeurs comme ci-dessous. On prend trois valeurs pratiques de X et on trouve les valeurs correspondantes de Y en remplaçant X par sa valeur dans l'équation puis en calculant Y.

X	Y
0	2
1	3
2	4



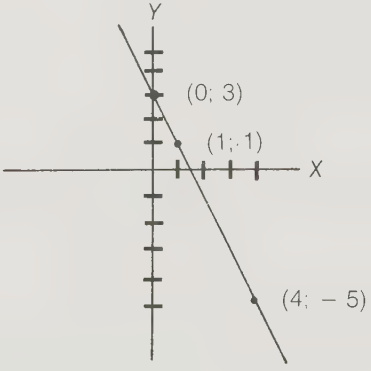
Exemple 1

Tracer la droite dont l'équation est $Y = 3 - 2X$. Donner à X les valeurs: 0; 1; 4.

X	Y
0	
1	
4	

Puis calculer les valeurs correspondantes de Y. Si $X = 0$ alors $Y = 3 - 2 * 0 = 3$.
De même, si $X = 1$, alors $Y = 1$; $X = 4$; $Y = -5$.

X	Y
0	3
1	1
4	−5



15.3 LA PENTE D'UNE DROITE

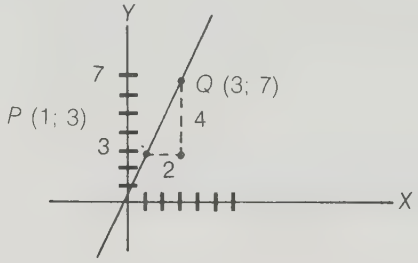
La pente d'une droite est un nombre qui nous indique si les valeurs de Y augmentent ou diminuent quand X augmente. La pente indique également la rapidité de cette variation. La pente d'une droite verticale n'est pas définie; la pente d'une droite horizontale est 0. Pour toute autre droite, on prend deux points de cette droite puis on trouve le rapport de la variation de Y à celui des X:

$$\text{Pente} = \frac{\text{Variation de } Y}{\text{Variation de } X}$$

Exemple 1

Quelle est la pente de la droite $Y = 1 + 2X$?
Choisir deux points de la droite; par exemple P (1; 3) et Q (3; 7).

La pente est $\frac{7-3}{3-1} = \frac{4}{2} = 2$.

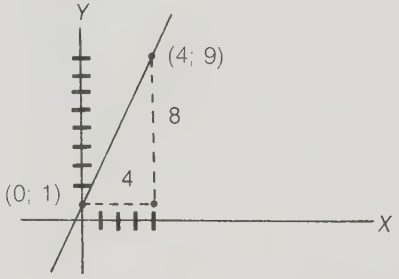


L'ordre des points n'importe pas:

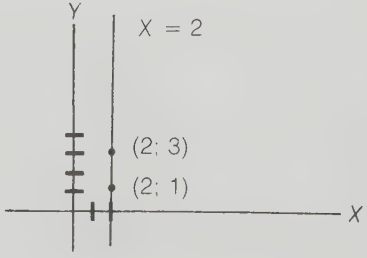
$$\text{Pente} = \frac{3-7}{1-3} = \frac{-4}{-2} = 2.$$

Si on prend l'ordonnée de Q en premier, il faut prendre l'abscisse de Q en premier.

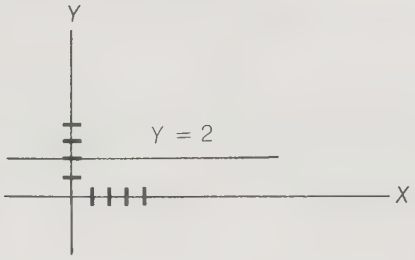
On peut choisir d'autres points de la droite; ainsi avec R (0; 1) et S (4; 9), la pente est $\frac{9-1}{4-0} = \frac{8}{4} = 2$.



Pourquoi la pente d'une droite verticale n'est-elle pas définie? Prenons deux points sur la droite verticale $X = 2$, par exemple les points (2; 1) et (2; 3), puis calculons la pente $\frac{3-1}{2-2} = \frac{2}{0} = ?$ La division par zéro n'est pas définie, la pente n'est pas définie.



Choisir deux points sur la droite horizontale $Y = 2$, puis vérifier que la pente est 0.

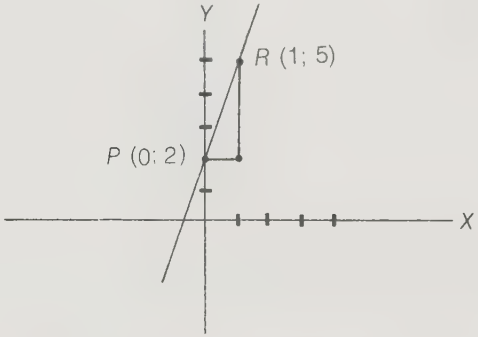


Exemple 2

Quelle est la pente de la droite dont l'équation est $Y = 2 + 3X$?
(0; 2) et (1; 5) sont deux points de la droite.

$$\text{La pente} = \frac{5-2}{1-0} = 3 \left(\text{ou } \frac{2-5}{0-1} = \frac{-3}{-1} = 3\right)$$

X	Y
0	2
1	5



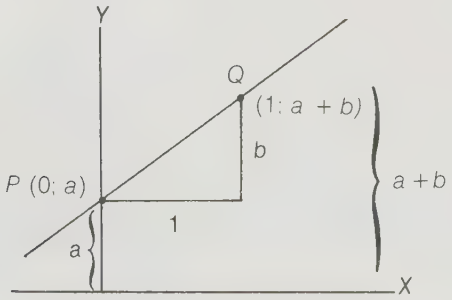
On constate que **3 est le coefficient de X dans l'équation $Y = 2 + 3X$.** Est-ce que ce sera toujours le cas?

Exemple 3

Quelle est la pente de la droite $Y = a + bX$ où a et b sont des constantes?
Supposons pour le graphique que a et b sont positifs. Si $X = 0$, $Y = a$; si $X = 1$, $Y = a + b$. Donc P (0; a) et Q (1; $a + b$) sont deux points de la droite.

$$\text{La pente} = \frac{(a+b)-a}{1-0} = \frac{b}{1} = b$$

X	Y
0	a
1	$a + b$



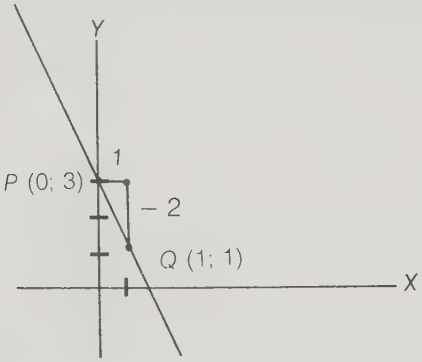
La pente de la droite $Y = a + bX$ est b.

Dans les exemples que nous venons de voir, le point de la droite s'élève si X augmente. Que se passe-t-il si le point **descend** quand X augmente?

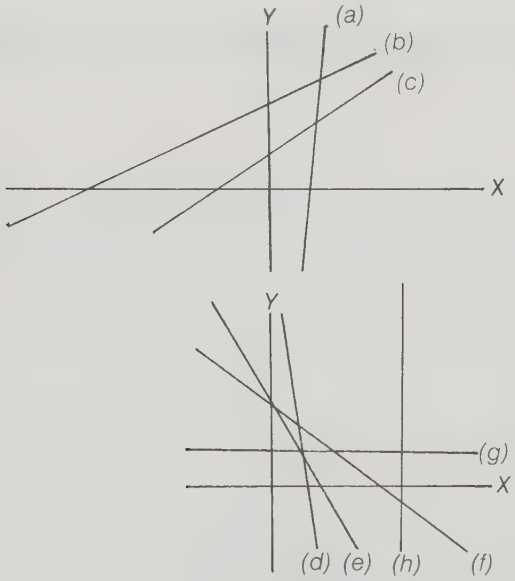
Exemple 4

Quelle est la pente de la droite $Y = 3 - 2X$? P (0; 3) et Q (1; 1) sont sur la droite.

$$\text{La pente} = \frac{3-1}{0-1} = \frac{2}{-1} = -2 \quad \left(\textbf{ou} = \frac{1-3}{1-0} = \frac{-2}{1} = -2\right)$$



Les droites (a) (b) (c) ci-dessous ont des pentes positives; les droites (d) à (f) ont des pentes négatives; la pente de (g) est 0; la pente de (h) n'est pas définie.



On peut trouver l'ordonnée à l'origine de la droite en posant $X = 0$ dans l'équation.

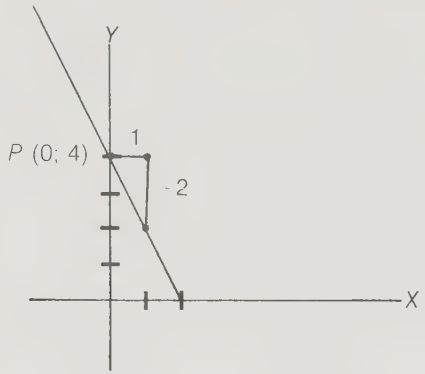
Équation	Pente	Ordonnée à l'origine
$Y = 1 + 2X$	2	1
$Y = 3 - 2X$	-2	3
$Y = -4 + X$	1	-4
$Y = a + bX$	b	a

Si la pente et l'ordonnée à l'origine sont connues, il est facile de tracer la droite. Il est à noter que dans l'équation $Y = a + bX$, « a » est l'ordonnée à l'origine et « b » la pente de la droite.

Exemple 5

Tracer la droite $Y = 4 - 2X$.

L'ordonnée à l'origine est 4; la droite traverse l'axe des Y au point (0; 4). La pente est -2; si X augmente de 1, Y décroît de 2. Les points P (0; 4) et Q (1; 2) déterminent la droite.



15.4 ÉQUATION D'UNE DROITE CONNAISSANT DEUX POINTS

Le graphique de l'équation $Y = a + bX$ est une droite. Pour trouver l'équation de la droite passant par deux points connus, on doit déterminer les valeurs de a et b. Il est facile de trouver b en calculant la pente de la droite comme à la section précédente. Il suffit de prendre un des deux points; sachant qu'il se trouve sur la droite, ses coordonnées vérifient l'équation de la droite. On peut alors déterminer la valeur de a.

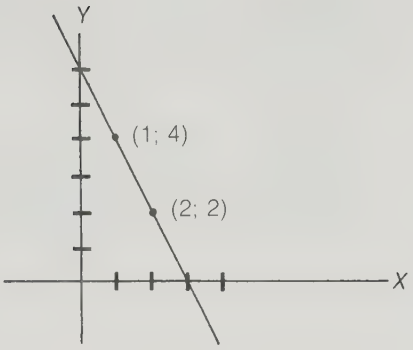
Examiner les exemples suivants.

Exemple 1

Trouver l'équation de la droite passant par les points (1; 4) et (2; 2).

L'équation est $Y = a + bX$, $a = ?$, $b = ?$

La pente $= \dfrac{4-2}{1-2} = -2$, donc $b = -2$ et l'équation devient $Y = a - 2X$. Le point P (1; 4) est sur la droite, donc les coordonnées vérifient l'équation; ainsi, $4 = a - 2 * 1$, donc $a = 6$; finalement $Y = 6 - 2X$ (on peut vérifier que cette droite passe par le point (2; 2)).

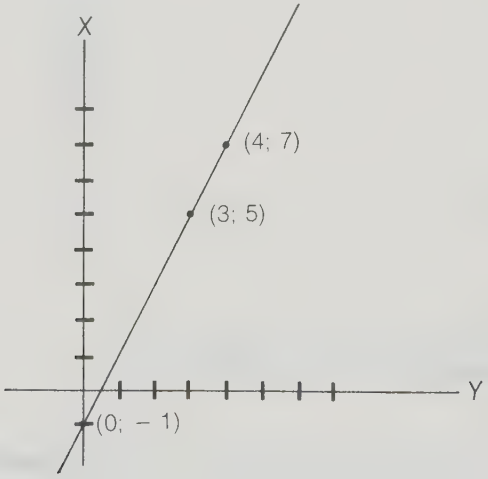


Exemple 2

Trouver (a) l'équation et (b) l'ordonnée à l'origine de la droite passant par (3; 5) et (4; 7).

(a) La pente $= \dfrac{5-7}{3-4} = 2 = b$; $Y = a + 2X$. Le point (3; 5) est sur la droite, donc $5 = a + 2 * 3$ et $a = -1$. L'équation de la droite est $Y = -1 + 2X$.

(b) Quand $X = 0$, $Y = -1$ qui est l'ordonnée à l'origine.

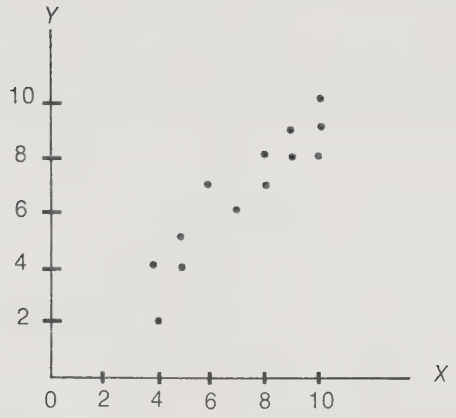


15.5 LA DROITE DE RÉGRESSION

Supposons que le nombre d'heures d'étude pour préparer l'examen final en statistiques et le nombre de bonnes réponses pour chaque étudiant sont donnés par le tableau suivant:

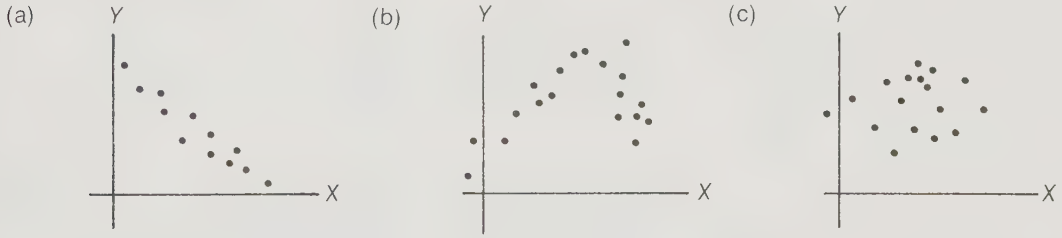
Heures d'étude (X)	5	8	6	9	10	8	5	4	10	4	10	7	9
Bonnes réponses (Y)	5	8	7	9	10	7	4	4	8	2	9	6	8

Ces données peuvent être représentées dans un graphique en associant à chaque paire de données un point de ce graphique. Un tel graphique s'appelle un **diagramme de dispersion.**



Nous verrons comment tracer la droite qui représente le mieux l'ensemble de ces points et comment trouver son équation. Cette droite s'appelle **droite de régression.** Elle fut découverte par Sir Francis Galton qui en comparant les grandeurs des parents et de leurs enfants a découvert que: des parents de petite taille ont des enfants plus petits que la moyenne, mais pas aussi petits que leurs parents et que des parents de grande taille ont des enfants plus grands que la moyenne, mais pas aussi grands que leurs parents. C'est une régression par rapport à la moyenne.

Il sera toujours possible de trouver une droite qui représente un ensemble de points d'un diagramme de dispersion. Quelquefois, il n'est pas bon de le faire. Plus tard, nous étudierons la corrélation, ce qui nous donnera une mesure permettant de savoir si une dispersion est bien représentée par une droite. D'ici là, nous devrons tracer un diagramme de dispersion pour voir s'il semble raisonnable de la représenter par une droite comme en (a) ci-dessous. Nous ne pourrons pas trouver de droite de régression dans des dispersions qui ressemblent aux figures (b) et (c) ci-dessous.

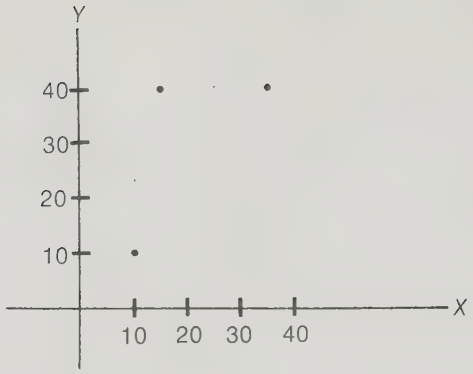


Dans un cours plus avancé, on montre comment trouver une courbe de régression de la forme $Y = a + bX + cX^2$ pour représenter la dispersion (b). Dans ce cours, on se limite à la droite de régression.

L'équation de régression (équation de la droite de régression) permet d'utiliser une variable pour prédire l'autre. Si on connaît les résultats d'un étudiant en secondaire V, peut-on prédire son résultat en première année du collégial? Est-ce qu'une compagnie qui fait l'assemblage de petits appareils peut, à la suite d'un test de dextérité manuelle, prédire qui parmi les nouveaux employés seront les assembleurs les plus rapides? La variable qui est connue (les résultats du secondaire V, les résultats du test de dextérité) est appelée **variable indépendante** et est habituellement représentée par X; la variable que l'on doit prédire (le résultat au collégial, le nombre d'unités assemblées) est appelée **variable dépendante** et est habituellement représentée par Y.

Supposons que l'on donne les paires X et Y suivantes:

X	Y
10	10
15	40
35	40



Nous devons trouver une équation $Y_c = a + bX$ qui permet de prédire Y quand X est connu. Par exemple, si $X = 20$, quelle sera la valeur de Y? Si a et b peuvent être déterminés, la réponse sera $a + 20b$. L'indice c est utilisé pour distinguer la valeur calculée de Y ou la valeur prédite (Y_c) des valeurs (Y) des paires données.

La différence entre la valeur donnée Y et la valeur prédite Y_c est appelée l'**erreur d'estimation.** La **droite de régression** ou la droite qui représente le mieux l'ensemble des paires est la droite pour laquelle la **somme des carrés de ces erreurs d'estimation est minimum.** C'est la **méthode des moindres carrés.**

Exemple 1

Laquelle parmi les équations suivantes représente le mieux les trois paires données précédemment?

(a) $Y_c = 30$, obtenue en posant $a = Y$ et $b = 0$.

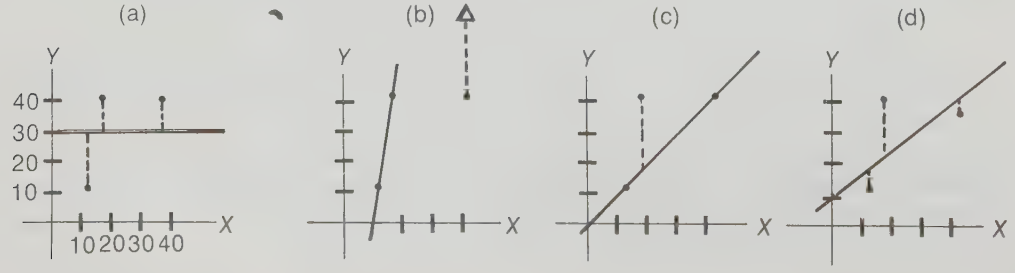
(b) $Y_c = -50 + 6X$, obtenue en utilisant les deux premiers points (10; 10) et (15; 40).

(c) $Y_c = -2 + 1{,}2X$, équation qui vérifie la première et la troisième paire: (10; 10) et (35; 40).

(d) $Y_c = 12{,}8 + 0{,}86X$: c'est l'équation de régression; nous verrons comment la trouver.

RÉPONSES

		(a) $Y_c = 30$			(b) $Y_c = -50 + 6X$			(c) $Y_c = -2 + 1{,}2X$			(d) $Y_c = 12{,}8 + 0{,}86X$		
X	Y	Y_c	$Y - Y_c$	$(Y - Y_c)^2$	Y_c	$Y - Y_c$	$(Y - Y_c)^2$	Y_c	$Y - Y_c$	$(Y - Y_c)^2$	Y_c	$Y - Y_c$	$(Y - Y_c)^2$
10	10	30	−20	400	10	0	0	10	0	0	21,4	− 11,4	130,0
15	40	30	10	100	40	0	0	16	24	576	25,7	14,3	204,5
35	40	30	10	100	160	−120	14 400	40	0	0	42,9	2,9	8,4
				600			14 400			576			342,9



La somme des carrés des erreurs est moindre pour la dernière équation $Y_c = 12{,}8 + 0{,}86X$; c'est donc le meilleur choix pour représenter l'ensemble des trois paires.

Pour trouver l'équation de régression, on doit utiliser des mathématiques plus avancées. On peut montrer que les valeurs de «*b*» et «*a*» sont données par:

$$b = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n} = \overline{Y} - b\overline{X}$$

Attention

(1) n est le nombre de **paires** $(X; Y)$ qui sont utilisées pour calculer la droite de régression. Dans l'exemple précédent, $n = 3$.

(2) Bien faire la distinction entre $(\Sigma X)^2$ et ΣX^2; $(\Sigma X)^2$ indique d'additionner en premier puis élever la somme au carré alors que ΣX^2 indique d'élever chaque X au carré puis d'additionner les carrés.

(3) Le dénominateur de b est toujours plus grand que 0. Il est $n(n-1)$ fois la variance des résultats X et la variance est toujours plus grande que 0 sauf si tous les X sont identiques.

Exemple 2

Trouver la droite de régression des paires $(X; Y)$ données ci-dessous.

X^2	X	Y	XY
100	10	10	100
225	15	40	600
1225	35	40	1400
1550	60	90	2100

Noter la disposition des colonnes X^2 et XY.

Réponse

$$b = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{3 * 2100 - 60 * 90}{3 * 1550 - 60^2} = \frac{6300 - 5400}{4650 - 3600} = \frac{900}{1050} = 0{,}857$$

$$a = \frac{90}{3} - \frac{60}{3} * 0{,}86 = 30 - 20 * 0{,}86 = 12{,}8$$

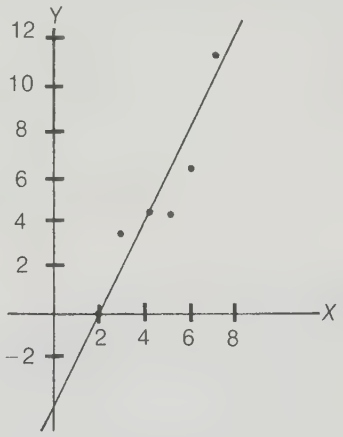
L'équation de la droite de régression est alors

$$Y_c = 12{,}8 + 0{,}86X.$$

Exemple 3

Trouver la droite de régression des 6 paires (X;Y) données ci-dessous.

X^2	X	Y	XY
4	2	0	0
9	3	3	9
16	4	4	16
25	5	4	20
36	6	6	36
49	7	11	77
139	27	28	158



$$b = \frac{6 * 158 - 27 * 28}{6 * 139 - 27^2} = \frac{948 - 756}{834 - 729} = \frac{192}{105} = 1{,}829$$

$$a = \frac{28}{6} - 1{,}829 * \frac{27}{6} = -3{,}56$$

$$Y_c = -3{,}56 + 1{,}83X$$

Ne pas oublier que $n = 6$.

Quelques fois $n\Sigma XY$ est approximativement égal à $(\Sigma X)(\Sigma Y)$, il faut éviter de faire des calculs approximatifs et d'arrondir ces résultats avant la soustraction quand on calcule la valeur de b. Voici un exemple:

Exemple 4

(a) Trouver l'équation de la droite de régression pour les valeurs de X et Y données ci-dessous.

(b) Quelle est la valeur de Y quand $X = 72$?

X^2	X	Y	XY
4900	70	127	8890
4356	66	112	7392
5625	75	146	10950
5041	71	131	9301
4489	67	116	7772
3844	62	97	6014
28 255	411	729	50 319

(a) $$b = \frac{6 * 50\,319 - 411 * 729}{6 * 28\,255 - 411^2} = \frac{301\,914 - 299\,619}{169\,530 - 168\,921} = \frac{2\,295}{609} = 3{,}77$$

$$a = \frac{729}{6} - \frac{411}{6} * 3{,}77 = -137$$

$$Y_c = -137 + 3{,}77X$$

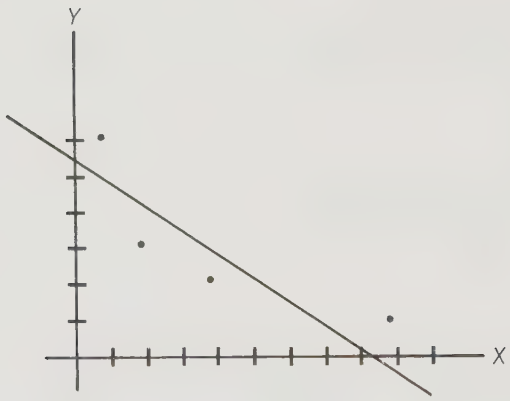
(b) Quand $X = 72$, $Y_c = -137 + 3{,}77 * 72 = 134$.

Il est à noter que cette valeur (134) est sans signification si la distribution de la population Y ne vérifie pas certaines hypothèses comme nous le verrons au chapitre 16. Dans les exemples précédents, b était toujours positif. La pente de la droite de régression peut être négative.

Exemple 5

Quelle est la valeur de Y quand $X = 6$, connaissant les valeurs ci-dessous?

X^2	X	Y	XY
1	1	6	6
4	2	3	6
16	4	2	8
64	8	1	8
85	15	12	28



Réponse

$$b = \frac{4 * 28 - 15 * 12}{4 * 85 - 15^2} = \frac{112 - 180}{340 - 225} = \frac{-68}{115} = -0{,}591$$

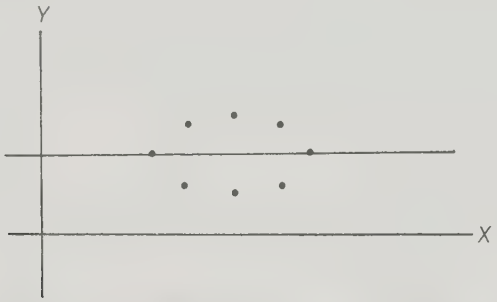
$$a = \frac{12}{4} - (-0{,}591) * \frac{15}{4} = 3 + 2{,}21 = 5{,}22$$

$$Y_c = 5{,}22 - 0{,}59X$$

Quand $X = 6$, $Y_c = 5{,}22 - 0{,}59 * 6 = 1{,}68$.

Une équation de régression est utilisée pour prédire un résultat (dans le dernier exemple $Y_c = 1{,}68$ quand $X = 6$). La prédiction est sans signification si on ne peut estimer sa validité et ceci peut être fait à partir d'hypothèses concernant la distribution de la population. Nous en discuterons au chapitre suivant. Pour l'instant, concentrez-vous sur ce qui suit.

1. Une équation de régression $Y_c = a + bX$ peut être trouvée même si les données sont loin d'être linéaires; par exemple:



chercher l'équation de régression dans un tel cas est une perte de temps. Pour éviter cette situation, il est recommandé de toujours tracer un diagramme de dispersion pour voir si les données sont suffisamment près d'une droite.

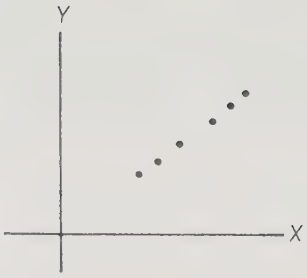
2. Nous avons trouvé la meilleure droite pour représenter un ensemble de paires. C'est la droite de régression trouvée par la méthode des moindres carrés. Dans un cours plus avancé, il est possible de trouver par cette même méthode des courbes de régression du type $Y_c = a + b \log X$ ou $Y_c = a + bX + cX^2$.

3. On peut toujours trouver une équation de régression même s'il n'y a pas de relation entre X et Y et l'existence de cette équation ne prouve pas qu'une telle relation existe.

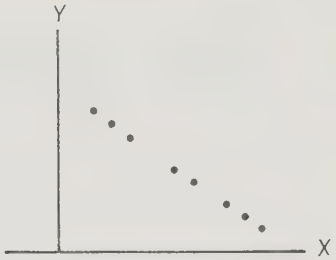
15.6 COEFFICIENT DE CORRÉLATION: APPROCHE INTUITIVE

Un coefficient de corrélation tout comme une équation de régression est déterminé pour des résultats X et Y qui se présentent sous forme de paires. Ce coefficient de corrélation r mesure la proximité de la linéarité d'une relation entre deux variables et de ce fait permet de déterminer si l'équation de régression étudiée à la section précédente peut être utilisée pour prédire des résultats. Il est à noter que dans les problèmes de corrélation, il n'est pas nécessaire de considérer une variable comme indépendante et l'autre comme dépendante. Par exemple, on peut considérer la relation qui existe entre le poids et la grandeur des individus ou vice-versa.

Quand toutes les paires $(X; Y)$ sont sur la droite de régression, le coefficient de corrélation est soit $+1$ soit -1; il est $+1$ si la pente de la droite de régression est positive (dans ce cas on dit que la corrélation est **parfaite et positive**). Si la pente de la droite de régression est négative le coefficient de corrélation est -1 (et la corrélation est **parfaite et négative**).



$r = +1$



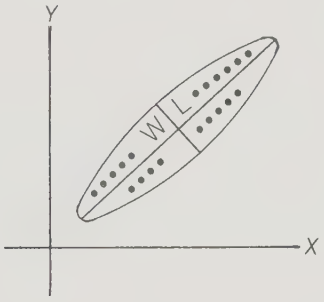
$r = -1$

Si quelques paires $(X; Y)$ ne sont pas sur la droite de régression, le coefficient de corrélation est compris entre -1 et 1. (Attention: si le calcul de coefficient de corrélation donne $r = 2{,}3$ ou $r = -1{,}5$ il faut vérifier ses calculs, il y a une erreur.)

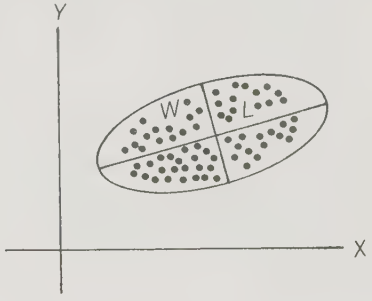
La valeur de r peut être estimée rapidement en encerclant l'ensemble des points du diagramme de dispersion puis en mesurant la longueur L et la largeur W de cet ensemble puis en calculant

$$r = \pm\left(1 - \frac{W}{L}\right)$$

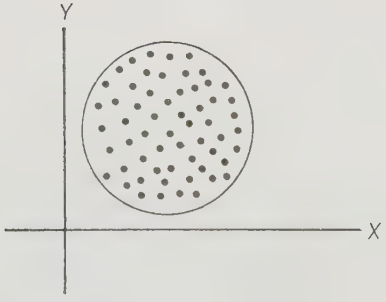
Le signe de r est déterminé par le signe de la pente de L. On peut utiliser les unités que l'on veut, mais il est plus facile de poser la largeur W comme valant 1 unité puis exprimer L comme un multiple de W. Voici quelques exemples:



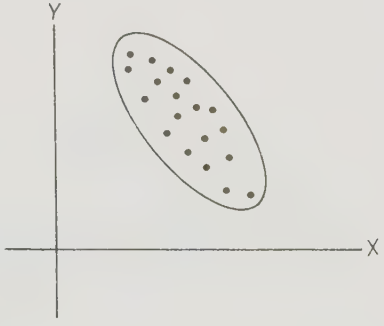
$W = 1, L = 4, r \approx 1 - \frac{1}{4} = \frac{3}{4}$



$W = 1, L = 2, r \approx 1 - 1/2 = 1/2$



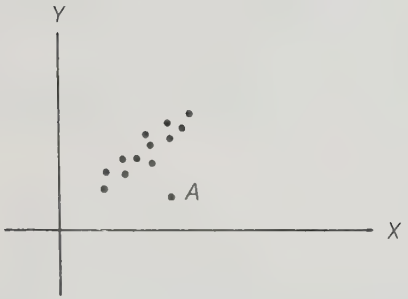
$W = 1, L = 1, r \approx 1 - \frac{1}{1} = 0$

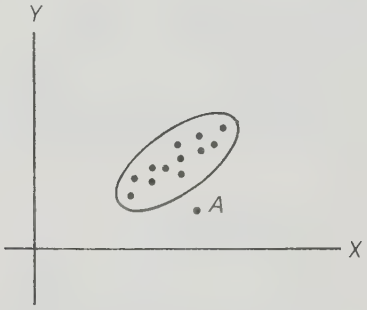


$W = 1, L = 3, r \approx -\left(1 - \frac{1}{3}\right) = -\frac{2}{3}$

Cette méthode donne un bon résultat seulement si la densité des points à l'intérieur de l'ellipse est assez grande. Malheureusement, en pratique les données ont quelquefois tendance à s'éparpiller un peu trop et des points isolés rendent difficile le tracé de l'ellipse. Si on inclut le point *A*, il en résulte une plus petite valeur de *r*. Dans ce cas, il sera préférable de l'exclure et d'avoir une meilleure estimation de *r*.

$$1 - \frac{W}{L} = 1 - \frac{1}{4} = \frac{3}{4}; \; r \approx 0{,}7$$





15.7 ERREUR TYPE D'UN ESTIMÉ, VARIATION TOTALE ET VARIATION EXPLIQUÉE

Le fait que quelques points ne sont pas sur la droite de régression ne peut nous empêcher d'affirmer que cette droite représente une relation entre *X* et *Y*. Si ces résultats proviennent d'une enquête psychologique, on prévoit une certaine erreur dans les observations. Le terme «erreur expérimentale» comporte des erreurs qui peuvent ne pas être dues à l'expérience (comme arrondir un résultat) et des erreurs de mesures (dues à l'imprécision d'un appareil). L'**erreur-type d'un estimé** (s_e) est une mesure de ces erreurs.

$$s_e = \sqrt{\frac{\Sigma (Y - Y_c)^2}{n - 2}}$$

(Il y a $n - 2$ degrés de liberté, car *a* et *b* de l'équation de régression sont utilisés dans le calcul de Y_c.)

Erreur de variation. Le numérateur $\Sigma (Y - Y_c)^2$ de l'équation ci-dessus est appelé l'erreur de variation. (Dans cette section, on se concentrera sur la variation au lieu de la variance ou de l'écart-type. Nous reviendrons à l'erreur-type d'un estimé au chapitre 16.)

Rappelons-nous que dans la recherche de la droite de régression, nous avions utilisé le critère suivant; la somme des carrés des erreurs devait être aussi petite que possible (moindres carrés). Ainsi si les Y_c sont calculés avec d'autres valeurs de *a* et *b* dans l'équation de régression, l'erreur-type de l'estimé et l'erreur de variation seront plus grandes.

Exemple 1

Trouver (a) l'erreur de variation (b) l'erreur-type d'un estimé pour les paires $(X; Y)$ ci-dessous. L'équation de régression est $Y_c = -3{,}564 + 1{,}829X$ (trouvée à l'exemple 3 de la section 15.5).

X	Y	Y_c	$Y - Y_c$	$(Y - Y_c)^2$
2	0	0,094	− 0,094	0,009
3	3	1,923	1,077	1,160
4	4	3,752	0,248	0,062
5	4	5,581	− 1,581	2,500
6	6	7,410	− 1,410	1,988
7	11	9,239	1,761	3,101
	28	28,00	0,001	8,820

(a) L'erreur de variation $= \Sigma (Y - Y_c)^2 = 8{,}82$.

(b) $s_e = \sqrt{8{,}82/4} = 1{,}48$.

Il est à remarquer que $\Sigma Y = \Sigma Y_c$ et que $\Sigma (Y - Y_c) = 0$, ce qui permet une vérification des calculs.

Variation expliquée. S'il n'y a pas de relation linéaire entre X et Y, il n'est pas nécessaire de prédire la valeur de Y à partir d'un résultat X. La meilleure estimation pour Y_c pour tout X est $Y_c = \overline{Y}$. Dans ce cas, $\Sigma (Y_c - \overline{Y})^2 = 0$. S'il y a une relation linéaire entre X et Y, $\Sigma (Y_c - \overline{Y})^2$ mesure l'avantage qu'on à utiliser Y_c à la place de $\overline{Y}$; $\Sigma (Y_c - \overline{Y})^2$ est appelée la **variation expliquée.**

Exemple 2

Pour les paires données à l'exemple 1, calculer la variation expliquée.

X	Y	Y_c	$Y_c - \overline{Y}$	$(Y_c - \overline{Y})^2$
2	0	0,094	− 4,572	20,90
3	3	1,923	− 2,746	7,54
4	4	3,752	− 0,914	0,84
5	4	5,581	0,915	0,84
6	6	7,410	2,744	7,53
7	11	9,239	4,573	20,91
	28		0,000	58,56

La variation expliquée $= \Sigma (Y_c - \overline{Y})^2 = 58{,}56$.

Remarquer que la somme des $Y_c - \overline{Y}$ est 0 ce qui permet une vérification des calculs.

Variation totale. La variation des Y par rapport à leur moyenne peut être calculée. La $\Sigma (Y - \overline{Y})^2$ est appelée **variation totale.**

Exemple 3

Pour les paires $(X; Y)$ données à l'exemple 1, trouver la variation totale.

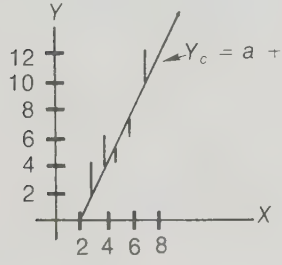
X	Y	$(Y - \overline{Y})$	$(Y - \overline{Y})^2$
2	0	− 4,667	21,78
3	3	− 1,667	2,78
4	4	− 0,667	0,44
5	4	− 0,667	0,44
6	6	1,333	1,78
7	11	6,333	40,11
	28	− 0,002	67,33

$$\overline{Y} = \frac{28}{6} = 4{,}67$$

$$\text{Variation totale} = \Sigma(Y - \overline{Y})^2 = 67{,}33$$

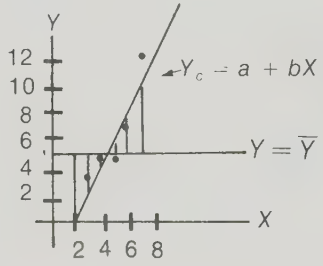
Encore une fois la colonne $Y - \overline{Y}$ a une somme nulle ce qui permet une vérification des calculs.

Examiner attentivement les figures suivantes, vous verrez les différentes longueurs qui sont utilisées dans le calcul des variations.



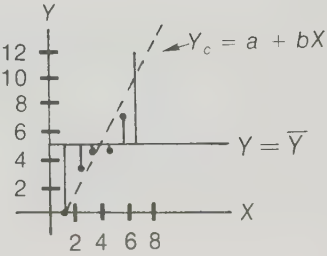
Erreur de Variation

$(Y - Y_c)^2$



Variation expliquée

$(Y_c - \overline{Y})^2$



Variation totale

$(Y - \overline{Y})^2$

On peut montrer que: $\Sigma(Y - Y_c)^2 + \Sigma(Y_c - \overline{Y})^2 = \Sigma(Y - \overline{Y})^2$:

Erreur de variation + variation expliquée = variation totale

La même relation s'applique aux variances.

Erreur de variance + variance expliquée = variance totale.

Exemple 4

Vérifier que la somme de l'erreur de variation et de la variation expliquée égale la variation totale pour les paires $(X; Y)$ de l'exemple 1.

	Variation	Formule
Erreur	8,82	$\Sigma(Y - Y_c)^2$
Expliquée	58,56	$\Sigma(Y_c - \overline{Y})^2$
Total	67,33	$\Sigma(Y - \overline{Y})^2$

$$8{,}82 + 58{,}56 = 67{,}33$$

À la précision des calculs près, la vérification est faite.

Si tous les points sont sur la droite de régression, chaque $Y - Y_c = 0$ et $\Sigma (Y - Y_c)^2 = 0$; la variation est égale à la variation totale. Si le fait de connaître X n'est d'aucune utilité pour prédire Y, alors le meilleur choix pour une équation de régression est $Y_c = \overline{Y}$; cette fois, la variation expliquée $\Sigma (Y_c - \overline{Y})^2 = 0$.

Le rapport de la variation expliquée à la variation totale mesure l'ajustement de la régression linéaire aux paires de valeurs données. Il est appelé **coefficient de détermination** et est noté r^2.

$$r^2 = \frac{\text{Variation expliquée}}{\text{Variation totale}}$$

Dans l'exemple précédent:

$$r^2 = \frac{58{,}56}{67{,}33} = 0{,}87$$

La variation expliquée n'est jamais négative (c'est la somme des carrés) et n'est jamais plus grande que la variation totale donc r^2 est toujours compris entre 0 et 1. Si la variation expliquée est nulle, $r^2 = 0$; ceci se produit quand la valeur de X ne permet pas de prédire la valeur de Y et la meilleure valeur de Y_c dans ce cas est $\overline{Y}$.

Si r^2 est connu, alors $r = \pm \sqrt{r^2}$ où le signe de r **est le même que celui de b de l'équation de régression.** «r» s'appelle **coefficient de corrélation** et est donné par:

$$r = \pm \sqrt{\frac{\text{Variation expliquée}}{\text{Variation totale}}}$$

Comme r^2 est compris entre 0 et 1, r est compris entre -1 et $+1$.

Dans l'exemple précédent, l'équation de régression était $Y_c = -3{,}57 + 1{,}83X$ comme $b = +1{,}83$ alors $r = +\sqrt{0{,}87} = 0{,}93$.

Une valeur de r près de 1 ou -1 indique un bon ajustement, car la variation expliquée est une grande partie de la variation totale; une valeur de r près de 0 indique un mauvais ajustement, l'erreur de variation étant la plus grande partie de la variation totale.

15.8 CALCUL DE r

Si vous avez bien compris la définition du coefficient de corrélation r donnée précédemment, vous savez à quoi il sert; il mesure la dépendance linéaire des variables X et Y

de chaque paire et donne le signe de la pente de régression. Pour trouver r, il faut normalement trouver l'équation de régression $Y_c = a + bX$; on peut éviter cet inconvénient en utilisant la formule ci-dessous:

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\ \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

Dans l'exemple qui suit, bien remarquer la disposition des colonnes et bien distinguer entre ΣXY et $(\Sigma X)(\Sigma Y)$.

Exemple 1

Trouver r pour les données des problèmes précédents.

X^2	X	Y	Y^2	XY
4	2	0	0	0
9	3	3	9	9
16	4	4	16	16
25	5	4	16	20
36	6	6	36	36
49	7	11	121	77
139	27	28	198	158

$$n = 6;\ r = \frac{6 * 158 - 27 * 28}{\sqrt{6 * 139 - 27^2} * \sqrt{6 * 198 - 28^2}} = \frac{948 - 756}{\sqrt{105} * \sqrt{404}} = \frac{192}{206} = 0{,}93$$

Exemple 2

Trouver r pour les paires $(X; Y)$ ci-dessous.

X^2	X	Y	XY	Y^2
1	1	6	6	36
4	2	3	6	9
16	4	2	8	4
64	8	1	8	1
85	15	12	28	50

$$r = \frac{4 * 28 - 15 * 12}{\sqrt{4 * 85 - 15^2}\ \sqrt{4 * 50 - 12^2}} = \frac{112 - 180}{\sqrt{340 - 225}\ \sqrt{200 - 144}} = \frac{-68}{\sqrt{115 * 56}} = \frac{-68}{80{,}2} = -0{,}85$$

Comme $n\Sigma X^2 - (\Sigma X)^2 = n(n-1)\left[\dfrac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}\right]$ et $\dfrac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}$

est la variance des résultats X, la quantité sous la racine carrée est positive. Il en est de même pour $n\Sigma Y^2 - (\Sigma Y)^2$. Si un signe négatif apparaît sous un radical, il faut vérifier ses calculs. Il faut se rappeler également que n représente le nombre de paires $(X; Y)$.

Exemple 3

Trouver r

X^2	X	Y	XY	Y^2
	1	4		
	2	5		
	3	6		

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\ \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}} = ?$$

Réponse

$$r = \frac{3 * 32 - 6 * 15}{\sqrt{3 * 14 - 6^2}\ \sqrt{3 * 77 - 15^2}} = \frac{6}{6} = 1$$

Est-ce surprenant? Non, les trois points sont sur la même droite.

Il existe d'autres formules pour trouver r. La similarité entre la formule pour trouver r ci-dessus et la formule pour trouver «b» dans l'équation de régression $Y_c = a + bX$ est assez frappante.

$$b = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

$$r = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} * \frac{\sqrt{n\Sigma X^2 - (\Sigma X)^2}}{\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}} = b * \frac{\sqrt{n\Sigma X^2 - (\Sigma X)^2}}{\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

Si le numérateur et le dénominateur de la fraction de droite sont divisés par $\sqrt{n(n-1)}$ ils deviennent respectivement les écarts-types des résultats X et Y. On utilise les symboles s_x et s_y. En substituant on obtient:

$$r = b * \frac{s_x}{s_y}$$

Cette formule est utile si on a déjà trouvé l'équation de régression et les écart-types.

Exemple 4

Dans l'exemple 5 de la section 15.5, on a déjà trouvé l'équation de régression $Y_c = 5{,}22 - 0{,}59X$. Trouver r. Se rappeler que $s_x = \frac{\Sigma X^2 - (\Sigma X)^2/n}{n-1}$.

X	Y	X^2	Y^2
1	6	1	36
2	3	4	9
4	2	16	4
8	1	64	1
15	12	85	50

$$s_X = \sqrt{\frac{85 - 15^2/4}{3}} = 3{,}10$$

$$s_Y = \sqrt{\frac{50 - 12^2/4}{3}} = 2{,}16$$

$$b = -0{,}59$$

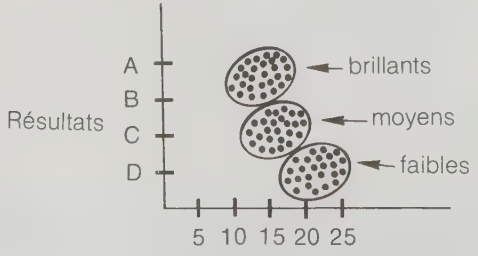
$$r = -0{,}59 * \frac{3{,}10}{2{,}16} = -0{,}85$$

15.9 QUELQUES REMARQUES

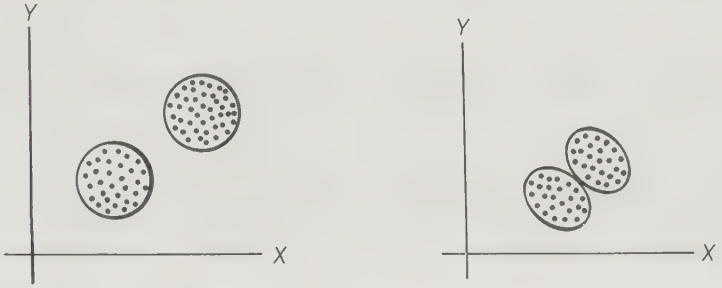
1. Même si deux variables ont une grande corrélation (positive ou négative) ceci n'implique pas qu'elles ont une relation de cause à effet. Une étude portant sur 100 villes choisies au hasard montre qu'il y a une corrélation positive entre le nombre d'églises et le nombre de discothèques. Doit-on en conclure que la religion porte à boire ou vice-versa?

   Il y a une forte corrélation entre le fait de porter des pantalons longs et le fait de bien lire; les élèves de 1re année portent en général des pantalons courts et lisent difficilement et ceux de 6e année portent des pantalons longs et lisent bien. Doit-on en conclure qu'il suffirait de faire porter des pantalons longs aux élèves de 1re année pour qu'ils lisent bien?

2. Quelquefois, un échantillon peut montrer une corrélation entre deux variables, alors qu'un autre échantillon peut ne pas montrer la même corrélation.
   Quelques personnes pensent qu'il y a une corrélation négative entre les résultats et le temps d'étude. (Il n'est probablement pas vrai que les étudiants brillants étudient moins que les faibles.) Même en supposant que la corrélation est négative, est-ce que vous obtiendrez de meilleurs résultats en étudiant moins? Dans un cours particulier, il peut arriver que les étudiants brillants étudient moins et obtiennent de meilleurs résultats que les étudiants faibles même s'ils ont étudié plus longtemps.



En général, si une population est composée de plusieurs groupes ayant des moyennes différentes, une corrélation apparente peut être obtenue ou perdue comme le montre les diagrammes suivants:

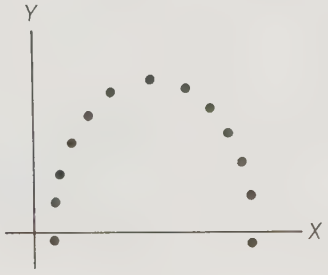


La meilleure façon de remédier à cet inconvénient est (a) de toujours tracer un diagramme d'éparpillement et (b) étudier attentivement la provenance de l'échantillon.

3. Attention à l'interprétation de *r*.

L'énoncé suivant est à la fois utile et dangereux. Si *r* est plus grand que 0,7, les variables *X* et *Y* sont fortement liées; si *r* est entre 0,3 et 0,7 il y a une certaine relation; si *r* est entre 0 et 0,3, l'association est négligeable. Le même énoncé peut être fait avec les valeurs négatives corresponsantes de *r*.

Pourquoi cet énoncé est-il dangereux? Premièrement, parce que *r* mesure seulement les relation linéaires. Dans le diagramme d'éparpillement ci-dessous, *r* sera près de 0 même s'il existe une étroite relation entre *X* et *Y* (cette relation n'est pas linéaire et est plutôt de forme quadratique). Quand on obtient une faible corrélation, le diagramme d'éparpillement peut montrer un autre type de relation (mais pas nécessairement).



Deuxièmement, il faut se rappeler que r^2 mesure la proportion de variation des résultats *Y* expliquée par la relation linéaire entre *X* et *Y*. Si $r = 0{,}3$, alors $r^2 = 0{,}09$ et seulement 9% de la variation des résultats *Y* est expliqué. Si $r = 0{,}7$, alors 49% est expliqué et ainsi de suite.

Finalement, quand peut-on dire qu'il y a une grande corrélation entre les variables? La réponse dépend de la façon dont *r* sera utilisé. Un sociologue qui étudie la délinquance juvénile veut savoir si le nombre d'arrestations est en corrélation avec différentes variables: attitude envers l'autorité, éducation des parents, Q.I., la pratique religieuse et ainsi de suite. Toute corrélation supérieure à 0,2 peut indiquer une relation suffisante pour poursuivre l'étude de la variable. Ce qui semble être une faible corrélation pour un physicien ou un chimiste peut être une forte corrélation pour un sociologue.

4. Jusqu'à maintenant nous n'avons pas fait d'hypothèses concernant les populations d'où sont prélevés les résultats X et Y. Même sans ces hypothèses, on peut déterminer l'équation de régression et le coefficient de corrélation d'un échantillon comme si c'était une partie de la statistique. Quand nous ferons de l'inférence sur les populations, ces hypothèses seront essentielles. Nous y reviendrons au prochain chapitre.

15.10 LA RELATION ENTRE LA RÉGRESSION ET LA CORRÉLATION

La corrélation mesure la force de la relation linéaire entre deux variables alors que la droite de régression décrit cette relation. Il y a d'autres différences.

Dans l'équation de régression, la variable (X) est indépendante et sert à prédire la variable dépendante (Y). Dans la corrélation, on ne fait pas de distinction entre les deux variables.

Un sociologue utilise plus fréquemment la corrélation; il recueille les données de plusieurs variables et cherche celles qui sont en relation linéaire. Un économiste utilise plus fréquemment la régression; les données recueillies sont souvent plus exactes et la théorie lui indique souvent qu'il peut exister une relation linéaire. Il utilisera la corrélation si aucune indication ne lui permet de déceler une relation linéaire; il déterminera par la suite l'équation de régression.

En corrélation, les deux variables peuvent avoir une grande étendue. Ceci est quelquefois vrai, mais pas toujours pour la prédiction à partir d'une droite de régression. Par exemple si on étudie la variation du point de congélation d'un produit chimique en fonction de la quantité d'eau ajoutée, les valeurs de X (eau ajoutée en cm^3) ne sont pas choisies au hasard, mais ont des valeurs fixes (10 cm^3, 20 cm^3, ...).

15.11 VOCABULAIRE ET SYMBOLES

Graphique d'une droite
Pente d'une droite
Relation linéaire
Droite de régression
Erreur d'estimation
Moindres carrés
$Y_c = a + bX$
Coefficient de corrélation r
Erreur-type d'estimation
s_e
Erreur de variation
Valeur expliquée
Variation totale
Coefficient de détermination r^2

15.12 EXERCICES

1. Pour chacune des équations suivantes, déterminer la pente, l'ordonnée à l'origine puis tracer le graphique.
   (a) $Y = -2 + 3X$.
   (b) $Y = 5 - 2X$.
   (c) $X = 4$.
   (d) $X + 2Y = 4$.
2. Trouver l'équation de la droite passant par les points suivants puis tracer le graphique.
   (a) P (0; 4) Q (4; 6).
   (b) P (1; 3) Q (2; 1).
   (c) P (3; 0) Q (3; 4).
   (d) P (-2; 5) Q (4; -3).

3. En calculant $\Sigma (Y - Y_c)^2$, déterminer laquelle des équations suivantes représente le mieux les paires de résultats suivants (1; 2) (2; 2) (2; 3) (3; 4).

(a) $Y_c = 1 + X$.

(b) $Y_c = 1 - 0{,}5X$.

(c) $Y_c = 3/4 + X$.

4. Dans un examen à choix multiples, le professeur fait une erreur en calculant l'équation de régression des points (1; 3) (1; 4) (2; 5) (3; 7) et donne les choix suivants (a) $Y_c = 1$ (b) $Y_c = 1 + X$ (c) $Y_c = 1 + 3X$ et (d) $Y_c = 3X$. Malheureusement, aucune de ces équations n'est celle de régression. En se basant sur la méthode des moindres carrés, laquelle de ces équations représente le mieux les points précédents?

5. Dessiner un diagramme d'éparpillement pour chaque cas.

(a)

X	3	5	6	2	3	8	6	4	7	1	2	4
Y	5	1	1	6	4	6	2	2	4	9	5	3

(b)

X	32	36	30	36	34	35	30	32	38	36	37	34	33
Y	5	10	4	8	7	7	2	7	12	9	10	6	5

Pour quel cas semble-t-il logique de trouver une équation de régression linéaire? Expliquer.

6. Utiliser la méthode des moindres carrés pour trouver la droite de régression. Pour chaque cas, tracer un graphique comportant les données et la droite.

(a)

X	Y
1	2
2	2
2	3
3	4

(b)

X	Y
1	14
3	12
6	9
7	8
13	2

Pour chaque cas prédire la valeur de Y pour $X = 4$.

7. Dans les données suivantes X représente le Q.I. et Y le résultat à un examen de statistique.

X^2	X	Y	XY
10000	100	10	1000
12100	110	12	1320
14400	120	15	1800
16900	130	18	2340
19600	140	20	2800
73000	600	75	9260

(a) Tracer un diagramme d'éparpillement.

(b) Trouver l'équation de la droite qui représente le mieux l'ensemble des données.

(c) Quelle est l'ordonnée à l'origine de la droite de régression? Sa pente?
(d) Quel serait le résultat d'un étudiant qui a un Q.I. de 125? de 105?
(e) Tracer la droite de régression dans le diagramme de dispersion.

8. Un étudiant en physique veut déterminer la force gravitationnelle; il enregistre la vitesse d'un corps en chute libre après un intervalle de temps. Voici les résultats obtenus.

Temps (s) (X)	Vitesse (Y)
1/4	3
1/2	7
3/4	10
1	11
5/4	17

(a) Utiliser la méthode des moindres carrés pour trouver la droite de régression.
(b) Faire un graphique représentant les points et la droite de régression.
(c) Quelle sera la valeur de Y si $X = 7/8$?

9. À la graduation, les résultats en chimie de six étudiants sont comparés à leur résultat moyen total.

Moyenne	Chimie
600	2,8
720	3,8
640	3,6
590	2,9
620	3,2
680	3,6

(a) Tracer un diagramme d'éparpillement.
(b) Trouver l'équation de la droite de régression.
(c) Donner l'ordonnée à l'origine et la pente de cette droite.
(d) Quel serait le résultat de chimie d'un étudiant ayant une moyenne de 700?

10.

X	0	1	5
Y	3	4	5

Pour ces résultats $\overline{X} = 2, \overline{Y} = 4$. Chacune des droites suivantes passent par $(\overline{X}; \overline{Y}) = (2; 4)$; (a) $Y_c = 4$ (b) $Y_c = 2X$ (c) $Y_c = 4 - 2b + bX$. (Si $Y_c = a + bX$ passe par (2; 4) alors $a = 4 - 2b$).
Pour chaque droite montrer que $\Sigma (Y - Y_c) = 0$.

11. La taille et le poids de 5 hommes sont les suivants:

X: Taille (cm)	160	166	170	175	180
Y: Poids (kg)	70	76	80	86	90

(a) Quel sera le poids d'un homme de 168 cm?
(b) Quelle sera la grandeur d'un homme pesant 83 kg?

12. Un économiste donne l'estimation du prix de vente et de la demande d'un produit:

Prix ($)	Demande (tonnes)
1	9
2	7
3	6
4	3
5	1

(a) Quelle sera la demande si le prix de vente est $1,50?

(b) Quel sera le prix de vente si la demande est de 8 t?

13. Donner la liste de 3 paires de données qui auront une grande corrélation positive.

14. Donner la liste de 3 paires qui n'auront pas de corrélation.

15. Donner la liste de 3 paires qui auront une grande corrélation négative.

16. Pour les données suivantes, tracer un diagramme de dispersion, estimer r en se servant de ce diagramme puis calculer r.

(a)

X	Y
1	1
3	3
5	4
7	8

(b)

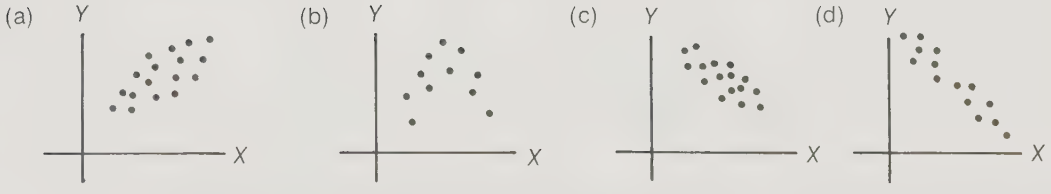
X	Y
1	2
2	5
3	3
4	5
5	2

(c)

X	Y
1	7
3	5
2	6
1	6
7	1
2	5
6	3
6	2

17. Estimer r.

(a) (b) (c) (d)



18. Pour lesquels des problèmes suivants pouvez-vous trouver l'équation de régression et trouver r?

(a) Y a-t-il une relation linéaire entre le taux d'échange de la Livre Sterling et le prix du Scotch au Canada?

(b) Quelle est la relation entre la longueur des pantalons des enfants et leur degré d'étude au primaire?

(c) Est-ce que des ouvriers qui boivent beaucoup ont un plus grand taux d'absentéisme?

(d) Peut-on prédire l'assistance à un match de baseball en matinée si la température à 08:00 est connue?

(e) Quel effet a l'âge d'une machine à laver la vaisselle sur le nombre de réparations?

(f) Peut-on prédire le résultat moyen d'un étudiant à sa graduation en examinant ses résultats de 1re année collégiale?

19. Pour les résultats suivants calculer:
(a) L'équation de régression.
(b) L'erreur de variation $\Sigma (Y - Y_c)^2$.
(c) La variation expliquée $\Sigma (Y_c - \overline{Y})^2$.
(d) La variation totale $\Sigma (Y - \overline{Y})^2$.
(e) r^2 en utilisant les résultats obtenus en (a) et (d).
(f) r.
Finalement,
(g) Tracer un diagramme d'éparpillement. Est-ce que le résultat de (f) vous surprend?

X	−1	0	1	−1	1
Y	1	0	1	−1	−1

20. Mêmes questions qu'au numéro précédent avec les données suivantes:

X	−1	0	2	−2	1
Y	1	0	−2	2	−1

21. (a) L'équation de régression des points (2; 9) (4; 6) (6; 0) est $Y_c = 14 - 2{,}25X$. Calculer (i) $\Sigma (Y - Y_c)^2$; (ii) $\Sigma (Y - \overline{Y})^2$; (iii) $\Sigma (Y - Y_c)^2 / \Sigma (Y - \overline{Y})^2$; (iv) r.
(b) Si chaque valeur Y dans (a) est multipliée par 10, la nouvelle équation de régression est $Y_c = 140 - 22{,}5X$. Calculer les nouvelles valeurs (i) (ii) (iii) (iv) définies en (a).
(c) Comparer les résultats (a) et (b).

22. Un agent d'une compagnie de publicité présente les résultats suivants à un client:

(X) Pourcentage des profits investis en publicité	(Y) Ventes (en milliers de $)
1	100
2	107
3	115
4	125
5	135

(a) Trouver la droite de régression.
(b) Trouver $(Y - Y_c)^2$.
(c) Trouver $(Y - \overline{Y})^2$.
(d) Utiliser (b) et (c) pour trouver r.
(e) Utiliser la formule pour calculer r.

23. Le nombre d'années de scolarité (X) et les salaires (Y) en milliers de $ de huit femmes choisies au hasard à Montréal est comme suit:

X	12	16	19	10	13	16	12	12
Y	10	9	20	8	6	18	9	7

(a) Tracer un diagramme d'éparpillement.
(b) Quel sera le salaire d'une femme ayant 17 années de scolarité?
(c) Trouver le coefficient de corrélation entre le nombre d'années de scolarité et le salaire.

24. Un médecin recueille les informations sur le nombre moyen de calories par jour et la perte de poids par semaine

Calories	2000	1200	1500	500	800
Perte de poids	1	3	2	5	4

(a) Trouver la corrélation entre le nombre de calories et la perte de poids.
(b) Quelle sera la perte de poids pour une diète de 700 calories par jour?

25. Le salaire des Ministres (en milliers de $) et la vente d'alcool (en billions de $) pour les années 1960 - 1964 sont comme suit:

	1960	61	62	63	64
Salaires	8,3	9,0	10,0	10,5	11,1
Vente d'alcool	7,9	8,3	8,7	9,2	9,6

(a) Calculer r.
(b) Pouvez-vous expliquer pourquoi le coefficient de corrélation est si élevé?

26. Un courtier qui transige quotidiennement un certain nombre d'actions compare leur activité à l'indice de la bourse. X représente la perte ou le gain moyen en points de ses transactions alors que Y représente l'indice de la bourse. Pour une semaine voici les résultats obtenus.

	X	Y
Lun.	+1 1/2	−1/8
Mar.	−1/4	−3/4
Mer.	+1/4	−3/4
Jeu.	+1 1/8	+1/2
Ven.	−1/2	−3/8

(a) Calculer r.
(b) L'ensemble des transactions est-il bon?

27. L'équation de régression des points (1; 2) (2; 2) (2; 3) (3; 4) est $Y_c = 0{,}75 + X$. Trouver la corrélation entre X et Y.

(a) En utilisant $r = b * \frac{s_X}{s_Y}$.

(b) En utilisant $r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\ \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$.

RÉPONSES

1.	Pentes	Ordonnée à l'origine	
(a)	3	−2	
(b)	−2	5	
(c)	pas définie	aucune	
(d)	−1/2	2	(Y=2−1/2 X)

3.

X	Y
1	2
2	2
2	3
3	4

(a) $Y_c = 1 + X$

Y_c	$(Y - Y_c)^2$
2	0
3	1
3	0
4	0
	1

(b) $Y_c = 1 + 0{,}5X$

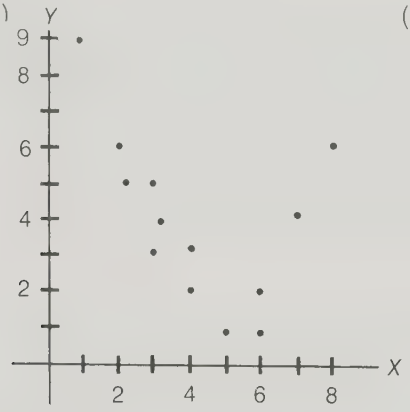
Y_c	$(Y - Y_c)^2$
1,5	0,25
2,0	0
2,0	1
2,5	2,25
	3,50

(c) $Y_c = 3/4 + X$

Y_c	$(Y - Y_c)^2$
1,75	1/16
2,75	9/16
2,75	1/16
3,75	1/16
	12/16 =0,75

(c) Donne la meilleure représentation des points 0,75 étant la plus petite valeur (c'est la droite de régression).

5. (a)



(b)

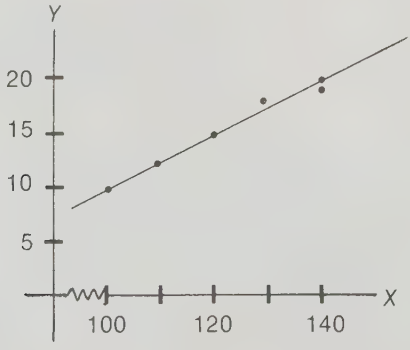


On peut trouver une droite de régression pour (b), pour (a) il serait mieux de tracer une parabole.

7. (b) $b = \dfrac{5 * 9260 - 600 * 75}{5 * 73000 - 600^2} = 0{,}26$, $a = 15 - 0{,}26 * 120 = -16{,}2$, $Y_c = -16{,}2 + 0{,}26X$.

(c) ordonnée à l'origine $a = -16{,}2$; pente $= b = 0{,}26$.

(d) $Y_c = -16{,}2 + 0{,}26 * 125 = 16{,}3$ quand $X = 125$.
$Y_c = -16{,}2 + 0{,}26 * 105 = 11{,}1$ quand $X = 105$.

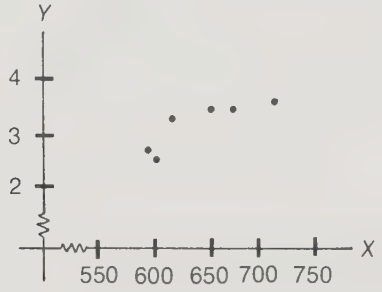


9. (a) Noter les coupures dans les échelles des axes X et Y.

(b) $\Sigma XY = 12\,863$, $\Sigma X = 3850$, $\Sigma Y = 19{,}9$, $\Sigma X^2 = 2\,482\,900$; $b = 0{,}007\,52$, $a = -1{,}51$; $Y_c = -1{,}51 + 0{,}007\,52X$.

(c) Ordonnée à l'origine $= -1{,}51$; pente $= 0{,}007\,52$.

(d) 3,75.



11. (a) L'équation de régression est: $Y_c = -93{,}2 + 1{,}02X$. Si $X = 168$ cm, alors $Y = 78{,}2$ kg.

(b) En isolant X, on obtient $X_c = 91{,}4 + 0{,}98Y$. Si $Y = 83$ kg, alors $X = 172{,}7$ cm.

17. (a) $r \approx 0{,}5$; (b) $r \approx 0$; (c) $r \approx -0{,}6$; (d) $r \approx -0{,}9$.

19. (a) $b = \dfrac{5 * 0 - 0 * 0}{5 * 5 - 0^2} = 0$; $a = 0 - 0 * \dfrac{0}{5} = 0$, $Y_c = 0$. (b) 4. (c) 0. (d) 4. (e) $\dfrac{0}{4} = 0$. (f) 0.

21. (a) (i) $\Sigma (Y - Y_c)^2 = (-0{,}5)^2 + 1^2 + (-0{,}5)^2 = 1{,}50$; (ii) $4^2 + 1^2 + (-5)^2 = 42$; (iii) 1,50/42 = 0,04; (iv) − 0,98.

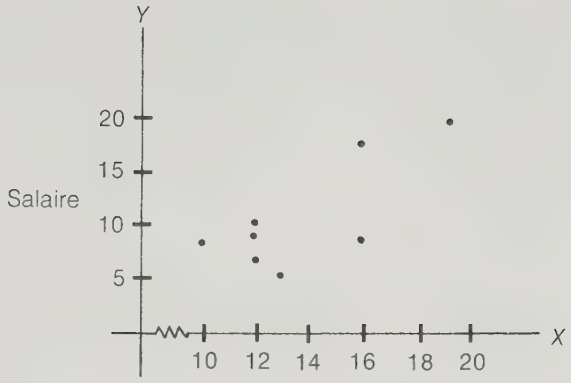
(b) (i) $(-5)^2 + 10^2 + (-5)^2 = 150$; (ii) $40^2 + 10^2 + (-50)^2 = 4200$; (iii) 150/4200 = 0,04; (iv) −0,98.

(c) Quand les valeurs de Y sont multipliées par 10, $\Sigma (Y - Y_c)^2$ et $\Sigma (Y - Y)^2$ sont multipliées par 100; leur quotient est inchangé et ainsi r ne change pas.

23. (a)

(b) $b = \dfrac{8 * 1282 - 110 * 87}{8 * 1574 - 110^2} = 1{,}394$

$a = \dfrac{87}{8} - 1{,}394 * \dfrac{110}{8} = -8{,}29$



$Y_c = -8{,}29 + 1{,}39X$; si $X = 17$, $Y_c = 15{,}3$. Elle gagnera approximativement \$15 300.

(c) $r = \dfrac{686}{\sqrt{492}\,\sqrt{1511}} = 0{,}80$

25. (a) $r = \dfrac{5 * 430{,}43 - 48{,}9 * 43{,}7}{\sqrt{25{,}54}\,\sqrt{9{,}26}} = 0{,}99$

(b) Parce que les salaires augmentent au même rythme que le coût de la vie; la consommation d'alcool augmente avec l'accroissement de la population et l'accroissement des salaires.

27. (a) $r = 1 * \dfrac{0{,}82}{0{,}96} = 0{,}85$. (b) $r = \dfrac{4 * 24 - 8 * 11}{\sqrt{4 * 18 - 8^2}\,\sqrt{4 * 33 - 11^2}} = 0{,}85$.

Chapitre 16

Inférence statistique: régression linéaire et corrélation.

On sait maintenant comment trouver un coefficient de corrélation et une équation de régression linéaire pour un échantillon de paires (X; Y). Il ne fut pas nécessaire de poser des hypothèses concernant la distribution de X ou de Y dans les populations. Nous devrons maintenant nous poser des questions concernant la relation qui existe entre les résultats X et Y d'une population et également sur la valeur prédite Y à partir d'un X donné. Il sera essentiel de poser certaines hypothèses pour les distributions des populations. Il y a une relation étroite entre la régression et la corrélation. À faire de l'inférence statistique sur la régression et la corrélation, on étudiera la différence qui existe entre les deux, de même que les différentes hypothèses sur les distributions des populations.

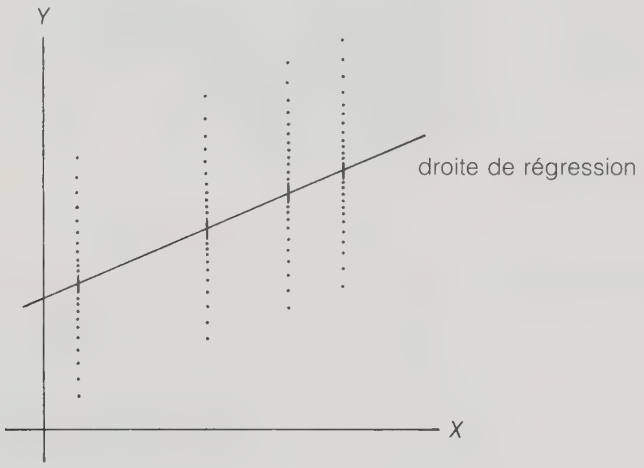
16.1 INFÉRENCES SUR LA RÉGRESSION: HYPOTHÈSES SUR LES POPULATIONS

Supposons qu'on trace un diagramme d'éparpillement pour n paires $(X; Y)$ de résultats, qu'on décide qu'ils sont approximativement en ligne droite et qu'on trouve l'équation de régression $Y_c = a + bX$. Examinons maintenant la population de laquelle les paires $(X; Y)$ sont prélevées. On peut se poser trois questions: est-ce que la relation demeure linéaire pour **toutes les paires** $(X; Y)$? Si oui, peut-on estimer la moyenne de tous les Y pour tout X de la population? Si plus tard on choisit un X dans la population, peut-on prédire la valeur correspondante de Y?

On doit faire les hypothèses suivantes sur la population.

1. Si un psychologue, qui fait une enquête, peut mesurer **tous** les résultats Y pour chaque X de la population, il s'attend d'avoir certaines erreurs d'observation, dues à l'imprécision des mesures, à des erreurs d'interprétation, et autres qui ne peuvent être contrôlées. Il peut affirmer, quand même, que la moyenne des Y pour un X donné est sur la droite $Y = A + BX$. Ceci confirme qu'il y a véritablement une relation linéaire entre les paires de résultats et que A et B provenant de l'équation de la population sont approximés par a et b dans l'équation de régression $Y_c = a + bX$.

2. On suppose que pour chaque X, les valeurs correspondantes de Y dans la population sont normalement distribuées et que toutes les distributions normales ont la même variance. Le graphique suivant permet d'illustrer les deux premières hypothèses. «On voit que pour une valeur de X, l'éparpillement des Y sur la verticale suit une loi normale.»



3. En général, les résultats X utilisés pour trouver la droite de régression ne sont pas choisis au hasard dans la population, d'où l'inutilité de formuler des hypothèses sur leur distribution. Les valeurs particulières de X sont souvent choisies par l'expérimentateur. Un exemple: un chimiste qui étudie les points d'ébullition de solutions peut ajouter 5 g de la substance à 100 cm³ d'eau puis 10; 15; 20.....g. Quand on étudie la précision d'une prédiction Y pour un X donné, on constate que celle-ci peut être plus grande pour un X pris près de la moyenne $\overline{X}$. Ainsi on ne voudra pas trop s'éloigner de $\overline{X}$, sauf si l'échantillon est suffisamment grand. Il ne faut pas confondre cette restriction pratique du choix de X avec une hypothèse concernant la distribution des X dans la population.

16.2 VÉRIFIER $H_0 : B = 0$

Existe-t-il une équation de régression linéaire $Y = A + BX$ pour la population? Si ce n'est pas le cas, le meilleur choix pour un X quelconque est $\overline{Y}$ et dans ce cas la meilleure équation de la population est $Y = \overline{Y}$. En posant comme hypothèse nulle H_0: $B = 0$, on obtient cette équation. Si cette hypothèse est rejetée on peut alors supposer que l'équation de régression de l'échantillon peut être utilisée pour prédire une valeur Y dans la population et ceci avec une précision que l'on peut déterminer.

Quelle est alors l'hypothèse rivale? Selon le type d'expérience il peut arriver que l'on sache que la pente est positive (ou négative); dans ce cas on utilise un test unilatéral en posant $H_1 : B > 0$ (ou $B < 0$). Dans les autres cas, on pose $H_1 : B \neq 0$ et on utilise un test bilatéral.

Avant de vérifier l'hypothèse nulle, nous devons, comme auparavant, examiner la distribution d'échantillonnage des pentes. Supposons que, d'une population qui vérifie toutes les hypothèses de la section précédente, nous prenons tous les échantillons de taille n et que nous calculons $\frac{b-B}{s_b}$ pour chacune, où

$$s_b = \frac{s_e}{\sqrt{\Sigma(X-\bar{X})^2}}$$

avec $s_e = \sqrt{\frac{\Sigma(Y-Y_c)^2}{n-2}}$ qui est l'erreur standard de l'estimation telle que définie à la section 15.7. Cette distribution d'échantillonnage est une distribution «t» avec $n-2$ degrés de liberté.

Heureusement, on peut développer les formules de $\Sigma(Y-Y_c)^2$ et $\Sigma(X-\bar{X})^2$

$$\Sigma(Y-Y_c)^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY$$

$$\Sigma(X-\bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/n$$

(le dernier développement est le numérateur de la variance). Après quelques transformations algébriques, on obtient la formule suivante pour s_b.

$$s_b = \sqrt{\frac{n(\Sigma Y^2 - a\Sigma Y - b\Sigma XY)}{(n-2)[n\Sigma X^2 - (\Sigma X)^2]}}$$

N.B.: $n\Sigma X^2 - (\Sigma X)^2$ est le dénominateur de b.

Pour vérifier $H_0 : B = 0$, il suffit de calculer $t = \frac{b-B}{s_b}$ et de comparer cette valeur avec la valeur critique de t pour α et $n-2$ degrés de liberté en utilisant un test unilatéral ou bilatéral selon l'hypothèse H_1.

Exemple 1

Au niveau de signification de 0,05, vérifier $H_0 : B = 0$ et $H_1 : B \neq 0$ pour les paires $(X; Y)$ données ci-dessous.

X^2	X	Y	XY	Y^2
4	2	0	0	0
9	3	3	9	9
16	4	4	16	16
25	5	4	20	16
36	6	6	36	36
49	7	11	77	121
139	27	28	158	198

$$b = \frac{6 * 158 - 27 * 28}{6 * 139 - 27^2} = \frac{192}{105} = 1{,}8286$$

$$a = \frac{28}{6} - 1{,}8286 * \frac{27}{6} = -3{,}5620$$

$$\Sigma Y^2 - a\Sigma Y - b\Sigma XY = 198 - (-3{,}5620) * 28 - 1{,}8286 * 158 = 8{,}82$$

N.B.: (Pour éviter la perte de précision, on évalue a et b à quatre décimales près.) «Il est à remarquer que l'erreur de variation $\Sigma (Y - Y_c)^2$ calculée à la section 15.7, exemple 1, pour les mêmes paires, donne le même résultat que la formule développée utilisée ci-dessus.»

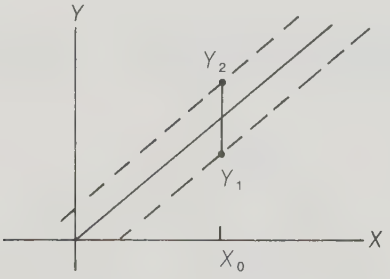
$$s_b = \sqrt{\frac{6 * 8{,}82}{4 * 205}} = 0{,}254$$

$$t = \frac{1{,}83 - 0}{0{,}254} = 7{,}20$$

Pour $\alpha = 0{,}05$; n = 6; $D = 4$; les valeurs critiques de t pour un test bilatéral sont $\pm$ 2,78. Comme $7{,}20 > 2{,}78$, H_0 est rejetée et H_1 acceptée; la pente de la droite de régression de la population n'est pas nulle.

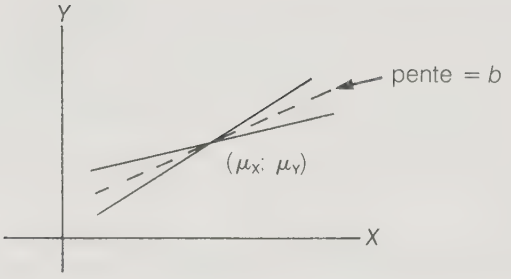
16.3 INTERVALLES DE CONFIANCE ET DE PRÉDICTION POUR L'ÉQUATION DE RÉGRESSION

À partir d'un échantillon de n paires (X; Y), on trouve l'équation de régression $Y_c = a + bX$ puis on vérifie $H_0 : B = 0$. Si elle est rejetée, on peut affirmer qu'il y a une relation linéaire entre toutes les paires (X; Y) de la population. On peut alors prédire le résultat Y moyen pour un X pris dans la population, mais quelle est la précision de cette estimation? Notre intuition nous suggère qu'il doit exister deux droites parallèles à la droite de régression nous permettant d'affirmer: «qu'il y a un intervalle de confiance au niveau de 95% pour un Y moyen correspondant à X_0; cet intervalle étant $Y_c \pm$ (une constante)».

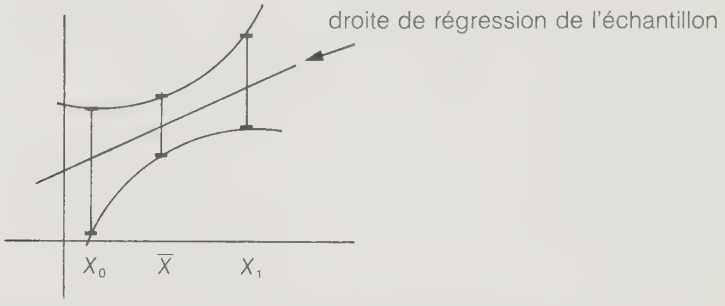


Ceci implique que cet intervalle de confiance est le même pour toutes valeurs de X. Malheureusement, ce n'est pas le cas: cet intervalle est plus étroit pour des valeurs de X pris près de la moyenne $\overline{X}$. Examinons pourquoi.

L'équation de régression de l'échantillon $Y_c = a + bX$ passe toujours par le point ($\overline{X}$; $\overline{Y}$): en effet, $a = \overline{Y} - b\overline{X}$ (section 15.5 exemple 1) et $Y_c = a + bX = \overline{Y} - b\overline{X} + bX$ si $X = \overline{X}$ on a $Y_c = \overline{Y}$. De même on doit admettre que la droite de régression de la population passera par le point (μ_x; μ_y). La pente de la droite de régression de la population n'est pas nécessairement la même que celle de l'échantillon; la droite représentant la population est dans une région qui a le point (μ_x; μ_y) comme pivot.



Deuxièmement, on peut déterminer des intervalles de confiance pour μ_x et μ_y, mais on ne connaît pas exactement les valeurs de μ_x et μ_y, ce qui rend difficile le tracé de la droite de régression de la population. Il résulte que l'intervalle de confiance pour ce résultat Y moyen est plus étroit quand X_0 est près de $\overline{X}$ et s'élargit quand X_0 s'éloigne de $\overline{X}$.



Supposons que X_0 est une valeur particulière de X et que pour cette valeur X_0 les résultats Y sont normalement distribués dans la population. On peut se poser deux questions:

1. Quelle est la valeur moyenne des résultats Y correspondants à X_0?

2. Si $X = X_0$, peut-on trouver des limites entre lesquelles Y se trouvera avec une certaine probabilité?

Considérons la première question. Il semble plausible d'utiliser la droite de régression de l'échantillon et de choisir $a + bX_0$ comme une estimation de la valeur moyenne Y pour $X = X_0$. Ici nous avons à faire face au même problème que nous avons eu dans la recherche de μ à partir de $\overline{X}$. Une estimation ponctuelle est sans signification si on n'a aucune idée de sa précision. C'est pour cette raison qu'on obtient une estimation par intervalle (revoir la discussion de la section 9.6). Soit μ_{Y/X_0} la moyenne de la valeur Y correspondante à X_0. On trouve un intervalle de confiance pour μ_{Y/X_0} en utilisant une distribution «t» avec $n - 2$ degrés de liberté, ainsi:

$$t = \frac{Y_c - \mu_{Y/X_0}}{s_{Y/X_0}}$$

où

$$s_{Y/X_0} = s_e \sqrt{\frac{1}{n} + \frac{n(X_0 - \overline{X})^2}{n\Sigma X^2 - (\Sigma X)^2}}$$

s_e représente l'éparpillement des valeurs Y autour de la droite de régression pour toutes valeurs de X (voir hypothèse 2 section 16.1); la racine carrée représente l'augmentation de l'éparpillement des valeurs Y au fur et à mesure que X_0 s'éloigne de $\overline{X}$ (il est à remarquer que ce facteur prend sa plus petite valeur quand $X_0 = \overline{X}$, car le dernier terme sous le radical devient nul).

Exemple 1

Pour les résultats donnés ci-dessous, l'équation de régression est:

$$Y_c = -3{,}56 + 1{,}83X,\ s_e = \sqrt{\frac{8{,}82}{4}} = 1{,}485$$

et $n\Sigma X^2 - (\Sigma X)^2 = 205$. (Voir l'exemple 1 section 15.7.)

Trouver un intervalle de confiance au niveau de 95% pour μ_{Y/X_0} quand X_0 est (a) 2,0 (b) 4,5.

X	Y
2	0
3	3
4	4
5	4
6	6
7	11

(a) Au niveau de 95%, avec $n = 6$ et $D = 4$, $t = \pm 2{,}78$; $\overline{X} = 4{,}5$.

$$s_{Y/X_0} = 1{,}485\sqrt{\frac{1}{6} + \frac{6(2-4{,}5)^2}{205}} = 0{,}878$$

Quand $X = 2$, $Y_c = -3{,}56 + 1{,}83 * 2 = 0{,}100$.

$$\pm 2{,}78 = \frac{0{,}100 - \mu_{Y/X_0}}{0{,}878}$$

$$\mu_{Y/X_0} = -0{,}10 \pm 2{,}44 = -2{,}3 \text{ ou } 2{,}5$$

Quand $X = 2$, l'intervalle de confiance au niveau de 95% pour la valeur moyenne des Y est $[-2{,}3;\ 2{,}5]$.

(b) $t = \pm 2{,}78$; $\overline{X} = 4{,}5$; $s_{Y/X_0} = 1{,}485 * \sqrt{1/6} = 0{,}606$. Quand $X = 4{,}5$; $Y_c = -3{,}56 + 1{,}83 * 4{,}5 = 4{,}68$.

Quand $X = 4{,}5$, l'intervalle de confiance pour le Y moyen au niveau de 95% est $[3{,}0;\ 6{,}4]$. Il est à noter que la largeur de l'intervalle de confiance quand $X = \overline{X}$ est $6{,}4 - 3{,}0 = 3{,}4$ alors qu'elle est de $2{,}5 - (-2{,}3) = 4{,}8$ quand $X = 2{,}5$.

Exemple 2

Trouver pour les paires $(X; Y)$ ci-dessous:

(a) l'équation de régression;

(b) un intervalle de confiance pour μ_{Y/X_0} au niveau de 90% quand X_0 est (i) 1; (ii) 2; (iii) 2,5; (iv) 3; (v) 4.

(c) tracer la droite de régression de l'échantillon et les limites de confiances pour μ_{Y/X_0}.

X^2	X	Y	XY	Y^2	Y_c
1	1	0	0	0	0,3
4	2	2	4	4	1,1
9	3	1	3	1	1,9
16	4	3	12	9	2,7
30	10	6	19	14	

Réponse

(a) $$b = \frac{4 * 19 - 10 * 6}{4 * 30 - 10^2} = \frac{16}{20} = 0{,}80$$

$$a = \frac{6}{4} - 0{,}8 * \frac{10}{4} = -0{,}50$$

$$Y_c = -0{,}50 + 0{,}80X$$

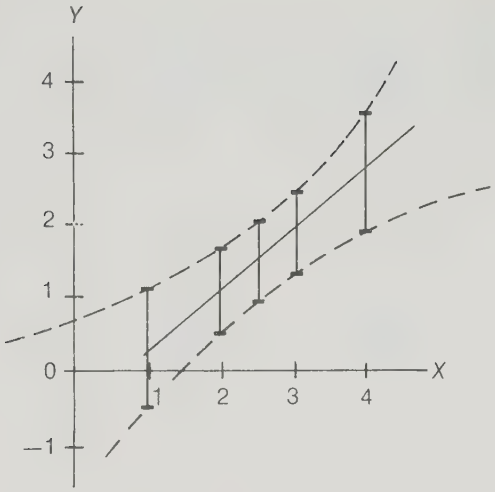
(b) Au niveau de 90% avec $n = 4$; $D = 2$; $t = \pm 2{,}92$.

$$\mu_{Y/X_0} = Y_c \pm 2{,}92\, s_{Y/X_0}$$

$$s_{Y/X_0} = \sqrt{\frac{14 + 0{,}5 * 6 + 0{,}80 * 19}{2}} \quad \sqrt{\frac{1}{4} + \frac{4(X_0 - 2{,}5)^2}{20}}$$

$$= 0{,}95\sqrt{0{,}25 + 0{,}20\,(X_0 - 2{,}5)^2}$$

	X_0	Y_c	s_{Y/X_0}	Intervalle de confiance pour μ_{Y/X_0}	Largeur de l'intervalle
(i)	1	0,3	0,79	−2,0 à 2,6	4,6
(ii)	2	1,1	0,52	−0,4 à 2,6	3,0
(iii)	2,5	1,5	0,48	0,1 à 2,9	2,8
(iv)	3	1,9	0,52	0,4 à 3,4	3,0
(v)	4	2,7	0,79	0,4 à 5,0	4,6



Revenons maintenant à la deuxième question posée au début de cette section. Si $X = X_0$, peut-on trouver des limites entre lesquelles Y se trouvera avec une certaine probabilité? Cette situation est presque identique à celle de la première question. Il n'est pas nécessaire d'estimer les paramètres de la population; on veut prédire la valeur de Y avec une certaine probabilité.

La notation Y_p indique la valeur probable (prédite) de Y pour une valeur particulière X_0. L'intervalle de prédiction pour Y_p est trouvé en utilisant une distribution «t».

$$t = \frac{Y_c - Y_p}{s_{Y_p}}$$

Avec $n - 2$ degrés de liberté, où

$$s_{Y_p} = s_e \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X - \bar{X})^2}}$$

C'est une formule qui ressemble beaucoup à celle de s_{Y/X_0} elle diffère seulement par le 1 qui est ajouté sous le radical. En ajoutant 1 sous le radical, on augmente la valeur de s_{Y/X_0} et par conséquent, l'intervalle de prédiction pour Y_p est plus large que l'intervalle de confiance de μ_{Y/X_0} Ce résultat n'est pas surprenant, car on doit s'attendre à avoir plus de variabilité avec une valeur prise dans une distribution normale que sur la moyenne de plusieurs valeurs.

Exemple 3

Pour les données de l'exemple 1, prédire l'intervalle dans lequel le résultat Y se trouvera avec une certitude de 95% si (a) $X_0 = 2$, (b) $X_0 = 4{,}5$.

(a) Comme précédemment $s_e = 1{,}485$; $\bar{X} = 4{,}5$; $t = \pm 2{,}78$ et $Y_c = 0{,}100$ quand $X = 2$

$$s_{Y_p} = 1{,}485\sqrt{1 + \frac{1}{6} + \frac{6(2 - 4{,}5)^2}{205}} = 1{,}72$$

$$\pm 2{,}78 = \frac{0{,}100 - Y_p}{1{,}72}$$

$$Y_p = 0{,}100 \pm 4{,}78 = -4{,}7 \text{ ou } 4{,}9$$

L'intervalle de prédiction au niveau de 95% pour Y quand $X = 2$ est $[-4{,}7;\ 4{,}9]$

(b) Quand $X = 4{,}5$

$$s_{Y_p} = 1{,}485\sqrt{1 + \frac{1}{6}} = 1{,}60 \text{ et } Y_c = 4{,}68$$

$$\pm 2{,}78 = \frac{4{,}68 - Y_p}{1{,}60}$$

$$Y_p = 4{,}68 \pm 4{,}45 = 0{,}2 \text{ ou } 9{,}1$$

L'intervalle de prédiction au niveau de 95% pour Y quand $X = 4{,}5$ $(= \overline{X})$ est $[0{,}2;\ 9{,}1]$. Comparons les intervalles de confiance et de prédiction pour Y.

X_0	Int. de confiance pour μ_{Y/X_0}	Largeur	Int. de prédiction pour Y_p	Largeur
2	$[-2{,}3;\ 2{,}5]$	4,8	$[-4{,}7;\ 4{,}9]$	9,6
4,5 $(= \overline{X})$	$[3{,}0;\ 6{,}4]$	3,4	$[0{,}2;\ 9{,}1]$	8,9

Exemple 4

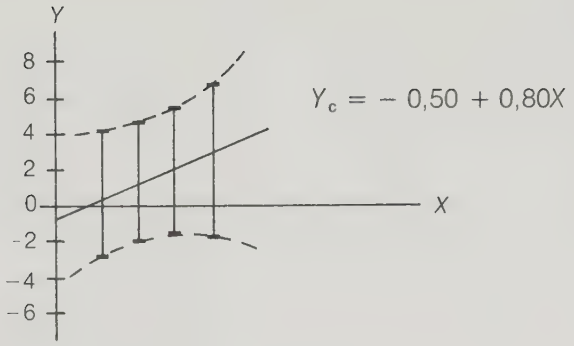
Pour les données de l'exemple 2, (a) trouver un intervalle de prédiction au niveau de 90% pour Y quand X_0 est (i) 1 (ii) 2 (iii) 2,5 (iv) 3 (v) 4; (b) tracer la droite de régression et l'intervalle de prédiction en plaçant les limites de l'intervalle pour les différentes valeurs de X_0.

Réponses

(a) Comme à l'exemple 2 $t = \pm 2{,}92$ et $s_e = 0{,}95$. $Y_p = Y_c \pm 2{,}92\, s_{Y_p}$, et $s_{Y_p} = \sqrt{1 + \frac{1}{4} + \frac{4(X_0 - 2{,}5)^2}{20}}$

	X_0	s_{Y_p}	Y_c	Intervalle de prédiction pour Y_p	Largeur
(i)	1	1,30	0,3	− 3,5 à 4,1	7,6
(ii)	2	1,14	1,1	− 2,2 à 4,4	6,6
(iii)	2,5	1,12	1,5	− 1,8 à 4,8	6,6
(iv)	3	1,14	1,9	− 1,4 à 5,2	6,6
(v)	4	1,30	2,7	− 1,1 à 6,5	7,6

(b)



L'intervalle de confiance pour μ_{Y/X_0} et l'intervalle de prédiction pour Y_p sont tous les deux très larges dans les exemples précédents. Comment pouvons-nous les retrécir? On peut y parvenir premièrement en augmentant la taille de l'échantillon puisque n et $n - 2$ apparaissent au dénominateur. On pourrait également augmenter l'éparpillement des résultats X puisque $\Sigma(X - \overline{X})^2$ apparaît au dénominateur. Si on a un échantillon aléatoire, on ne peut intervenir sur les facteurs précédents. Lors de l'étude d'un phénomène où intervient une régression linéaire, on fait souvent une expérience où les valeurs de X sont préalablement fixées (le chimiste qui ajoute 100 cm³ d'eau à sa solution, 5 cm³ à la fois; le physicien qui augmente la tention électrique de 100 volts par intervalles de 10 volts). Il ne faut pas être trop ambitieux sur l'écart des valeurs X (l'éparpillement), sauf si l'échantillon est passablement grand.

16.4 LA DISTRIBUTION D'ÉCHANTILLONNAGE DE *r*

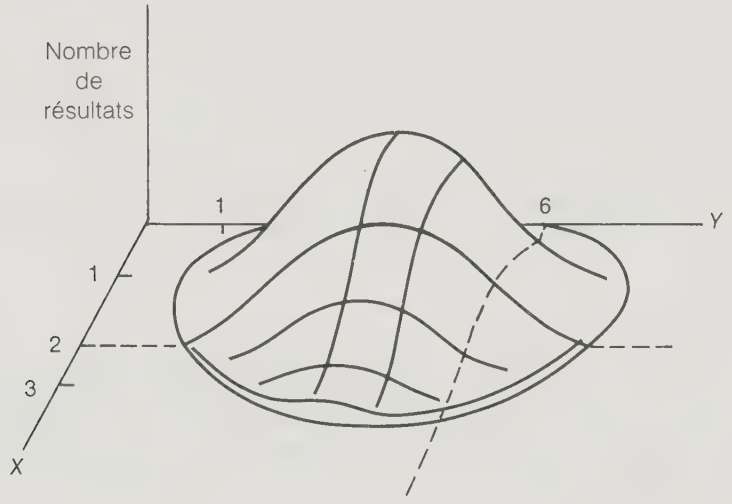
Supposons qu'une population est formée des paires suivantes:

X:	1	1	3	5	5
Y:	1	5	3	1	5

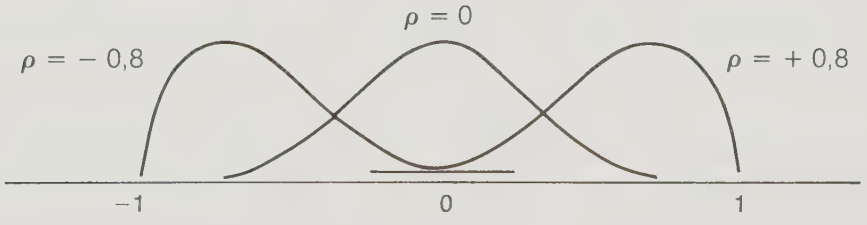
le coefficient de corrélation de la population est 0. Prenons un échantillon de taille 3, en prenant la 1re, la 3e et la 5e paire. Pour cet échantillon le coefficient de corrélation est 1. Cet exemple montre la nécessité de vérifier une hypothèse sur le coefficient de corrélation de la population après que le coefficient r d'un échantillon a été déterminé. Nous devons examiner la distribution d'échantillonnage des r.

Appelons ρ (prononcer **rhô**) le coefficient de corrélation de la population. Si tous les échantillons de taille n sont pris dans la population et que le coefficient de corrélation r est calculé pour chaque échantillon, quelle sera la distribution des r? On ne peut répondre à cette question avant d'affirmer que les **résultats X et Y sont tous les deux normalement distribués.** Ceci veut dire en particulier que pour chaque X, les valeurs correspondantes de Y sont normalement distribuées et que pour chaque Y, les valeurs correspondantes de X le sont également. La figure ci-dessous représente en trois dimensions l'aspect de la distribution (la ligne pointillée qui devient pleine montre la distribution normale des Y pour $X = 2$; la ligne pointillée montre la distribution normale des X pour $Y = 6$).

Une telle distribution s'appelle une **distribution binormale.**



Malheureusement, la distribution de r ne dépend pas seulement de l'échantillon de taille n, mais également de la corrélation de la population. Quand $\rho = 0$, la distribution de r est symétrique par rapport à 0. La distribution est biaisée si ρ est près de + 1 ou − 1. Si $\rho = 0{,}80$, par exemple, les échantillons pris dans la population auront des valeurs différentes de r, mais aucune d'elles ne sera supérieure à + 1 et l'échantillon des r ne peut être supérieur à 0,20 au-dessus du coefficient ρ de la population mais peut être bien en-dessous. La distribution de r est biaisée négativement si ρ est près de 1 et positivement si ρ est près de − 1.



Distribution d'échantillonnage de r

Pour cette raison, il sera nécessaire d'utiliser d'autres méthodes pour vérifier l'hypothèse nulle $\rho = A$ *selon que* $A = 0$ ou $A \neq 0$.

16.5 VÉRIFIER $H_0 : \rho = 0$

Supposons que r est calculé pour chaque échantillon de taille n dans une population binormale avec $\rho = 0$. La distribution des résultats r n'est ni normale, ni une distribution

«t». Calculons pour chaque échantillon la valeur $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. Ces valeurs forment une distribution «t» avec $n - 2$ degrés de liberté.

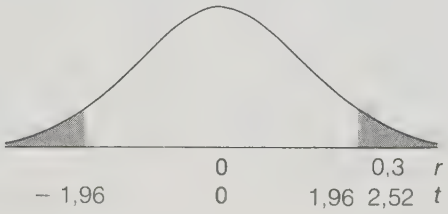
> Vérifier $H_0 : \rho = 0$ en calculant $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ pour l'échantillon, puis comparer avec la valeur critique de t pour $n - 2$ degrés de liberté avec un α donné.
> Utiliser un test unilatéral si $H_1 : \rho < 0$ ou $H_1 : \rho > 0$.
> Utiliser un test bilatéral si $H_1 : \rho \neq 0$.

Exemple 1

Une étude portant sur 66 étudiants choisis au hasard dans le collège Rosemont montre qu'il y a une corrélation de 0,3 entre les résultats des cours d'arts et celui de mathématiques. Existe-t-il une corrélation significative entre les résultats de ces deux cours pour tous les étudiants du collège ($\alpha = 0{,}05$)? Supposer que les résultats de tous les étudiants dans ces deux cours sont normalement distribués. $H_0 : \rho = 0$; $H_1 : \rho \neq 0$; $\alpha = 0{,}05$; $D = 66 - 2 = 64$, test bilatéral, donc $t = \pm 1{,}96$ (utiliser la table z puisque $t > 40$); $r = 0{,}3$; $n = 66$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t = \frac{0{,}3\sqrt{66-2}}{\sqrt{1-0{,}3^2}} = \frac{2{,}4}{\sqrt{0{,}91}} = 2{,}52$$



Comme $2{,}52 > 1{,}96$; H_0 est rejetée; il y a une corrélation significative entre les résultats des deux cours.

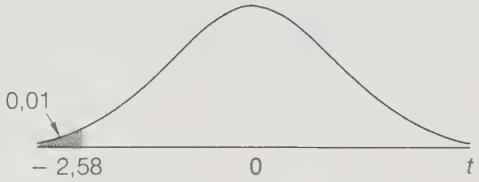
N.B.: Si l'une ou l'autre ou les deux variables sont approximativement normales, il faudra augmenter la taille de l'échantillon. Si l'une ou l'autre **n'est pas** normalement distribuée, le test ne s'applique pas.

Exemple 2

Un échantillon de 18 enfants choisis au hasard dans une classe de 3e année de l'école primaire Christ-Roy montre une corrélation de $-0{,}4$ entre le revenu familial et le nombre d'heures d'écoute de télévision par semaine. Au niveau de signification de 0,01, existe-t-il une corrélation significative pour tous les enfants de 3e année de l'école Christ-Roy?
$H_0 : \rho = 0$; $H_1 : \rho < 0$ test unilatéral $\alpha = 0{,}01$; $r = -0{,}4$; $n = 18$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0{,}4\sqrt{18-2}}{\sqrt{1-0{,}4^2}} = \frac{-1{,}6}{0{,}917} = -1{,}74$$

La valeur critique de t pour $n - 2 = 16$ degrés de liberté avec $\alpha = 0{,}01$ est $-2{,}58$.



Comme $-1{,}74 > -2{,}58$, H_0 est acceptée; il n'y a pas de corrélation entre le revenu familial et le nombre d'heures d'écoute de télévision.

16.6 VÉRIFIER H_0: $\rho = A$, $A \neq 0$

Si $\rho \neq 0$, la distribution d'échantillonnage des coefficients de corrélation r est biaisée. Un statisticien R.A. Fisher a découvert une relation permettant de transformer les résultats r en de nouveaux résultats Z qui sont distribués de façon approximativement normale pour une grande valeur de r. (Ne pas confondre ces résultats Z avec les résultats standards z.) Heureusement, il n'est pas nécessaire de calculer cette transformation; la table 8 appendice C donne les valeurs correspondantes de r et Z.

Exemple 1

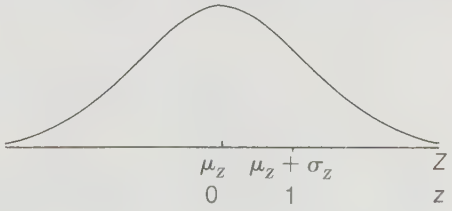
Quelle est la valeur de Z pour $r = 0{,}15$?
Pour $r = -0{,}70$?
Pour $r = 0{,}15$ alors $Z = 0{,}151$.
Pour $r = -0{,}70$ alors $Z = -0{,}867$.

Si le résultat Z est trouvé pour chaque paire d'échantillons de taille n pris dans la population, la distribution est approximativement normale pour $n > 30$. Sa moyenne μ_Z est la valeur de Z correspondante à ρ et son écart-type est

$$\sigma_Z = \frac{1}{\sqrt{n - 3}}$$

L'hypothèse est alors vérifiée par un test z d'une distribution normale.

$$z = \frac{Z - \mu_Z}{\sigma_Z}$$



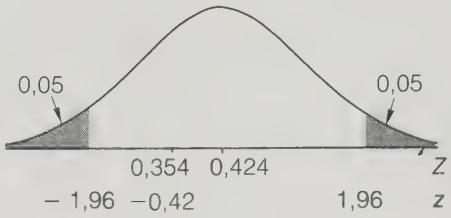
Exemple 2

Une étude antérieure montre qu'il y a une corrélation de 0,40 entre les résultats du cours d'économie et celui d'histoire à l'université Laval. Cette

année un échantillon aléatoire de 39 étudiants suivant les deux cours montre une corrélation de 0,34. La corrélation de cette année est-elle différente de façon significative au niveau de 0,05? H_0 : $\rho = 0{,}40$; H_1 : $\rho \neq 0{,}40$; $\alpha = 0{,}05$; si $r = 0{,}34$ alors $Z = 0{,}354$; si $\rho = 0{,}40$ alors $\mu_Z = 0{,}424$; la transformation de ρ à μ_Z est du même type que celle de r à Z.

$$\sigma_Z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{36}} = 0{,}167$$

$$z = \frac{Z - \mu_Z}{\sigma_Z} = \frac{0{,}354 - 0{,}424}{0{,}167} = -0{,}42$$



Pour un test bilatéral avec $\alpha = 0{,}05$; $z = \pm 1{,}96$; $-0{,}42$ est dans la région d'acceptation; la corrélation demeure $\rho = 0{,}40$ pour tous les étudiants qui suivent ces deux cours, la différence n'est pas significative.

Exemple 3

Trouver un intervalle de confiance pour ρ au niveau de 95% si un échantillon de 103 paires $(X; Y)$ a une corrélation de 0,58. Au niveau de 95%, $z = \pm 1{,}96$; si $r = 0{,}58$, alors $Z = 0{,}663$; $n = 103$ donc $\sigma_Z = 1/10 = 0{,}10$;

$$z = \frac{Z - \mu_Z}{\sigma_Z} \text{ ou } 1{,}96 = \frac{0{,}663 - \mu_Z}{0{,}1}; \text{ donc } \mu_Z = 0{,}467 \text{ ou } 0{,}859.$$

En utilisant la table 8 appendice C, on cherche dans la colonne Z les valeurs 0,467 et 0,859 et on trouve les valeurs correspondantes de ρ dans la colonne r.

Si $\mu_Z = 0{,}467$ alors $\rho = 0{,}44$.

Si $\mu_Z = 0{,}859$ alors $\rho = 0{,}70$.

L'intervalle de confiance au niveau de 95% pour ρ est [0,44 ; 0,70].

Comparer ce résultat à celui obtenu avec un échantillon plus petit pour la même corrélation.

Exemple 4

Trouver un intervalle de confiance pour ρ au niveau de 95% si $r = 0{,}58$ et $n = 39$. Comme à l'exemple précédent, $Z = \pm 1{,}96$ et $Z = 0{,}663$.

$$\sigma_Z = \frac{1}{\sqrt{39-3}} = \frac{1}{\sqrt{36}} = 0{,}167$$

$$z = \frac{Z - \mu_Z}{\sigma_Z}; \pm 1{,}96 = \frac{0{,}663 - \mu_z}{0{,}167}; \mu_z = 0{,}336 \text{ ou } 0{,}990$$

Si $\mu_Z = 0{,}336$ alors $\rho = 0{,}32$. Si $\mu_Z = 0{,}990$ alors $\rho = 0{,}76$.

L'intervalle de confiance pour ρ au niveau de 95% est [0,32; 0,76]. Comme on pouvait s'y attendre, une plus petite valeur de n donne un intervalle plus large.

16.7 VOCABULAIRE ET SYMBOLES

$Y = A + BX$
Distribution des pentes
Y_p
σ_{Y/X_0}
ρ
Distribution des r
Distribution binormale
Distribution Z
μ_Z, σ_Z
Transformation de Fisher de r à Z

16.8 EXERCICES

1. L'équation de régression des paires $(X; Y)$ ci-dessous est $Y_c = 5{,}22 - 0{,}59X$. Au niveau de signification de 0,05, trouver si $B = 0$ ou non.

X	1	2	4	8
Y	6	3	2	1

2. Après la collation des grades, les résultats d'un test d'aptitudes de six étudiants sont comparés à leur résultat de chimie.

(X) T.A.	600	720	640	590	620	680
(Y) chimie	2,8	3,8	3,6	2,9	3,2	3,6

L'équation de régression est $Y_c = -1{,}51 + 0{,}0075X$. Si $\alpha = 0{,}01$, vérifier $H_0 : B = 0$; $H_1 : B \neq 0$.

3.

X	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3
Y	−1	1	0	2	4	3	1	2	6	4	5	6	7	8	6	8

(a) Vérifier $H_0 : B = 0$; $H_1 : B \neq 0$ ($\alpha = 0{,}05$).
(b) Vérifier $H_0 : B = 2{,}5$; $H_1 : B \neq 2{,}5$ ($\alpha = 0{,}05$).
(c) Trouver un intervalle de confiance pour B au niveau de 95%.

4. Se référant au problème 23 de la section 15.12, l'équation de régression est $Y_c = -8{,}29 + 1{,}39X$; $\Sigma XY = 1\,282$; $\Sigma X = 110$; $\Sigma X^2 = 1\,574$; $\Sigma Y = 87$; $\Sigma Y^2 = 1\,135$.
(a) Vérifier, avec $\alpha = 0{,}05$, $H_0 : B = 0$; $H_1 : B \neq 0$.
(b) Trouver un intervalle de confiance pour B au niveau de 95%.

5.

X	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6
Y	8	6	6	4	4	2	2	0	0	−2	−2	−4	−4

$\Sigma X = 0$, $\Sigma X^2 = 182$, $\Sigma Y = 20$, $\Sigma Y^2 = 216$, $\Sigma XY = -182$, $n = 13$.
(a) Trouver l'équation de la droite de régression.
(b) Vérifier $H_0 : B = 0$; $H_1 : B < 0$ ($\alpha = 0{,}01$).
(c) Vérifier $H_0 : B = -1{,}1$; $H_1 : B \neq -1{,}1$ ($\alpha = 0{,}01$).
(d) Trouver un intervalle de confiance pour B au niveau de 99%.

6. Se référer au problème 3. Former un nouvel ensemble de 64 paires en répétant 4 fois chacune des 16 paires du problème 3. Trouver un intervalle de confiance pour B au niveau de 95% et comparer le résultat à celui du problème 3(c).
Noter: ΣX, ΣY, ΣX^2, ΣY^2, ΣXY et n sont tous multipliés par 4.

7. Supposons qu'un nouvel ensemble de 1 600 paires est formé en répétant 100 fois chacune des 16 paires du problème 3. Trouver un intervalle de confiance pour B au niveau de 95% et comparer votre réponse à celle du problème 3 (c) et à celle du problème 6.

8. Se référant au problème 4, trouver (a) un intervalle de confiance au niveau de 95% pour la moyenne de tous les Y dans la population si (i) $X = 0{,}0$ (ii) $X = 1{,}0$ (iii) $X = 1{,}5$. (b) un intervalle de prédiction pour Y au niveau de 95% si (i) $X = 0$ (ii) $X = 1$ (iii) $X = 1{,}5$.

9. Un sociologue étudie le comportement social (Y) de 80 enfants avec un test. Il croit que le nombre d'enfants dans la famille (X) peut expliquer le résultat du test. Il obtient les résultats suivants $\Sigma X = 200$; $\Sigma X^2 = 600$; $\Sigma Y = 440$; $\Sigma Y^2 = 4\,000$; $\Sigma XY = 1\,300$; $a = 0{,}500$; $b = 2{,}00$.

(a) Trouver un intervalle de confiance au niveau de 95% pour le résultat moyen du test s'il y a 2 enfants dans la famille.

(b) Un autre enfant se présente; il a un frère et pas de soeur. Quel résultat peut prédire le sociologue avec une certitude de 95%?

10. Un homme doit classer des pommes en 2, 3 ou 4 catégories. Le temps pris pour le classement dépend du nombre de catégories comme suit:

Temps (secondes)	3;1;2;1;1	2;3;2;1;2	2;3;3;1;4	3;4;4;5;2
Nombre de catégories	2	3	4	5

(a) Quel temps peut-on prédire pour six catégories avec une probabilité de 0,90?

(b) Pouvez-vous expliquer pourquoi il ne serait pas prudent d'extrapoler pour une prédiction d'une valeur de X qui serait en dehors de l'étendue de l'échantillon?

11. Cinquante et un élèves du niveau primaire passent un test de lecture et la taille de chaque élève est mesurée. La valeur calculée de r entre les résultats du test et la grandeur des élèves est 0,80.

(a) Vérifier l'hypothèse qu'il n'y a pas de corrélation entre le résultat du test de lecture et la taille de tous les élèves du primaire ($\alpha = 0{,}05$).

(b) Vérifier l'hypothèse que la corrélation est au moins 0,85 pour tous les étudiants de niveau primaire ($\alpha = 0{,}05$).

(c) Pouvez-vous expliquer cette grande corrélation?

(d) Pouvez-vous expliquer pourquoi les vérifications suggérées en (a) et (b) ne doivent pas être utilisées?

12. Vérifier $H_0: \rho = 0$; $H_1: \rho \neq 0$ pour les données du problèmes 3 ($\alpha = 0{,}01$).

13. Vérifier $H_0: \rho = 0{,}90$; $H_1: \rho > 0{,}90$ pour les données du problème 3 ($\alpha = 0{,}05$).

14. Vérifier $H_0: \rho = 0{,}85$; $H_1: \rho < 0{,}85$. Voir les données du problème 3 ($\alpha = 0{,}05$).

15. La corrélation entre deux échantillons est 0,30. Vérifier $H_0: \rho = 0$; $H_1: \rho \neq 0$. Si (a) $n = 11$ (b) $n = 102$ ($\alpha = 0{,}05$).

16. La corrélation entre deux échantillons est 0,30. Vérifier $H_0: \rho = 0{,}50$; $H_1: \rho < 0{,}50$. Si (a) $n = 11$ (b) $n = 102$ ($\alpha = 0{,}05$).

17. La corrélation entre l'âge et la pression sanguine pour un échantillon de 228 hommes choisis au hasard est 0,2. Trouver au niveau de 95% un intervalle de confiance pour la corrélation entre l'âge et la pression sanguine de tous les hommes.

18. La corrélation entre la superficie et la population de 10 comtés du Québec choisis au hasard est − 0,2. Au niveau de signification de 0,10: (a) existe-t-il une corrélation positive entre la superficie et la population de tous les comtés du Québec? (b) Est-ce que la corrélation est − 0,3 ou moins pour les comtés? (c) Trouver un intervalle de confiance au niveau de 95% pour ρ. Est-ce que ceci explique la contradiction qui semble exister entre (a) et (b)?

RÉPONSES

1. $s_b = \sqrt{\frac{4(50 - 5{,}22 * 12 + 0{,}59 * 28)}{2(4 * 85 - 15^2)}} = 0{,}26$; $t = \frac{-0{,}59 - 0}{0{,}26} = -2{,}27$; $\alpha = 0{,}05$ $D = 2$, test bilatéral, $t = \pm 4{,}3$. H_0 est acceptée, $B = 0$.

3. (a) $\Sigma X = 24$; $\Sigma X^2 = 56$; $\Sigma Y = 62$; $\Sigma Y^2 = 362$; $\Sigma XY = 139$; $n = 16$; $b = 2{,}30$; $a = 0{,}43$; $s_b = \sqrt{\frac{16(362 - 0{,}43 * 62 - 2{,}30 * 139)}{14(16 * 56 - 24^2)}} = 0{,}236$; $t = \frac{2{,}30 - 0}{0{,}236} = 9{,}75$; pour $\alpha = 0{,}05$; test bilatéral, $D = 14$, les valeurs critiques de t sont $\pm 2{,}15$; H_0 est rejetée.

(b) $t = \frac{2{,}30 - 2{,}50}{0{,}236} = -0{,}85$; H_0 est acceptée.

(c) $\pm 2{,}15 = \frac{2{,}30 - B}{0{,}236}$; $B = 2{,}30 \pm 0{,}51 = 1{,}8$ ou $2{,}8$; l'intervalle de confiance pour B est 1,8 à 2,8.

5. (a) $b = \frac{13 * (-182) - 0 * 20}{13 * 182 - 0^2} = -1{,}00$; $a = \frac{20}{13} - 0 = 1{,}54$; $Y_c = 1{,}54 - 1{,}00X$.

(b) $s_b = \sqrt{\frac{13[216 - 1{,}54 * 20 + 1{,}00 * (-182)]}{11 * 13 * 182}} = 0{,}040$; $t = \frac{-1{,}00 - 0}{0{,}040} = -25$.

La valeur critique de t est $-2{,}72$; donc H_0 est rejetée.

(c) $t = \frac{-1{,}00 - (-1{,}10)}{0{,}040} = 2{,}50$; les valeurs critiques de t sont $\pm 3{,}01$; H_0 est acceptée.

(d) $\pm 3{,}01 = \frac{-1{,}00 - B}{0{,}040}$, $B = -1{,}00 \pm 0{,}12 = -1{,}12$ ou $-0{,}88$; l'intervalle de confiance pour B est $-1{,}12$ à $-0{,}88$.

7. L'équation de régression est la même que celle du problème 3: $a = 0{,}43$; $b = 2{,}30$. $s_b = \sqrt{\frac{1\,600(100 * 362 - 0{,}43 * 100 * 62 - 2{,}30 * 100 * 139)}{1\,598(1\,600 * 100 * 56 - 100^2 * 24^2)}} = 0{,}022$; $\pm 1{,}96 = \frac{2{,}30 - B}{0{,}022}$. $B = 2{,}30 \pm 0{,}043 = 2{,}257$ ou $2{,}343$; la largeur de l'intervalle est 0,09: pour 64 étudiants, elle était de 0,48 et pour 16 résultats de 1,00. Pour 100 fois plus de résultats, on obtient une largeur qui est 100 fois plus petite.

9. (a) $s_e = \sqrt{\frac{(4\,000 - 0{,}500 * 440 - 2{,}00 * 1\,300)}{78}} = 3{,}89$, $\overline{X} = 2{,}5$; si $X_0 =$

2, $Y_c = 0{,}50 + 2{,}00 * 2 = 4{,}50$; $s_{Y/X_0} = 3{,}89\sqrt{\frac{1}{80} + \frac{80(2-2{,}5)^2}{8\,000}} = 0{,}476$; $D = 78$; $t = \pm 1{,}96 = \frac{4{,}50 - \mu_{Y/X_0}}{0{,}476}$; $\mu_{Y/X_0} = 3{,}57$ ou $5{,}43$

(b) $s_{Y_p} = 3{,}89\sqrt{1 + \frac{1}{80} + \frac{80(2-2{,}5)^2}{8\,000}} = 3{,}92$; $\pm 1{,}96 = \frac{4{,}50 - Y_p}{3{,}92}$, $Y_p = -3{,}18$ à $12{,}18$.

11. (a) $t = \frac{0{,}80\sqrt{49}}{\sqrt{1 - 0{,}80^2}} = 9{,}33$; $\alpha = 0{,}05$; $D = 49$, les valeurs critiques de t sont $\pm 1{,}96$; H_0 est rejetée.

(b) $H_0: \rho = 0{,}85$; $H_1: \rho < 0{,}85$; $\sigma_z = 1/\sqrt{48}$; $Z = \frac{1{,}10 - 1{,}26}{0{,}144} = -1{,}11$; la valeur critique de t est $-1{,}64$; donc H_0 est acceptée.

(c) L'habileté à lire et la taille augmentent avec l'âge.

(d) Les grandeurs des élèves d'un même niveau peuvent être distribuées normalement, mais il se peut que pour le primaire elles ne soient pas distribuées normalement; il en est de même pour l'habileté à lire.

13. $r = \frac{16 * 139 - 24 * 62}{\sqrt{16 * 56 - 24^2}\sqrt{16 * 362 - 62^2}} = 0{,}932$, $Z = 1{,}66$; $\mu_z = 1{,}47$; $z = \frac{1{,}66 - 1{,}47}{1/\sqrt{13}} = 0{,}69$. La valeur critique de z est $1{,}64$; H_0 est acceptée.

15. (a) $r = \frac{0{,}30\sqrt{9}}{\sqrt{1 - 0{,}30^2}} = 0{,}94$; si $D = 9$, $\alpha = 0{,}05$; alors $t_c = \pm 2{,}26$; H_0 est acceptée.

(b) $t = \frac{0{,}30\sqrt{100}}{\sqrt{1 - 0{,}30^2}} = 3{,}14$; si $D = 100$, $\alpha = 0{,}05$; alors $t = z = \pm 1{,}96$; H_0 est rejetée.

17. $\pm 1{,}96 = \frac{0{,}203 - \mu_z}{1/\sqrt{225}}$, $\mu_z = 0{,}072$ ou $0{,}334$, $\rho = 0{,}07$ à $0{,}32$.

Chapitre 17

Tests non-paramétriques

Un test non-paramétrique est utilisé quand (a) il n'est pas nécessaire de poser que la population est normalement distribuée (ou de poser d'autres hypothèses concernant la distribution de la population, c'est pour cette raison qu'on parle de distributions «sans modèle standard») (b) on a des données à caractères spécifiques ou des données ordinales. Les calculs sont simples puisque des données ordinales ne possèdent pas de moyenne et d'écart-type.

17.1 INTRODUCTION

Pour la plupart des tests utilisés jusqu'à maintenant, la population était considérée comme normale. On pouvait estimer l'erreur commise en faisant une inférence statistique sur la population en partant d'un échantillon. Si la distribution n'est pas normale, l'erreur peut être grande et ne peut être estimée. Il sera donc utile de connaître quelques tests où la normalité de la distribution n'est pas essentielle; ces tests sont appelés **tests non-paramétriques**. On a déjà étudié un des plus importants: le test du χ^2. Il en existe plusieurs autres; dans ce chapitre nous en étudierons quatre.

L'expression «non-paramétrique» demande des explications: originellement, elle était utilisée quand il n'était pas nécessaire de poser des hypothèses concernant la distribution de la population (tout comme pour le test du χ^2 au chapitre 13). On utilise maintenant l'expression «non-paramétrique» à chaque fois qu'une hypothèse est moins restrictive que la normalité de la distribution. Par exemple, si la distribution est continue, alors on peut comparer les résultats de deux échantillons (le 3^e résultat du premier échantillon est plus grand que le 3^e résultat du deuxième échantillon). De plus, on utilise un test non-paramétrique à chaque fois qu'on a des données à caractères spécifiques (nominales) ou des données ordinales. Ainsi la comparaison des tailles de deux échantillons de femmes se fera par un test non-paramétrique; il ne sera pas nécessaire de mesurer la taille de chaque femme, il suffira de les disposer par ordre de grandeur. Même si la distribution des tailles est distribuée normalement, le fait de considérer des données ordinales nous oblige à utiliser un test non-paramétrique.

Les tests non-paramétriques ont plusieurs avantages: on élimine l'erreur qui peut être faite en affirmant que la population est distribuée normalement alors qu'elle ne l'est pas; les calculs à effectuer sont souvent très simples; la cueillette des données est simplifiée

(surtout avec des données ordinales ou à caractères spécifiques). Pourquoi alors n'utilise-t-on pas toujours un test non-paramétrique? Supposons qu'effectivement la population est distribuée normalement et qu'on a le choix entre un test «t» et un test non-paramétrique. Le test «t» donnera en général une plus petite valeur de β pour une même valeur de α que le test non-paramétrique; il faudra donc augmenter la taille de l'échantillon du test non-paramétrique pour les mêmes valeurs de α et β. Pour le même échantillon de taille n, le test «t» sera donc plus fiable.

17.2 LE TEST DU SIGNE

Le **test du signe** est utilisé dans une expérience du type «avant-après» dans laquelle l'expérimentateur découvre si l'expérience est un succès ou non. Un exemple: on fait passer un test de réflexe à 60 conducteurs choisis au hasard avant et après avoir bu 30 cm³ d'alcool. On définit + si le temps de réaction augmente, 0 s'il est inchangé et − s'il diminue. Le nombre de résultats + et − est compté, les résultats 0 sont éliminés de l'expérience. Si on veut vérifier que l'absorption d'alcool augmente le temps de réaction, H_0 affirme qu'il n'y a pas de changement dans le temps de réaction et H_1 qu'il y a plus de + que de − . Si l'échantillon est aléatoire et l'hypothèse nulle, la probabilité qu'un individu choisi au hasard dans la population ait un temps de réaction inférieur est 1/2; le test du signe utilise une distribution binômiale avec $H_0 : P = 1/2$.

Exemple 1

Onze enfants de race blanche sont choisis au hasard en première secondaire de la polyvalente Beloeil. Ils passent un test concernant leurs préjugés raciaux. On leur administre à nouveau le test après plusieurs contacts et discussions avec huit élèves de couleur. Six d'entre eux montrent moins de préjugés, 1 en montre plus et 4 n'ont pas modifié leur opinion. Décidez si ce genre de discussions et rencontres diminuent les préjugés raciaux de tous les étudiants de secondaire I ($\alpha = 0{,}10$).

H_0 : Il n'y a pas de changement d'opinion ($P = 0{,}5$ où P est la proportion de résultats −).

H_1 : Il y a moins de préjugés ($P < 0{,}5$).

Si $P = 0{,}5$, la probabilité que 6 étudiants ou plus sur 7 (noter que les 4 étudiants qui ne présentent pas de changement sont ignorés) montrent moins de préjugés est ${}_7C_6\,(0{,}5)^6\,(0{,}5)^1 + {}_7C_7\,(0{,}5)^7 = 0{,}055 + 0{,}008 = 0{,}063$ (voir table 2, appendice C). Ce résultat est inférieur à 0,10 donc H_0 est rejetée. Au niveau de signification de 10%, les rencontres et les discussions produisent un changement dans l'attitude des étudiants de secondaire I.

Il est à noter que l'utilisation de la distribution binômiale dans le test du signe présuppose que les hypothèses de la section 6.6 sont satisfaites. Ceci veut dire que les éléments de l'échantillon sont indépendants; les étudiants blancs ne s'influencent pas les uns les autres.

Il faut également se rappeler que la distribution normale est une approximation de la distribution binômiale si n est suffisamment grand ($n \geqslant 30$), mais qu'une correction de 0,5

pour la continuité doit être utilisée (voir section 7.11). La moyenne d'une distribution binômiale avec $P = 1/2$ est $n\,P = n/2$ et l'écart-type est $\sqrt{nPQ} = \sqrt{n}/2$. Pour vérifier H_0: $P = 1/2$, pour une grande valeur de n; il suffit de calculer $z = \dfrac{X \pm 0{,}5 - n/2}{\sqrt{n}/2}$; le terme 0,5 au numérateur est la correction pour la continuité, le signe $-$ est choisi si $X > n/2$, le signe $+$ si $X < n/2$ (le choix du signe est conservateur, puisque H_0 a moins de chance d'être rejetée).

Exemple 2

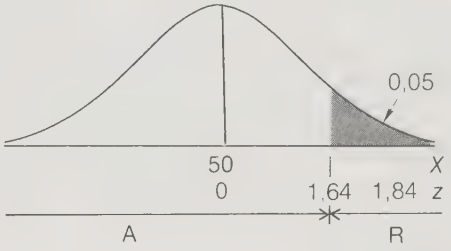
On enregistre le temps de réaction de 60 conducteurs après qu'ils aient bu 30 cm³ d'alcool. Trente-deux ont un temps de réaction plus lent, 10 ont le même temps et 18 un temps de réaction plus rapide. Le temps de réaction est-il plus lent?

$H_0 : P = 0{,}5$.

$H_1 : P > 0{,}5$ (ceci veut dire que P est la proportion des conducteurs ayant un temps de réaction plus lent).

Test unilatéral $\alpha = 0{,}05$, donc $z = 1{,}64$

$$z = \frac{X \pm 1/2 - n/2}{\sqrt{n}/2}$$



$X = 32$ = nombre de conducteurs ayant un temps de réaction plus lent.

$n = 32 + 18 = 50$ (les conducteurs ayant le même temps de réaction sont éliminés).

$$z = \frac{32 - 1/2 - 50/2}{\sqrt{50}/2} = \frac{6{,}50}{3{,}54} = 1{,}84$$

H_0 est rejetée; au niveau de 5% les conducteurs ont un temps de réaction plus lent après avoir bu 30 cm³ d'alcool.

Il sera toujours préférable d'utiliser un test t pour les différences entre les résultats (voir section 12.9) si les données sont numériques et que l'hypothèse de normalité est vérifiée.

17.3 LE TEST DES SÉQUENCES «RUNS»

Le test des **séquences** est très simple et est utilisé pour vérifier si un échantillon est

aléatoire. La médiane de l'échantillon est trouvée puis chaque résultat est marqué X s'il est inférieur à la médiane ou Y s'il lui est supérieur. Puis le nombre de séquences (runs) de X adjacents ou de Y adjacents est compté.

Exemple 1

Déterminer le nombre de suites (runs) des résultats suivants: 494; 486; 443; 419; 494; 436; 416; 481; 408; 451; 434; 488; 488; 415; 453; 401; 454; 403; 454; 456.
La médiane est 452 (le vérifier). En appelant X les résultats inférieurs à 452 et Y ceux qui lui sont supérieurs, on ignore le résultat 452, on obtient 13 séquences: $\overline{YY}$ $\underline{XX}$ $\overline{Y}$ $\underline{XX}$ $\overline{Y}$ $\underline{XXX}$ $\overline{YY}$ $\underline{X}$ $\overline{Y}$ $\underline{X}$ $\overline{Y}$ $\underline{X}$ $\overline{YY}$. Chaque séquence est indiquée par un trait horizontal.

Si l'échantillon est aléatoire, il est peu probable que les premiers résultats, pris dans l'ordre où ils sont sélectionnés, soient tous inférieurs à la médiane et que les derniers lui soient tous supérieurs. Si la sélection est aléatoire, on doit s'attendre à obtenir une alternance raisonnable des séquences sous la médiane et au-dessus de la médiane. Un test bilatéral est alors utilisé pour vérifier que l'échantillon est aléatoire.

Les symboles R, n_X, n_Y seront utilisés pour représenter le nombre de séquences (runs), le nombre de X et le nombre de Y respectivement. Les valeurs critiques de R pour $\alpha = 0{,}05$ et pour n_X et n_Y compris entre 4 et 20 sont données dans la table 9, appendice C. On rejette H_0 : l'échantillon est aléatoire et on accepte H_1 : l'échantillon n'est pas aléatoire, si le nombre de séquences n'est pas à l'intérieur des limites données par la table 9. (H_0 est rejetée si le nombre de séquences est à l'extérieur ou égal aux limites.) Le problème 11 section 17.7 donne une idée de la façon dont la table est construite.

Exemple 2

Les résultats de l'exemple 1 sont les poids en grammes de 20 boîtes de pêches. Utiliser le test des séquences par rapport à la médiane pour vérifier que l'échantillon est aléatoire ($\alpha = 0{,}05$).

$R = 13, n_X = 10, n_Y = 10$ (voir exemple 1). Les valeurs critiques de R sont 6 et 16, donc H_0 est acceptée. Si n_X ou $n_Y > 20$, on ne peut pas utiliser la table 9. Heureusement, dans ce cas, la distribution de R est approximativement normale, on peut alors calculer

$$\mu_R = 1 + \frac{2\,n_X n_Y}{n_X + n_Y}$$

$$\sigma_R = \sqrt{\frac{2\,n_X n_Y\,(2\,n_X n_Y - n_X - n_Y)}{(n_X + n_Y)^2\,(n_X + n_Y - 1)}}$$

Un exemple illustrera ce cas et montrera une autre application du test des séquences; une suite de nombres peut être vérifiée en posant X les nombres pairs et Y les impairs. Un test bilatéral est utilisé et il est peu probable que tous les nombres pairs se suivent ou encore qu'il y ait une alternance parfaite entre les nombres pairs et impairs.

Exemple 3

Les nombres suivants sont pris dans la 3e ligne de la table 3, appendice C (table de nombres aléatoires).

54222 56179 09833 34227 43897 38517 11617 30338

Utiliser un test de séquences pour les nombres pairs et impairs pour vérifier que ces nombres sont aléatoires ($\alpha = 0{,}05$).

H_0 : les nombres sont aléatoires. H_1 : ils ne sont pas aléatoires.

En posant X les nombres pairs et Y les impairs, on obtient 22 séquences

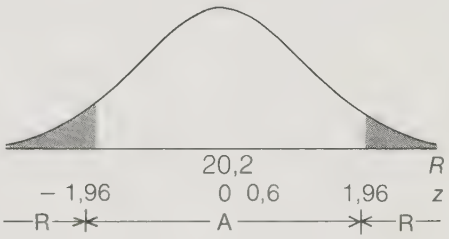
$\overline{Y}\underline{XXXX}\ \overline{Y}\underline{X}\overline{YYY}\ \underline{X}\overline{Y}\underline{X}\overline{YY}\ \overline{Y}\underline{XXX}\overline{Y}\ \underline{X}\overline{Y}\underline{X}\overline{YY}\ \overline{Y}\underline{X}\overline{YYY}\ \overline{YY}\underline{X}\overline{YY}\ \overline{Y}\underline{X}\overline{YY}\underline{X}$

$n_X = 16,\ n_Y = 24;$

$$\mu_R = \frac{2 * 16 * 24}{16 + 24} + 1 = 20{,}2$$

$$\sigma_R = \sqrt{\frac{2 * 16 * 24(2 * 16 * 24 - 16 - 24)}{40^2(40 - 1)}} = 2{,}99$$

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{22 - 20{,}2}{2{,}99} = 0{,}60$$



$\alpha = 0{,}05$, les valeurs critiques de z sont $\pm$ 1,96. H_0 est acceptée. L'échantillon est aléatoire.

17.4 LE TEST DE MANN-WHITNEY

Ce test aussi appelé «test de Wilcoxon» ou «test de somme ordinale» est utilisé pour vérifier si deux échantillons aléatoires proviennent de populations identiques. On prend pour acquis que les populations ont la même forme et la même variance; doivent-elles avoir la même médiane?

Il est possible de considérer les résultats de deux ensembles puis de les ordonner. La statistique U est calculée en comparant les sommes des rangs des données de chaque échantillon dans l'ensemble total. La table 10 appendice C peut être utilisée pour de petits échantillons (inférieurs à 20) ou un test z pour de grands échantillons.

Exemple 1

Au niveau de 0,05, est-ce que les échantillons suivants proviennent de la même population?

X:	7	9	10	11	13		
Y:	12	13	14	15	16	17	18

H_0 : Les deux échantillons proviennent de la même population.

H_1 : Ils proviennent de populations différentes.

Les deux échantillons sont réunis puis les résultats ordonnés en identifiant par X ceux qui proviennent du 1er échantillon et Y ceux du second.

	Échantillons											
	X	X	X	X	Y	X	Y	Y	Y	Y	Y	Y
Résultats	7	9	10	11	12	13	13	14	15	15	17	18
Rang	1	2	3	4	5	6,5	6,5	8	9	10	11	12

Il est à noter que les deux résultats 13 devraient avoir les rangs 6 et 7; on attribue à chacun d'eux le même rang, à savoir la moyenne entre 6 et 7 donc 6,5.

Somme des rangs des résultats $X = \Sigma R_X = 1 + 2 + 3 + 4 + 6{,}5 = 16{,}5$.
(Ne pas confondre R_X avec le nombre de séquences du test des séquences "Runs".)
Somme des rangs des résultats $Y = \Sigma R_Y = 5 + 6{,}5 + 8 + 9 + 10 + 11 + 12 = 61{,}5$.
$n_X = 5$; $n_Y = 7$.

Une vérification peut être faite en posant $R_X + R_Y = \dfrac{n(n+1)}{2}$, où $n = n_X + n_Y = 5 + 7 = 12$:

$$16{,}5 + 61{,}5 = \frac{12 * 13}{2} = 78 \quad \checkmark$$

Trouvons maintenant

$$U_X = n_X n_Y + \frac{n_X (n_X + 1)}{2} - \Sigma R_X$$

$$U_Y = n_X n_Y + \frac{n_Y (n_Y + 1)}{2} - \Sigma R_Y$$

Vérifier que $U_X + U_Y = n_X n_Y$

Soit U le plus petit de U_X ou U_Y

$$U_X = 5 * 7 + \frac{5 * 6}{2} - 16{,}5 = 33{,}5$$

$$U_Y = 5 * 7 + \frac{7 * 8}{2} - 61{,}5 = 1{,}5$$

$$33{,}5 + 1{,}5 = 5 * 7 = 35 \quad \checkmark$$

$$U = 1{,}5$$

Les valeurs critiques de U (table 10, appendice C) pour $n_X = 5$ et $n_Y = 7$ est 5. H_0 est acceptée si la valeur calculée de U est supérieure à la valeur critique. Dans cet exemple, la valeur calculée est plus petite que la valeur critique donc H_0 est rejetée; les échantillons proviennent de populations différentes.

Nous avons utilisé des données ordonnées, mais dans ce problème, les résultats sont connus. On peut donc calculer leur moyenne et leur écart-type: $\overline{X} = 10$; $s_X = 2{,}2$; $\overline{Y} = 15$; $s_Y = 2{,}2$; et on peut utiliser un test «t» pour la différence de moyenne (revoir l'exemple 2 de la section 12.7). Quels sont les avantages des tests «t» et Mann-Whitney? Pour les deux tests, il est admis que les deux populations ont la même forme et la même dispersion; pour le test «t», il est également nécessaire que la distribution soit normale. Si on doute de la normalité de la distribution, ce qui est souvent le cas quand un échantillon est petit, il est préférable d'utiliser alors le test de Mann-Whitney. Pour un α donné, la probabilité de rejeter H_0 quand elle est fausse (cette probabilité est appelée **puissance** du test et est mesurée par $1 - \beta$) est légèrement supérieure avec un test «t». Il faudra donc augmenter la taille de l'échantillon pour conserver la même puissance si on utilise un test de Mann-Whitney.

Il est possible de raccourcir les calculs en trouvant U_X ou U_Y puis, en utilisant la formule de vérification, on trouve l'autre valeur (par exemple $U_Y = n_X n_Y - U_X$).

Si les échantillons sont plus grands que 20, on ne peut plus utiliser la table 10: la distribution de U devient approximativement normale avec une moyenne de:

$$\mu_U = \frac{n_X n_Y}{2}$$

et un écart-type de

$$\sigma_U = \sqrt{\frac{n_X n_Y (n + 1)}{12}}$$

Exemple 2

Deux échantillons choisis au hasard et indépendamment sont ordonnés de la façon suivante:

X: 1; 2; 4; 5; 10; 11; 12; 14; 15; 16; 20; 21; 24; 27; 28; 29; 34; 35; 36.

Y: 3; 6; 7; 8; 9; 13; 17; 18; 19; 22; 23; 25; 26; 30; 31; 32; 33; 37; 38; 39; 40; 41; 42; 43.

Au niveau de 0,05, est-ce que ces deux échantillons proviennent de la même population?

H_0 : Les deux échantillons proviennent de la même population.

H_1 : Ils ne proviennent pas de la même population.

$$\Sigma R_X = 344,\ n_X = 19,\ n_Y = 24,\ n = 19 + 24 = 43$$

$$\Sigma R_Y = \frac{n(n+1)}{2} - \Sigma R_X = \frac{43 * 44}{2} - 344 = 602$$

$$U_X = 19 * 24 + \frac{19 * 20}{2} - 344 = 302$$

dans cet exemple, $\mu_U = \frac{19 * 24}{2} = 228$, $\sigma_U = \sqrt{\frac{19 * 24 * 44}{12}} = 40{,}9$.

Le résultat z pour U_X est $z = \dfrac{U_X - \mu_U}{\sigma_U} = \dfrac{302 - 228}{40{,}9} = 1{,}81$

Calculons maintenant U_Y: $U_Y = 19 * 24 + \dfrac{24 * 25}{2} - 602 = 154$

Le résultat z pour U_Y est $\dfrac{154 - 228}{40{,}9} = -1{,}81$ et cette valeur est la valeur négative du résultat z pour U_X. Ce sera toujours le cas, il suffit alors de calculer l'une des deux valeurs.

Pour $\alpha = 0{,}05$, et pour un test bilatéral, les valeurs critiques de z sont $\pm 1{,}96$; H_0 est acceptée: les deux échantillons viennent de la même population.

17.5 COEFFICIENT DE CORRÉLATION DE RANG DE SPEARMAN

Ce coefficient r_S mesure la corrélation entre deux paires d'échantillons de données ordonnées. Par exemple, deux juges peuvent ordonner un échantillon de n étudiants selon leur habileté à écrire. Les classements des juges sont-ils comparables? Les habiletés d'un individu en dessin, comptabilité, lecture etc. peuvent être ordonnées avant et après un cours d'entraînement. Est-ce que l'individu est meilleur en dessin et mauvais en lecture? Tout comme pour le coefficient de Pearson r, si les classements sont accordés (les classements des juges sont identiques) on obtient une grande corrélation positive; $r_S = +1$. Si les juges sont en désaccord complet, on obtient alors une grande corrélation négative.

Le coefficient de rang de Spearman est le coefficient de Pearson r appliqué à deux échantillons ordonnés. La formule pour calculer r s'applique, mais on peut simplifier les calculs pour r_S:

1. Faire une liste de n paires ordonnées $(X; Y)$.
2. Trouver les différences d de chaque paire $d = X - Y$.
3. *Élever au carré les différences puis additionner* (Σd^2).
4. Calculer r_S.

$$r_S = 1 - \frac{6 \Sigma d^2}{n(n^2 - 1)}$$

Le coefficient de corrélation de rang de Spearman est utile (1) quand les rangs sont connus (2) quand les résultats sont connus, mais qu'il est plus facile de déterminer r_S. Il y a un inconvénient à utiliser r_S; si r_S est près de 1 ou de (-1), on sait que si X augmente, Y augmente (ou diminue), mais on ne sait pas de quelle façon cette augmentation (ou diminution) se fait. Avec le coefficient de Pearson r, on sait que si r est près de 1 (ou de -1), il existe une relation linéaire (régression) entre les X et les Y, ce qui permet de dire que les paires $(X; Y)$ sont bien représentées par la droite de régression.

Exemple 1

Six peintures sont classées de la façon suivante par deux juges; calculer r_S: (a) en utilisant les 4 étapes décrites précédemment; (b) en utilisant la formule de Pearson r.

	X	Y	(a)		(b)		
Peintures	1er Juge	2e Juge	d	d^2	XY	X^2	Y^2
A	2	2	0	0	4	4	4
B	1	3	−2	4	3	1	9
C	4	4	0	0	16	16	16
D	5	6	−1	1	30	25	36
E	6	5	1	1	30	36	25
F	3	1	2	4	3	9	1
	(21)	(21)		10	86	91	91

(a) $\Sigma d^2 = 10$, $n = 6$.

$$r_S = 1 - \frac{6 * 10}{6 * (6^2 - 1)} = 0{,}71$$

(b) $$r = \frac{6 * 86 - 21 * 21}{\sqrt{6 * 91 - 21^2}\ \sqrt{6 * 91 - 21^2}} = \frac{75}{105} = 0{,}71 = r_S$$

Si deux résultats ont le même rang, on remplace ceux-ci par la moyenne des rangs qu'ils devraient avoir (voir exemple 1, section 17.4). Ainsi si deux éléments sont ex aequo en 3e place (1; 2; 3; 3; 5) on écrira (1; 2; 3,5; 3,5; 5).

Exemple 2

Calculer r_S pour les classements de 6 individus à deux tests:

	Individus					
	A	B	C	D	E	F
Test X	1	2	2	4	5	6
Test Y	3	2	1	4	6	5

Il y a égalité en deuxième place dans le test X, donc chaque 2 sera remplacé par 2,5.

X	Y	d	d^2
1	3	− 2	4
2,5	2	0,5	0,25
2,5	1	1,5	2,25
4	4	0	0
5	6	− 1	1
6	5	1	1
			8,5

$$n = 6$$

$$r_S = 1 - \frac{6 * 8{,}5}{6 * (6^2 - 1)} = 1 - 0{,}24 = 0{,}76$$

Exemple 3

Huit ouvriers de l'usine Bombardier sont choisis au hasard et sont classés comme suit par un test de dextérité manuelle et par un test de connaissance de la mécanique. Existe-t-il une corrélation entre la dextérité et la connaissance?

ouvrier	dextérité	connaissances	X	Y	d	d^2	
A	1	2	1	3	−2	4	
B	2	1	2	1	1	1	
C	3	8	4	8	−4	16	$n = 8$
D	3	2	4	3	1	1	
E	3	2	4	3	1	1	
F	6	6	6	6,5	− 0,5	0,25	
G	7	6	7	6,5	0,5	0,25	
H	8	5	8	5	3	9	
						32,50	

La colonne X: la dextérité manuelle est ajustée en tenant compte de la triple égalité en 3e place; la colonne Y: la connaissance de la mécanique est ajustée en tenant compte des doubles égalités en 2e et 6e places.

$$r_S = 1 - \frac{6 * 32{,}50}{8 * (8^2 - 1)} = 1 - 0{,}39 = 0{,}61$$

Peut-on vérifier une hypothèse concernant le coefficient de corrélation de rang de Spearman pour la population ρ_S? Si l'hypothèse nulle est que $\rho_S = 0$, un test «t» peut être utilisé tout comme pour vérifier $H_0 : \rho = 0$:

$$t = \frac{r_S\sqrt{n - 2}}{\sqrt{1 - r_S^2}}$$

Avec $(n - 2)$ degrés de liberté. Dans ce cas, n doit être supérieur ou égal à 10.

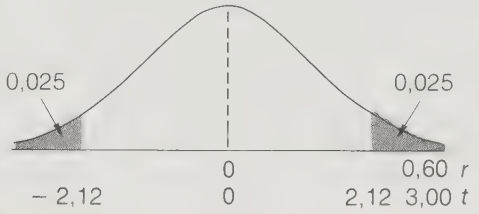
Exemple 4

Dix-huit ouvriers de l'usine ABC sont choisis au hasard et donnent un coefficient de corrélation de rang de Spearman de 0,60 entre la dextérité manuelle et la connaissance de la mécanique. Existe-t-il une corrélation significative ($\alpha = 0{,}05$) pour tous les ouvriers de l'usine?

$$H_0 : \rho_S = 0,\ H_1 : \rho_S \neq 0;\ \alpha = 0{,}05$$

$$r_s = 0{,}60;\ n = 18;\ \text{alors}\ t = \frac{0{,}6 * \sqrt{18-2}}{\sqrt{1-0{,}6^2}} = 3{,}00$$

Pour un test bilatéral avec 16 degrés de liberté; $\alpha = 0{,}05$; les valeurs critiques de «t» sont $\pm$ 2,12. H_0 est rejetée; H_1est acceptée, il y a une corrélation ($\rho_s \neq 0$).



17.6 VOCABULAIRE ET SYMBOLES

Tests non-paramétriques
Test du signe
Test des séquences (runs) R, μ_R, σ_R
Séquences (runs)
Test de Mann-Whitney R_x, R_y, U_x, U_y, μ_U, σ_U
Test de Wilcoxon
Test de sommes ordinales
Coefficient de corrélation de rang r_s, d, ρ_s

17.7 EXERCICES

1. Un horticulteur veut comparer l'efficacité de deux insecticides le Képope et le Myrex sur les fourmis. Il va dans un champ où il y a plusieurs fourmillières et en choisit 36 au hasard. Il forme 18 paires en tenant compte des grosseurs des nids. Pour chaque paire, il alimente un nid avec du Képope et l'autre avec la même quantité de Myrex. Après une semaine, il retourne dans le champ et constate que dans 8 paires, le nid traité avec du Myrex n'a plus que quelques fourmis; dans 6 paires toutes les fourmis sont mortes; dans 4 paires, le Képope a mieux réussi. Utiliser un test du signe pour savoir si les deux insecticides produisent le même effet ($\alpha = 0{,}10$).

2. Deux alpinistes se demandent lesquels des clous (pitons) étroits ou larges sont les plus populaires. Ils s'adressent à plusieurs clubs d'alpinistes pour mener une enquête. Quinze équipes de deux alpinistes chacune s'attaquent à la pente abrupte du Mont St-Hilaire; une première fois, avec des pitons étroits et une deuxième fois avec des pitons larges. Des trente alpinistes 20 préfèrent les pitons étroits et 10 préfèrent les larges. Au niveau de signification de 0,05, y a-t-il une différence dans la préférence?

3. Soixante-dix-sept vendeurs de la compagnie Balayeuse Miracle suivent une semaine de cours intensifs de vente. Dans les mois qui suivirent, 40 améliorèrent de façon sensible leur nombre de ventes; 28 n'ont pas amélioré et 9 ont vendu moins que précédemment. Au niveau de 0,02, est-ce que les ventes ont augmenté après le cours?

4. Douze enfants d'une même classe sont classés par ordre de popularité (a) par le groupe (b) par un test de sociabilité. Les résultats sont les suivants:

		Enfants											
		A	B	C	D	E	F	G	H	I	J	K	L
(a)	groupe	6	2	7	3	4	11	1	5	10	8	9	12
(b)	test	7	1	8	4	6	3	2	5	10	9	11	12

Utiliser un test de signe sur les différences de rangs pour vérifier si les deux méthodes de classement sont différentes au niveau de 0,05.

5. Vingt-cinq enfants d'une même classe sont classés par ordre de popularité (a) par le groupe (b) par un test de sociabilité. Dix-neuf ont un meilleur classement par le groupe, 11 ont un moins bon classement et 5 ont le même classement. Utiliser un test du signe pour vérifier si les classements par les deux méthodes diffèrent de façon significative ($\alpha = 0{,}05$).

6. Une discussion avec les étudiants de couleurs de l'exemple 1 de la section 17.2 est étendue à 261 étudiants blancs. Après un test, 113 étudiants blancs ont moins de préjugés, 83 en ont plus et 65 ont la même opinion. Peut-on affirmer que ce type de discussion diminue les préjugés raciaux de tous les enfants de l'école ($\alpha = 0{,}05$)?

7. Les nombres suivants proviennent de la première ligne, colonnes 3, à 6 de la table 3 appendice C.

81536 09686 26743 87001

Utiliser un test des séquences, en comptant les séquences inférieures ou supérieures à la médiane, pour vérifier si ces nombres sont aléatoires ($\alpha = 0{,}05$).

8. Utiliser un test des séquences en comptant les suites de nombres pairs et impairs pour vérifier si les nombres de l'exercice 7 sont aléatoires.

9. La médiane des nombres de l'exemple 3 de la section 17.3 est 3,5. Utiliser un test des séquences par rapport à la médiane pour vérifier si ces nombres sont aléatoires ($\alpha = 0{,}05$).

10. Un échantillon de dossiers pris au Palais de Justice de Montréal donne les peines d'emprisonnements suivantes en années: 3; 1; 1/2; 2; 20; 5; 1; 6; 3; 1; 1/2; 1/6; 4; 2; 15; 7; 1; 1/2; 2; 3; 1; 4; 1; 2; 3; 1; 2; 18; 8; 1; 2; 3; 2; 5; 1; 2; 1; 7; 12; 9; 2; 1; 3; 4; 2; 1 1/2; 1/2; 6; 1/2; 3; 1; 5; 2; 7; 2; 1; 1; 4; 1; 1.
Est-ce un échantillon aléatoire au niveau de 0,10?

11. (a) Si $n_x = 2$; $n_y = 1$ et X et Y sont ordonnés au hasard dans une liste double, quelle est la probabilité que (i) $R = 3$? (ii) $R = 2$? (Conseil: (i) $p(XYX) = ?$ (ii) p (*XXY* ou *YXX*) $= ?$)

(b) Si $n_x = 3$; $n_y = 2$ et X *et* Y sont ordonnés au hasard dans une liste double, quelle est la probabilité que (i) $R = 2$? (ii) $R = 3$? (iii) $R = 4$? (iv) $R = 5$?

(c) Si $\alpha = 0{,}40$, pour quelles valeurs de R, H_0 sera-t-elle rejetée par un test bilatéral de séquences si $n_x = 3$ et $n_y = 2$?
Ce problème doit vous donner une bonne idée de la façon dont les valeurs critiques de R sont déterminées pour un test de séquences. Pour de plus grandes valeurs de n_x et n_y les calculs peuvent être longs si une autre méthode n'est pas utilisée.

12. Au niveau de 0,05, est-ce que les échantillons suivants proviennent de populations identiques?

1er échantillon (X) : 6; 9; 5; 8; 6; 4.
2e échantillon (Y) : 3; 6; 10; 7; 4; 9; 11; 7.

13. Un échantillon aléatoire d'hommes mariés est choisi puis les questions suivantes leur sont posées: êtes-vous satisfait ou insatisfait de votre mariage et depuis combien de temps êtes-vous mariés?

	Années de mariage
Satisfait	1; 3; 3; 2; 7; 10.
Insatisfait	16; 4; 9; 20; 16; 7; 4; 6; 18.

Au niveau de signification de 0,05, y a-t-il une différence entre les deux groupes?

14. Six cours choisis au hasard à l'université de Montréal sont évalués de 1 à 6 (où 6 est excellent) par les étudiants.
Deux items sont évalués: le travail demandé en regard des crédits accordés et le prix des manuels utilisés. Les résultats moyens des six cours sont les suivants:

	Cours					
	A	B	C	D	E	F
Travail demandé	3,1	2,4	5,0	3,0	1,8	4,2
Coût des livres	4,0	5,7	5,4	3,4	1,2	5,0

(a) Utiliser un test «t» pour vérifier au niveau de 0,05 que tous les cours ont le même résultat moyen à chacune des deux questions.
(b) Utiliser un test de Mann-Whitney pour vérifier que tous les cours ont la même distribution que l'évaluation de chacune des deux questions ($\alpha = 0{,}05$).
(c) Si les résultats moyens des six cours sont ordonnés, trouver le coefficient de corrélation de rang de Spearman entre les deux questions.

15. Six étudiants sont choisis au hasard parmi ceux qui étudient l'Économie 101 et la Psychologie 201; le premier avec le professeur A et le second avec le professeur B. Les étudiants évaluent leurs professeurs selon la cote 1 à 6 (6 étant excellent). Voici les résultats:

	Étudiants					
	A	B	C	D	E	F
Professeur A	5	4	6	4	4	5
Professeur B	4	3	5	2	4	3

Il n'y a aucune raison de croire que les évaluations de tous les étudiants sont normalement distribuées. Y a-t-il une différence significative entre les évaluations des deux professeurs ($\alpha = 0{,}05$)?

16. Cinquante hommes de science sont choisis au hasard et, pour chacun, le nombre de publications est enregistré. Des 50 scientistes, 27 sont les aînés de la famille. Se basant sur les classements suivants, y a-t-il une différence entre le nombre de publications des aînés et des autres scientistes ($\alpha = 0{,}01$)?

	Rangs (nombre de publications)
Scientistes aînés	2; 2; 2; 4; 7; 7; 7; 11; 11; 11; 16; 16; 16; 16; 16; 22; 22; 22; 28; 31; 33; 33; 33; 38; 38; 38; 44.
Autres scientistes	5; 9; 13; 19; 20; 25; 25; 25; 27; 29, 5; 29,5; 35; 36; 41; 41; 41; 43; 47,5; 47,5; 47,5; 47,5; 47,5; 47,5.

17. La compagnie Chemilin croit qu'un additif au caoutchouc de ses pneus diminuera l'usure. Deux échantillons de 25 pneus sont vérifiés sous les mêmes conditions. Voici les résultats:

	Usure (en mm)
Sans additif	23; 14; 12; 11; 13; 18; 15; 16; 13; 10; 15; 20; 14; 9; 7; 20; 14; 18; 24; 19; 13; 16; 15; 20; 15.
Avec additif	14; 19; 8; 10; 16; 12; 11; 9; 18; 12; 7; 6; 8; 11; 17; 13; 13; 15; 10; 8; 14; 7; 5; 13; 14.

(a) Utiliser un test «t» pour vérifier si les pneus avec additif ont sensiblement moins d'usure ($\alpha = 0{,}05$). Supposons que les variances des mesures sont les mêmes pour les deux types de pneus et que les mesures sont normalement distribuées. $\overline{X} = 15{,}4$; $s^2_x = 17{,}4$; $\overline{Y} = 11{,}6$; $s^2_y = 14{,}5$.

(b) Utiliser un test de Mann-Whitney pour vérifier que l'additif diminue l'usure au niveau de 0,05.

18. Septs film furent évalués par les étudiants du département de cinéma de l'U.Q.A.M. Le classement fut fait pour la qualité technique et pour l'intrigue. Trouver la corrélation entre leur évaluation.

	Films						
	A	B	C	D	E	F	G
Technique	1	2	2	4	5	6	6
Intrigue	3	4	2	1	7	5	6

19. Dix couturières sont choisies au hasard dans une manufacture. Elles sont classées selon leur habileté et leur vitesse. Trouver la corrélation entre les deux classements.

	Couturières									
	A	B	C	D	E	F	G	H	I	J
Habileté	10	2	4	5	9	5	1	3	7	8
Vitesse	10	5	2	4	8	6	2	1	8	7

20. Pour les paires suivantes (a) déterminer r (b) r_s (c) tracer un diagramme de dispersion puis expliquer la différence entre r et r_s.

X :	3	5	0	2	4	1
Y :	3	25	0	2	16	1

21. Cinquante et un ouvriers de manufactures montrent un coefficient de rang de Spearman de 0,50 entre leur absentéisme et leurs problèmes familiaux. Pour tous les ouvriers existe-t-il une corrélation significative ($\alpha = 0,05$)?

22. Se référer au problème 14.

(a) Pour tous les cours de l'université existe-t-il un coefficient de corrélation significatif entre les évaluations des deux questions?

(b) Supposons que 100 cours au lieu de 6 furent évalués, mais que le coefficient de corrélation de Spearman est le même. Est-ce qu'il y a une corrélation significative entre les deux questions ($\alpha = 0,05$)?

RÉPONSES

1. P = proportion des cas où Myrex est supérieur $H_0: P = 0,5$; $H_1: P > 0,5$; $n = 12$ (on ignore les 6 où les deux insecticides sont efficaces) $X = 8$; p (8 ou plus en 12 essais avec $P = 0,5$) $= 0,121 + 0,054 + 0,016 + 0,003 = 0,194$; qui est plus grand que 0,10 donc H_0 est acceptée; il n'y a pas de différence.

3. $H_0: P = 0,5$; $H_1: P > 0,5$; $n = 49$; $\mu = 24,5$; $\sigma = \sqrt{49 * 0,5 * 0,5} = 3,5$; $z = \dfrac{40 - 0,5 - 24,5}{3,5} = 4,29$. $\alpha = 0,02$ la valeur critique de z est 2,05. H_0 est rejetée; les ventes ont augmenté.

5. $n = 30$; $\mu = 15$; $\sigma = 2,7$; $z = \dfrac{19 - 0,5 - 15}{2,7} = 1,30$. $\alpha = 0,05$; la valeur critique de z est 1,96. H_0 est acceptée: P = 0,5; H_1 est rejetée: P ≠ 0,5.

7. $Md = 5,5$; 81536 09686 26743 87001; $R = 10$; $n_x = 10$; $n_Y = 10$. $\alpha = 0,05$; les valeurs critiques de R sont 6 et 16. H_0 acceptée les nombres sont aléatoires.

9. 54222 56179 09833 34227 43897 38517 11617 30338; $R = 25$; $n_x = n_Y = 20$; Les valeurs critiques de R sont 14 et 28. H_0 est acceptée les nombres sont aléatoires.

11. (a) (i) $\dfrac{2}{3} * \dfrac{1}{2} * \dfrac{1}{1} = \dfrac{1}{3}$; (ii) $\dfrac{2}{3} * \dfrac{1}{2} * \dfrac{1}{1} + \dfrac{1}{3} * \dfrac{2}{2} * \dfrac{1}{1} = \dfrac{2}{3}$.

(b) Il y a 5! permutations de 5 lettres et 3! * 2! de 3 (X) et 2 (Y) donc la probabilité de chaque permutation est 3! * 2!/5! = 1/10. (i) p ($XXXYY$ ou $YYXXX$) = 0,2; (ii) p ($YXXXY$ ou $XYYXX$ ou $XXYYX$) = 0,3; (iii) p ($XYXXY$ ou $XXYXY$ ou $YXXYX$ ou $YXYXX$) = 0,4; (iv) p ($XYXYX$) = 0,1.

(c) Rejeter si $R = 2$ ou 5.

13.

Résultats:	1	2	3	3	4	4	6	7	7	9	10	16	16	18	20
Rangs:	1	2	3	4	5	6	7	8.5	8.5	10	11	12	13	14	15

$n_x = 6$, $n_Y = 9$; $R_x = 29,5$; $U_x = 6 * 9 + 6 * 7/2 - 29,5 = 45,5$; $U_Y = 6 * 9 - 45,5 = 8,5 = U$. La valeur critique de U est 10 donc H_0 est rejetée; les deux groupes sont différents.

15.

Résultats:	2	3	3	4	4	4	4	4	5	5	5	6
Rangs:	1	2,5	2,5	6	6	6	6	6	10	10	10	12

$n_x = n_Y = 6$; $R_x = 28$; $U_x = 6 * 6 + 6 * 7/2 - 28 = 29$, $U_Y = 6 * 6 - 29 = 7$; $\alpha = 0,05$; la

valeur critique de U est 5 donc H_0 est acceptée.

17. (a) $H_0 : \mu_X - \mu_Y = 0$, $H_1 : \mu_X - \mu_Y > 0$ (X est sans additif) , $\sigma_{(\bar{X}-\bar{Y})} = \sqrt{\frac{24 * 17{,}4 + 24 * 14{,}5}{48} * \frac{2}{25}} = 1{,}13$; $t = \frac{15{,}4 - 11{,}6}{1{,}13} = 3{,}36$. $D = 48$, la valeur critique($= z$) $= 1{,}64$; H_0 est rejetée.

(b)

Résultats		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	23	24
Fréquences	X	0	0	1	0	1	1	1	1	3	3	4	2	0	2	1	3	1	1
	Y	1	1	2	3	1	2	2	2	3	3	1	1	1	1	1	0	0	0

Rangs X: 4; 9,5; 12; 15; 18; 22,5; 22,5; 22,5; 28,5; 28,5; 28,5; 34; 34; 34; 34; 38; 38; 42; 42; 44,5; 47; 47; 47; 49; 50.

Rangs Y: 1; 2; 4; 4; 7; 7; 7; 9,5; 12; 12; 15; 15; 18; 18; 22,5; 22,5; 22,5; 28,5; 28,5; 28,5; 34; 38; 40; 42; 44,5.

$R_Y = 483$, $U_Y = 25 * 25 + 25 * 26/2 - 483 = 467$, $U_X = 25^2 - 467 = 158 = U$, $\mu_U = 25^2/2 = 312{,}5$; $\sigma_U = \sqrt{25 * 25 * 51/12} = 51{,}54$; $z = \frac{158 - 312{,}5}{51{,}54} = -3{,}00$; $\alpha = 0{,}05$; la valeur critique de z est $-1{,}64$. H_0 est rejetée.

19. d : 0; -3; 2;1; 1; -1; -1; 2; -1; 1; $\Sigma d^2 = 23$, $r_S = 1 - \frac{6 * 23}{10 * 99} = 0{,}86$.

21. $H_0 : \rho_S = 0$, $H_1 : \rho_S \neq 0$, $t = \frac{0{,}50\sqrt{49}}{\sqrt{1 - 0{,}50^2}} = 4{,}04$; $D = 49$; les valeurs critiques de «t» sont $\pm$ 1,96. H_0 est rejetée, il y a une corrélation significative.

Appendice A

Symboles

Les nombres entre parenthèses indiquent la section où le symbole est utilisé la première fois.

A	Un événement, un sous-ensemble (5.3)
$\overline{A}$	Le complément d'un ensemble événement complémentaire. (5.5)
$A(z)$	Aire sous la courbe normale entre 0 et z (7.10)
α (alpha)	Probabilité de l'erreur de première espèce, niveau de signification (11.4)
β (beta)	Probabilité de l'erreur de deuxième espèce (11.4)
$1 - \beta$	Puissance d'un test statistique (17.4)
${}_nC_X$	Nombre de combinaisons de n objets pris X à la fois (6.5)
D, D_X, D_Y	Degrés de liberté (12.5 — 14.1)
E	Erreur (10.7)
E	Fréquence théorique (13.1)
f	Fréquence (2.4)
$F(D_X, D_Y)$	Résultat F (14.2)
$F_\alpha(D_X, D_Y)$	Valeur critique de F pour un test unilatéral avec α donné (14.4)
H_0	Hypothèse nulle (11.2)
H_1	Hypothèse rivale (11.3)
k	Nombre d'échantillons (14.4)
Md	Médiane (3.3)
Mo	Mode (3.2)
μ (mu)	Moyenne de la population (6.1)
μ_d	Moyenne des différences des paires d'une population (12.9)
μ_p	Moyenne des proportions d'une population (10.1)

$\mu_{p_X - p_Y}$	Moyenne de différences de proportions d'un échantillon (10.5)
$\mu_{\bar{X}}$	Moyenne de la distribution d'échantillonnage des moyennes (9.2)
μ_R	Moyenne de la distribution d'échantillonnage des séquences (17.3)
$\mu_{\bar{X}-\bar{Y}}$	Moyenne de la distribution d'échantillonnage des différences de moyennes (10.3)
μ_{Y/X_0}	Moyenne de la population Y pour un X_0 donné (16.3)
n	Taille de l'échantillon (3.4)
$n!$	Factoriel n (6.4)
O	Fréquence observée (13.1)
p	Probabilité de succès (6.7)
p	Proportion de réussite d'un événement dans un échantillon (10.1)
$p(a < X < b)$	Probabilité que X soit compris entre a et b (7.2)
$p(A)$	Probabilité de l'événement A (5.2)
$p(B/A)$	Probabilité de l'événement B étant donné la réussite de A (5.7)
P	Proportion d'un événement dans la population (10.1)
P_a	La a^e centile (3.6)
${}_nP_X$	Permutation de n objets pris X à la fois (6.6)
Q_1	Premier quartile (3.6)
r	Coefficient de corrélation linéaire (15.6)
r^2	Coefficient de détermination (15.7)
r_S	Coefficient de corrélation de rang de Spearman (17.5)
R	Nombre de séquences (runs) (17.3)
R_X $[R_Y]$	Somme des séquences X (17.4)
ρ (rho)	Coefficient de corrélation linéaire de la population (16.4)
s	Écart-type d'un échantillon (4.6)
s^2	Variance de l'échantillon (4.6)
s_b	Erreur standard de l'estimation (16.2)
s_d	Écart-type de différences dans un échantillon de paires (12.9)
s_e	Erreur type d'estimation (15.7)
s_{Y_p}	Écart-type de la moyenne des Y pour un X_0 (16.3)
S	Espace échantillonnal (5.3)
S	Test de Scheffé (14.7)
σ (sigma)	Écart-type de la population (6.1)
σ^2	Variance de la population (6.1)

$\hat{\sigma}^2$ intérieur	Estimation de σ^2 intérieur pour les échantillons (14.4)
$\hat{\sigma}^2$ extérieur	Estimation de σ^2 entre les échantillons (14.4)
σ_p	Écart-type des proportions (10.1)
$\sigma_{(p_X - p_Y)}$	Écart-type de différences de proportions (10.5)
σ_R	Écart-type du nombre de séquences (17.3)
$\sigma_{\bar{X}}$	Écart-type de la moyenne (9.2)
$\sigma_{(\bar{X} - \bar{Y})}$	Écart-type de différences de moyennes (10.3)
Σ	Sommation (1.4)
t	Distribution t de Student (12.1)
U	Test de Mann-Whitney (17.4)
X	Résultat, donnée (2.4)
$\overline{X}$	Moyenne de l'échantillon (3.4)
Y_c	Valeur de Y calculée (15.5)
Y_p	Valeur prédite de Y (16.3)
z	Résultat standard, cote (7.7)
Z	Résultat Z de Fisher (16.6)
$\neq$	Non égal
$*$	Multiplication
$>$	Plus grand que
$<$	Plus petit que
$\leqslant$	Plus petit ou égal
$\simeq$	Approximativement égal
∞	Infinie
$\lvert a \rvert$	Valeur absolue de a.

Appendice B

Formules

Les nombres à droite indiquent la section.

Moyenne d'un échantillon	$\overline{X} = \frac{\Sigma X}{n}$	3.4
Moyenne d'un échantillon (données regroupées)	$\overline{X} = \frac{\Sigma fX}{n}$	3.5
Moyenne d'une population	$\mu = \frac{\Sigma X}{N}$	7.6
Moyenne d'une distribution de probabilité	$\mu = \Sigma X p(X)$	6.1
Moyenne d'une distribution binômiale	$\mu = np$	6.9
Rapport de variation	$V = \frac{n - f_{Mo}}{n}$	4.2
Intervalle semi-interquartile	$Q = \frac{Q_3 - Q_1}{2}$	4.4
Variance d'un échantillon	$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$	4.6
	$s^2 = \frac{\Sigma X^2 - (\Sigma X)^2/n}{n - 1}$	4.6
Variance d'un échantillon (données regroupées)	$s^2 = \frac{\Sigma f(X - \overline{X})^2}{n - 1}$	4.7
	$s^2 = \frac{\Sigma fX^2 - (\Sigma fX)^2/n}{n - 1}$	4.7

Variance d'une population	$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$	7.6
	$\sigma^2 = \frac{\Sigma X^2 - (\Sigma X)^2/N}{N}$	7.6
Variance d'une population (données regroupées)	$\sigma^2 = \frac{\Sigma fX^2 - (\Sigma fX)^2/N}{N}$	9.2
Variance d'une distribution de probabilité	$\sigma^2 = \Sigma(X - \mu)^2 p(X)$	6.1
Variance d'une distribution binômiale	$\sigma^2 = npq$	6.9
Écart-type d'un échantillon	$s = \sqrt{s^2}$	4.6
Écart-type d'une population	$\sigma = \sqrt{\sigma^2}$	7.6
Dissymétrie	$\frac{3(\bar{X} - Md)}{s}$	4.8
Probabilité d'événements équiprobables	$p(A) = \frac{s}{n}$	5.2
	$p(A \text{ ou } B) = p(A) + p(B) \quad p(A \text{ et } B)$	5.6
Probabilité si A et B sont mutuellement exclusifs	$p(A \text{ ou } B) = p(A) + p(B)$	5.4
Probabilité si A et B sont indépendants	$p\ (A \text{ et } B) = p(A)p(B)$	5.4
Probabilité si A et B ne sont pas indépendants	$p\ (A \text{ et } B) = p(A)p(B/A)$	5.4
Pente d'une droite	$b = \frac{Y_2 - Y_1}{X_2 \quad X_1}$	15.3
Droite	$Y = a + bX$	15.3
Équation de régression	$Y_c = a + bX$	15.5
	$b = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$	15.5
	$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n}$	15.5

Erreur type d'estimation

$$s_e = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n - 2}} \qquad 15.5$$

$$\Sigma(Y - Y_c)^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY$$

Erreur de variation

$$\Sigma(Y - Y_c)^2$$

Variation expliquée

$$(\Sigma Y_c - \overline{Y})^2$$

Variation totale

$$\Sigma(Y - \overline{Y})^2 \qquad 15.7$$

Coefficient de détermination

$$r^2 = \frac{\text{variation expliquée}}{\text{variation totale}} \qquad 15.7$$

Coefficient de corrélation linéaire

$$r = (\text{signe de } b)\sqrt{r^2} \qquad 15.7$$

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\ \sqrt{n\Sigma Y^2 - (\Sigma Y)^2}} \qquad 15.8$$

$$r = b\frac{s_X}{s_Y} \qquad 15.8$$

Estimation rapide

$$r = \pm\left(1 - \frac{W}{L}\right) \qquad 15.5$$

Coefficient de corrélation de rang de Spearman

$$r_S = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} \qquad 17.5$$

Inférence pour	*Formule*	*Erreur type* †	
Moyenne si σ est connu **et** la population normale où $n \geqslant 30$	$z = \dfrac{\overline{X} - \mu}{\sigma_{\overline{X}}}$	$\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$	9.6
Moyenne si σ est inconnu **et** la population normale	$t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}, D = n - 1$		12.6
Proportion si p est connu	$z = \dfrac{p - P}{\sigma_p}$	$\sigma_p = \sqrt{\dfrac{PQ}{n}}$	11.4
Proportion si p est inconnu	$z = \dfrac{p - P}{\sigma_p}$	$\sigma_p = \sqrt{\dfrac{pq}{n}}$	10.6
Différence de moyennes, deux échantillons indépendants si σ_X et σ_Y sont connus et les deux populations sont normales où $n_X \geqslant 30$, $n_Y \geqslant 30$	$z = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\overline{X}-\overline{Y})}}$	$\sigma_{(\overline{X}-\overline{Y})} = \sqrt{\sigma_X^2 + \sigma_Y^2}$	10.4
Différence de moyennes, deux échantillons indépendants si σ_X et σ_Y sont inconnus mais égaux et les deux populations sont normales	$\begin{cases} t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\overline{X}-\overline{Y})}} \\ D = n_X + n_Y - 2 \end{cases}$	$\sigma_{(\overline{X}-\overline{Y})} = \sqrt{\dfrac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}\left(\dfrac{1}{n_X} + \dfrac{1}{n_Y}\right)}$	12.7
Différence de moyennes, deux échantillons indépendants, si σ_X et σ_Y sont inconnus et différents et les populations sont normales	$\begin{cases} t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sigma_{(\overline{X}-\overline{Y})}} \\ D = \dfrac{(s_X^2/n_X + s_Y^2/n_Y)^2}{\left(\dfrac{s_X^2}{n_X}\right)^2 \dfrac{1}{n_X + 1} + \left(\dfrac{s_Y^2}{n_Y}\right)^2 \dfrac{1}{n_Y + 1}} - 2 \end{cases}$	$\sigma_{(\overline{X}-\overline{Y})} = \sqrt{\dfrac{s_X^2}{n_X} + \dfrac{s_Y^2}{n_Y}}$	12.7

Différence de moyennes de deux échantillons ou plus si les populations sont normales et ont la même variance	$F = \dfrac{\hat{\sigma}^2 \text{ intérieur}}{\hat{\sigma}^2 \text{ extérieur}}$ $\hat{\sigma}^2 \text{ intérieur} = \dfrac{A - B^2/n_T}{k-1}$, $\hat{\sigma}^2 \text{ extérieur} = \dfrac{C-A}{n_T - k}$ (voir section 14.6 pour A, B et C)	14.5
Quelles moyennes sont différentes? (test de Scheffé)	Comparer $\lvert\overline{X}_i - \overline{X}_j\rvert$ et $S\sqrt{1/n_i + 1/n_j}$ $S = \sqrt{(k-1)F_\alpha(k-1, n-k)(\hat{\sigma}^2 \text{ extérieur})}$	14.6
Différences de moyennes, échantillons de paires si les différences sont normales	$t = \dfrac{d - \mu_d}{s_d}$, $D = n - 1$	12.7
Différences de moyennes, échantillons de paires, test du signe	$z = \dfrac{X \pm 0{,}5 - n/2}{\sqrt{n}/2}$	17.2
Différences de proportions	$z = \dfrac{(p_X - p_Y) - (P_X - P_Y)}{\sigma_{(p_X\ p_Y)}}$	
Variances si les deux populations sont normales	$F = \dfrac{{s_X}^2}{{s_Y}^2}$, $D_X = n_X - 1$, $D_Y = n_Y - 1$	14.3
Ajustement $D > 1$	$\begin{cases} {\chi_P}^2 = \Sigma \dfrac{(O-E)^2}{E} \\ D = (\text{nombres d'entrées } E) - 1 \end{cases}$	13.1

† Le facteur $\sqrt{\dfrac{N-n}{N-1}}$ a été omis de certaines formules. Voir sections 9.2 et 10.1

$D = 1$	$\chi_P^2 = \Sigma \frac{(\lvert O - E \rvert - 0{,}5)^2}{E}$		13.7
Les deux variables sont-elles indépendantes?			13.4
$D > 1$	$\begin{cases} \chi_P^2 = \Sigma \frac{(O - E)^2}{E} \\ D = (n-1)(m-1) \end{cases}$		
$D = 1$	$\chi_P^2 = \Sigma \frac{(\lvert O - E \rvert - 0{,}5)^2}{E}$		13.7
Droite de régression si les X et les Y sont distribués normalement avec la même variance	$t = \frac{b - B}{s_b}, D = n - 2$	$s_b = \sqrt{\frac{n(\Sigma Y^2 - a\Sigma Y - b\Sigma XY)}{(n-2)[n\Sigma X^2 - (\Sigma X)^2]}}$	15.12
Moyenne des Y pour X_0	$\begin{cases} t = \frac{Y_c - \mu_{Y X_0}}{s_{Y X_0}} \\ D = n - 2 \end{cases}$	$s_{Y/X_0} = s_e \sqrt{\frac{1}{n} + \frac{n(X_0 - \bar{X})^2}{n\Sigma X^2 - (\Sigma X)^2}}$ $s_e = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n-2}}$	16.3
Intervalle de prédiction pour Y pour un X_0 donné	$\begin{cases} t = \frac{Y_c - Y_p}{s_{Y_p}} \\ D = n - 2 \end{cases}$	$s_{Y_p} = s_e \sqrt{1 + \frac{1}{n} + \frac{n(X_0 - \bar{X})^2}{n\Sigma X^2 - (\Sigma X)^2}}$	16.3
Corrélation si X et Y sont distribués normalement	$H_0 : \rho = 0$	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, D = n - 2$	16.5

$H_0: \rho = A,\ A \neq 0$ $\quad z = \dfrac{Z - \mu_Z}{\sigma_Z} \quad \sigma_Z = \dfrac{1}{\sqrt{n-3}}$ 16.6

$H_0: \rho_S = 0$ (si $n > 10$) $\quad t = \dfrac{r_S\sqrt{n-2}}{\sqrt{1-r_S{}^2}},\ D = n-2$ 17.5

Échantillon aléatoire, test de séquences

$\mu_R = 1 + \dfrac{2n_Xn_Y}{n_X + n_Y} \quad \sigma_R = \sqrt{\dfrac{2n_Xn_Y(2n_Xn_Y - n_X - n_Y)}{(n_X + n_Y)^2(n_X + n_Y - 1)}}$ 17.3

Les populations sont-elles identiques?

Test de Mann-Whitney si les deux populations ont la même forme et la même variance

$U_X = n_Xn_Y + \dfrac{n_X(n_X + 1)}{2} - \Sigma R_X$
$= n_Xn_Y - U_Y$ 17.4

$U_Y = n_Xn_Y + \dfrac{n_Y(n_Y + 1)}{2} - \Sigma R_Y$
$= n_Xn_Y - U_X$
$U =$ le plus petit de U_X, U_Y

$\mu_U = \dfrac{n_Xn_Y}{2} \quad \sigma_U = \sqrt{\dfrac{n_Xn_Y(n_X + n_Y + 1)}{12}}$ 17.4

Appendice C

tables

Table 1 CARRÉS ET RACINES CARRÉES

	Carré	Racine carrée	
Si n est entre	Multiplier n par:	Utiliser la colonne	Multiplier par
0,000100 et 0,000999	0,00000001	$\sqrt{n}$	0,01
0,00100 et 0,00999	0,000001	$\sqrt{10n}$	0,01
0,0100 et 0,0999	0,0001	$\sqrt{n}$	0,1
0,100 et 0,999	0,01	$\sqrt{10n}$	0,1
1,00 et 9,99	1	$\sqrt{n}$	1
10,0 et 99,9	100	$\sqrt{10n}$	1
100 et 999	10 000	$\sqrt{n}$	10
1 000 et 9 990	1 000 000	$\sqrt{10n}$	10
10 000 et 99 900	100 000 000	$\sqrt{n}$	100
100 000 et 999 000	10 000 000 000	$\sqrt{10n}$	100
1 100 000 et 9 990 000	10 000 000 000	$\sqrt{n}$	1 000

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$
1,00	1,0000	1,000	3,162	1,25	1,5625	1,118	3,536	1,50	2,2500	1,225	3,873
1,01	1,0201	1,005	3,178	1,26	1,5876	1,122	3,550	1,51	2,2801	1,229	3,886
1,02	1,0404	1,010	3,194	1,27	1,6129	1,127	3,564	1,52	2,3104	1,233	3,899
1,03	1,0609	1,015	3,209	1,28	1,6384	1,131	3,578	1,53	2,3409	1,237	3,912
1,04	1,0816	1,020	3,225	1,29	1,6641	1,136	3,592	1,54	2,3716	1,241	3,924
1,05	1,1025	1,025	3,240	1,30	1,6900	1,140	3,606	1,55	2,4025	1,245	3,937
1,06	1,1236	1,030	3,256	1,31	1,7161	1,145	3,619	1,56	2,4336	1,249	3,950
1,07	1,1449	1,034	3,271	1,32	1,7424	1,149	3,633	1,57	2,4649	1,253	3,962
1,08	1,1664	1,039	3,286	1,33	1,7689	1,153	3,647	1,58	2,4964	1,257	3,975
1,09	1,1881	1,044	3,302	1,34	1,7956	1,158	3,661	1,59	2,5281	1,261	3,987
1,10	1,2100	1,049	3,317	1,35	1,8225	1,162	3,674	1,60	2,5600	1,266	4,000
1,11	1,2321	1,054	3,332	1,36	1,8496	1,166	3,688	1,61	2,5921	1,269	4,012
1,12	1,2544	1,058	3,347	1,37	1,8769	1,170	3,701	1,62	2,6244	1,273	4,025
1,13	1,2769	1,063	3,362	1,38	1,9044	1,175	3,715	1,63	2,6569	1,277	4,037
1,14	1,2996	1,068	3,376	1,39	1,9321	1,179	3,728	1,64	2,6896	1,281	4,050
1,15	1,3225	1,072	3,391	1,40	1,9600	1,183	3,742	1,65	2,7225	1,285	4,062
1,16	1,3456	1,077	3,406	1,41	1,9881	1,187	3,755	1,66	2,7556	1,288	4,074
1,17	1,3689	1,082	3,421	1,42	2,0164	1,192	3,768	1,67	2,7889	1,292	4,087
1,18	1,3924	1,086	3,435	1,43	2,0449	1,196	3,782	1,68	2,8224	1,296	4,099
1,19	1,4161	1,091	3,450	1,44	2,0736	1,200	3,795	1,69	2,8561	1,300	4,111
1,20	1,4400	1,095	3,464	1,45	2,1025	1,204	3,808	1,70	2,8900	1,304	4,123
1,21	1,4641	1,100	3,479	1,46	2,1316	1,208	3,821	1,71	2,9241	1,308	4,135
1,22	1,4884	1,105	3,493	1,47	2,1609	1,212	3,834	1,72	2,9584	1,311	4,147
1,23	1,5129	1,109	3,507	1,48	2,1904	1,217	3,847	1,73	2,9929	1,315	4,159
1,24	1,5376	1,114	3,521	1,49	2,2201	1,221	3,860	1,74	3,0276	1,319	4,171

Table 1 *(Suite)* CARRÉS ET RACINES CARRÉES

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$
1,75	3,0625	1,323	4,183	2,25	5,0625	1,500	4,743	2,75	7,5625	1,658	5,244
1,76	3,0976	1,327	4,195	2,26	5,1076	1,503	4,754	2,76	7,6176	1,661	5,254
1,77	3,1329	1,330	4,207	2,27	5,1529	1,507	4,764	2,77	7,6729	1,664	5,263
1,78	3,1684	1,334	4,219	2,28	5,1984	1,510	4,775	2,78	7,7284	1,667	5,273
1,79	3,2041	1,338	4,231	2,29	5,2441	1,513	4,785	2,79	7,7841	1,670	5,282
1,80	3,2400	1,342	4,243	2,30	5,2900	1,517	4,796	2,80	7,8400	1,673	5,292
1,81	3,2761	1,345	4,254	2,31	5,3361	1,520	4,806	2,81	7,8961	1,676	5,301
1,82	3,3124	1,349	4,266	2,32	5,3824	1,523	4,817	2,82	7,9524	1,679	5,310
1,83	3,3489	1,353	4,278	2,33	5,4289	1,526	4,827	2,83	8,0089	1,682	5,320
1,84	3,3856	1,356	4,290	2,34	5,4756	1,530	4,837	2,84	8,0656	1,685	5,329
1,85	3,4225	1,360	4,301	2,35	5,5225	1,533	4,848	2,85	8,1225	1,688	5,339
1,86	3,4596	1,364	4,313	2,36	5,5696	1,536	4,858	2,86	8,1796	1,691	5,348
1,87	3,4969	1,367	4,324	2,37	5,6169	1,539	4,868	2,87	8,2369	1,694	5,357
1,88	3,5344	1,371	4,336	2,38	5,6644	1,543	4,879	2,88	8,2944	1,697	5,367
1,89	3,5721	1,375	4,347	2,39	5,7121	1,546	4,889	2,89	8,3521	1,700	5,376
1,90	3,6100	1,378	4,359	2,40	5,7600	1,549	4,899	2,90	8,4100	1,703	5,385
1,91	3,6481	1,382	4,370	2,41	5,8081	1,552	4,909	2,91	8,4681	1,706	5,394
1,92	3,6864	1,386	4,382	2,42	5,8564	1,556	4,919	2,92	8,5264	1,709	5,404
1,93	3,7249	1,389	4,393	2,43	5,9049	1,559	4,930	2,93	8,5849	1,712	5,413
1,94	3,7636	1,393	4,405	2,44	5,9536	1,562	4,990	2,94	8,6436	1,715	5,422
1,95	3,8025	1,396	4,416	2,45	6,0025	1,565	4,950	2,95	8,7025	1,718	5,431
1,96	3,8416	1,400	4,427	2,46	6,0516	1,568	4,960	2,96	8,7616	1,720	5,441
1,97	3,8809	1,404	4,438	2,47	6,1009	1,572	4,970	2,97	8,8209	1,723	5,450
1,98	3,9204	1,407	4,450	2,48	6,1504	1,575	4,980	2,98	8,8804	1,726	5,459
1,99	3,9601	1,411	4,461	2,49	6,2001	1,578	4,990	2,99	8,9401	1,729	5,468
2,00	4,0000	1,414	4,472	2,50	6,2500	1,581	4,000	3,00	9,0000	1,732	5,477
2,01	4,0401	1,418	4,483	2,51	6,3001	1,584	4,010	3,01	9,0601	1,735	5,486
2,02	4,0804	1,421	4,494	2,52	6,3504	1,587	4,020	3,02	9,1204	1,738	5,495
2,03	4,1209	1,425	4,506	2,53	6,4009	1,591	4,030	3,03	9,1809	1,741	5,505
2,04	4,1616	1,428	4,517	2,54	6,4516	1,594	4,040	3,04	9,2416	1,744	5,514
2,05	4,2025	1,432	4,528	2,55	6,5025	1,597	4,050	3,05	9,3025	1,746	5,523
2,06	4,2436	1,435	4,539	2,56	6,5536	1,600	4,060	3,06	9,3636	1,749	5,532
2,07	4,2849	1,439	4,550	2,57	6,6049	1,603	4,070	3,07	9,4249	1,752	5,541
2,08	4,3264	1,442	4,561	2,58	6,6564	1,606	4,079	3,08	9,4864	1,755	5,550
2,09	4,3681	1,446	4,572	2,59	6,7081	1,609	4,089	3,09	9,5481	1,758	5,559
2,10	4,4100	1,449	4,583	2,60	6,7600	1,612	5,099	3,10	9,6100	1,761	5,568
2,11	4,4521	1,453	4,593	2,61	6,8121	1,616	5,109	3,11	9,6721	1,764	5,577
2,12	4,4944	1,456	4,604	2,62	6,8644	1,619	5,119	3,12	9,7344	1,766	5,586
2,13	4,5369	1,459	4,615	2,63	6,9169	1,622	5,128	3,13	9,7969	1,769	5,595
2,14	4,5796	1,463	4,626	2,64	6,9696	1,625	5,138	3,14	9,8596	1,772	5,604
2,15	4,6225	1,466	4,632	2,65	7,0225	1,628	5,148	3,15	9,9225	1,775	5,612
2,16	4,6656	1,470	4,648	2,66	7,0756	1,631	5,158	3,16	9,9856	1,778	5,621
2,17	4,7089	1,473	4,658	2,67	7,1289	1,634	5,167	3,17	10,0489	1,780	5,630
2,18	4,7524	1,476	4,669	2,68	7,1824	1,637	5,177	3,18	10,1124	1,783	5,639
2,19	4,7961	1,480	4,680	2,69	7,2361	1,640	5,187	3,19	10,1761	1,786	5,648
2,20	4,8400	1,483	4,690	2,70	7,2900	1,643	5,196	3,20	10,2400	1,789	5,657
2,21	4,8841	1,487	4,701	2,71	7,3441	1,646	5,206	3,21	10,3041	1,792	5,666
2,22	4,9284	1,490	4,712	2,72	7,3984	1,649	5,215	3,22	10,3684	1,794	5,674
2,23	4,9729	1,493	4,722	2,73	7,4529	1,652	5,225	3,23	10,4329	1,797	5,683
2,24	5,0176	1,497	4,733	2,74	7,5076	1,655	5,235	3,24	10,4976	1,800	5,692

Table 1 *(Suite)* CARRÉS ET RACINES CARRÉES

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$
3,25	10,5625	1,803	5,701	3,77	14,2129	1,942	6,140	4,29	18,4041	2,071	6,550
3,26	10,6276	1,806	5,710	3,78	14,2884	1,944	6,148	4,30	18,4900	2,074	6,557
3,27	10,6929	1,808	5,718	3,79	14,3641	1,947	6,156	4,31	18,5761	2,076	6,565
3,28	10,7584	1,811	5,727	3,80	14,4400	1,949	6,164	4,32	18,6624	2,078	6,573
3,29	10,8241	1,814	5,736	3,81	14,5161	1,952	6,173	4,33	18,7489	2,081	6,580
3,30	10,8900	1,817	5,745	3,82	14,5924	1,954	6,181	4,34	18,8356	2,083	6,588
3,31	10,9561	1,819	5,753	3,83	14,6689	1,957	6,189	4,35	18,9225	2,086	6,595
3,32	11,0224	1,822	5,762	3,84	14,7456	1,960	6,197	4,36	19,0096	2,088	6,603
3,33	11,0889	1,825	5,771	3,85	14,8225	1,962	6,205	4,37	19,0969	2,090	6,611
3,34	11,1556	1,828	5,779	3,86	14,8996	1,965	6,213	4,38	19,1844	2,093	6,618
3,35	11,2225	1,830	5,788	3,87	14,9769	1,967	6,221	4,39	19,2721	2,095	6,626
3,36	11,2896	1,833	5,797	3,88	15,0544	1,970	6,229	4,40	19,3600	2,098	6,633
3,37	11,3569	1,836	5,805	3,89	15,1321	1,972	6,237	4,41	19,4481	2,100	6,641
3,38	11,4244	1,838	5,814	3,90	15,2100	1,975	6,245	4,42	19,5364	2,102	6,648
3,39	11,4921	1,841	5,822	3,91	15,2881	1,977	6,253	4,43	19,6249	2,105	6,656
3,40	11,5600	1,844	5,831	3,92	15,3664	1,980	6,261	4,44	19,7136	2,107	6,663
3,41	11,6281	1,847	5,840	3,93	15,4449	1,982	6,269	4,45	19,8025	2,110	6,671
3,42	11,6964	1,849	5,848	3,94	15,5236	1,985	6,277	4,46	19,8916	2,112	6,678
3,43	11,7649	1,852	5,857	3,95	15,6025	1,987	6,285	4,47	19,9809	2,114	6,686
3,44	11,8336	1,855	5,865	3,96	15,6816	1,990	6,293	4,48	20,0704	2,117	6,693
3,45	11,9025	1,857	5,874	3,97	15,7609	1,992	6,301	4,49	20,1601	2,119	6,701
3,46	11,9716	1,860	5,882	3,98	15,8408	1,995	6,309	4,50	20,2500	2,121	6,708
3,47	12,0409	1,863	5,891	3,99	15,9201	1,997	6,317	4,51	20,3401	2,124	6,716
3,48	12,1104	1,865	5,899	4,00	16,0000	2,000	6,325	4,52	20,4304	2,126	6,723
3,49	12,1801	1,868	5,908	4,01	16,0801	2,002	6,332	4,53	20,5209	2,128	6,731
3,50	12,2500	1,871	5,916	4,02	16,1604	2,005	6,340	4,54	20,6116	2,131	6,738
3,51	12,3201	1,874	5,925	4,03	16,2409	2,007	6,348	4,55	20,7025	2,133	6,745
3,52	12,3904	1,876	5,933	4,04	16,3216	2,010	6,356	4,56	20,7936	2,135	6,753
3,53	12,4609	1,879	5,941	4,05	16,4025	2,012	6,364	4,57	20,8849	2,138	6,760
3,54	12,5316	1,881	5,950	4,06	16,4836	2,015	6,372	4,58	20,9764	2,140	6,768
3,55	12,6025	1,884	5,958	4,07	16,5649	2,017	6,380	4,59	21,0681	2,142	6,775
3,56	12,6736	1,887	5,967	4,08	16,6464	2,020	6,387	4,60	21,1600	2,145	6,782
3,57	12,7449	1,889	5,975	4,09	16,7281	2,022	6,395	4,61	21,2521	2,147	6,790
3,58	12,8164	1,892	5,983	4,10	16,8100	2,025	6,403	4,62	21,3444	2,149	6,797
3,59	12,8881	1,895	5,992	4,11	16,8921	2,027	6,411	4,63	21,4369	2,152	6,804
3,60	12,9600	1,897	6,000	4,12	16,9744	2,030	6,419	4,64	21,5296	2,154	6,812
3,61	13,0321	1,900	6,008	4,13	17,0569	2,032	6,427	4,65	21,6225	2,156	6,819
3,62	13,1044	1,903	6,017	4,14	17,1396	2,035	6,434	4,66	21,7156	2,159	6,826
3,63	13,1769	1,905	6,025	4,15	17,2225	2,037	6,442	4,67	21,8089	2,161	6,834
3,64	13,2496	1,908	6,033	4,16	17,3056	2,040	6,450	4,68	21,9024	2,163	6,841
3,65	13,3225	1,911	6,042	4,17	17,3889	2,042	6,458	4,69	21,9961	2,166	6,848
3,66	13,3956	1,913	6,050	4,18	17,4724	2,044	6,465	4,70	22,0900	2,168	6,856
3,67	13,4689	1,916	6,058	4,19	17,5561	2,047	6,473	4,71	22,1841	2,170	6,863
3,68	13,5424	1,918	6,066	4,20	17,6400	2,049	6,481	4,72	22,2784	2,173	6,870
3,69	13,6161	1,921	6,075	4,21	17,7241	2,052	6,488	4,73	22,3729	2,175	6,878
3,70	13,6900	1,924	6,083	4,22	17,8084	2,054	6,496	4,74	22,4676	2,177	6,885
3,71	13,7641	1,926	6,091	4,23	17,8929	2,057	6,504	4,75	22,5625	2,179	6,892
3,72	13,8384	1,929	6,099	4,24	17,9776	2,059	6,512	4,76	22,6576	2,182	6,899
3,73	13,9129	1,931	6,107	4,25	18,0625	2,062	6,519	4,77	22,7529	2,184	6,907
3,74	13,9876	1,934	6,116	4,26	18,1476	2,064	6,527	4,78	22,8484	2,186	6,914
3,75	14,0625	1,936	6,124	4,27	18,2329	2,066	6,535	4,79	22,9441	2,189	6,921
3,76	14,1376	1,939	6,132	4,28	18,3184	2,069	6,542	4,80	23,0400	2,191	6,928

Table 1 *(Suite)* CARRÉS ET RACINES CARRÉES

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$
4,81	23,1361	2,193	6,935								
4,82	23,2324	2,195	6,943								
4,83	23,3289	2,198	6,950	5,33	28,4089	2,309	7,301	5,81	33,7561	2,410	7,622
4,84	23,4256	2,200	6,957	5,34	28,5156	2,311	7,308	5,82	33,8724	2,412	7,629
4,85	23,5225	2,202	6,964	5,35	28,6225	2,313	7,314	5,83	33,9889	2,415	7,635
4,86	23,6196	2,205	6,971	5,36	28,7296	2,315	7,321	5,84	34,1056	2,417	7,642
4,87	23,7169	2,207	6,979	5,37	28,8369	2,317	7,328	5,85	34,2225	2,419	7,648
4,88	23,8144	2,209	6,986	5,38	28,9444	2,319	7,335	5,86	34,3396	2,421	7,655
4,89	23,9121	2,211	6,993	5,39	29,0521	2,322	7,342	5,87	34,4569	2,423	7,662
4,90	24,0100	2,214	7,000	5,40	29,1600	2,324	7,348	5,88	34,5744	2,425	7,668
4,91	24,1081	2,216	7,007	5,41	29,2681	2,326	7,355	5,89	34,6921	2,427	7,675
4,92	24,2064	2,218	7,014	5,42	29,3764	2,328	7,362	5,90	34,8100	2,429	7,681
4,93	24,3049	2,220	7,021	5,43	29,4849	2,330	7,369	5,91	34,9281	2,431	7,688
4,94	24,4036	2,223	7,029	5,44	29,5936	2,332	7,376	5,92	35,0464	2,433	7,694
4,95	24,5025	2,225	7,036	5,45	29,7025	2,334	7,382	5,93	35,1649	2,435	7,701
4,96	24,6016	2,227	7,043	5,46	29,8116	2,337	7,389	5,94	35,2836	2,437	7,707
4,97	24,7009	2,229	7,050	5,47	29,9209	2,339	7,396	5,95	35,4025	2,439	7,714
4,98	24,8004	2,232	7,057	5,48	30,0304	2,341	7,403	5,96	35,5216	2,441	7,720
4,99	24,9001	2,234	7,064	5,49	30,1401	2,343	7,408	5,97	35,6409	2,443	7,727
5,00	25,0000	2,236	7,071	5,50	30,2500	2,345	7,416	5,98	35,7604	2,445	7,733
5,01	25,1001	2,238	7,078	5,51	30,3601	2,347	7,423	5,99	35,8801	2,447	7,740
5,02	25,2004	2,241	7,085	5,52	30,4704	2,349	7,430	6,00	36,0000	2,449	7,746
5,03	25,3009	2,243	7,092	5,53	30,5809	2,352	7,436	6,01	36,1201	2,452	7,752
5,04	25,4016	2,245	7,099	5,54	30,6916	2,354	7,443	6,02	36,2404	2,454	7,759
5,05	25,5025	2,247	7,106	5,55	30,8025	2,356	7,450	6,03	36,3609	2,456	7,765
5,06	25,6036	2,249	7,113	5,56	30,9136	2,358	7,457	6,04	36,4816	2,458	7,772
5,07	25,7049	2,252	7,120	5,57	31,0249	2,360	7,463	6,05	36,6025	2,460	7,778
5,08	25,8064	2,254	7,127	5,58	31,1364	2,362	7,470	6,06	36,7236	2,462	7,785
5,09	25,9081	2,256	7,134	5,59	31,2481	2,364	7,477	6,07	36,8449	2,464	7,791
5,10	26,0100	2,258	7,141	5,60	31,3600	2,366	7,483	6,08	36,9664	2,466	7,797
5,11	26,1121	2,261	7,148	5,61	31,4721	2,369	7,490	6,09	37,0881	2,468	7,804
5,12	26,2144	2,263	7,155	5,62	31,5844	2,371	7,497	6,10	37,2100	2,470	7,810
5,13	26,3169	2,265	7,162	5,63	31,6969	2,373	7,503	6,11	37,3321	2,472	7,817
5,14	26,4196	2,267	7,169	5,64	31,8096	2,375	7,510	6,12	37,4544	2,474	7,823
5,15	26,5225	2,269	7,176	5,65	31,9225	2,377	7,517	6,13	37,5769	2,476	7,829
5,16	26,6256	2,272	7,183	5,66	32,0356	2,379	7,523	6,14	37,6996	2,478	7,836
5,17	26,7289	2,274	7,190	5,67	32,1489	2,381	7,530	6,15	37,8225	2,480	7,842
5,18	26,8324	2,276	7,197	5,68	32,2624	2,383	7,537	6,16	37,9456	2,482	7,849
5,19	26,9361	2,278	7,204					6,17	38,0689	2,484	7,853
5,20	27,0400	2,280	7,211					6,18	38,1924	2,486	7,861
5,21	27,1441	2,283	7,218	5,69	32,3761	2,385	7,543	6,19	38,3161	2,488	7,868
5,22	27,2484	2,285	7,225	5,70	32,4900	2,387	7,550	6,20	38,4400	2,490	7,874
5,23	27,3529	2,287	7,232	5,71	32,6041	2,391	7,556	6,21	38,5641	2,492	7,880
5,24	27,4576	2,289	7,239	5,72	32,7184	2,392	7,563	6,22	38,6884	2,494	7,887
5,25	27,5625	2,291	7,246	5,73	32,8329	2,394	7,570	6,23	38,8129	2,496	7,893
5,26	27,6676	2,293	7,253	5,74	32,9476	2,396	7,576	6,24	38,9376	2,498	7,899
5,27	27,7729	2,296	7,259	5,75	33,0625	2,398	7,583	6,25	39,0625	2,500	7,906
5,28	27,8784	2,298	7,266	5,76	33,1776	2,400	7,589	6,26	39,1876	2,502	7,912
5,29	27,9841	2,300	7,273	5,77	33,2929	2,402	7,596				
5,30	28,0900	2,302	7,280	5,78	33,4084	2,404	7,603	6,27	39,3129	2,504	7,918
5,31	28,1961	2,304	7,287	5,79	33,5241	2,406	7,609	6,28	39,4384	2,506	7,925
5,32	28,3024	2,307	7,294	5,80	33,6400	2,408	7,616	6,29	39,5641	2,508	7,931

Table 1 *(Suite)* CARRÉS ET RACINES CARRÉES

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$
6,30	39,6900	2,510	7,937	6,80	46,2400	2,608	8,246	7,30	53,2900	2,702	8,544
6,31	39,8161	2,512	7,944	6,81	46,3761	2,610	8,252	7,31	53,4361	2,704	8,550
6,32	39,9424	2,514	7,950	6,82	46,5124	2,612	8,258	7,32	53,5824	2,706	8,556
6,33	40,0689	2,516	7,956	6,83	46,6489	2,613	8,264	7,33	53,7289	2,707	8,562
6,34	40,1956	2,518	7,962	6,84	46,7856	2,615	8,270	7,34	53,8756	2,709	8,567
6,35	40,3225	2,520	7,969	6,85	46,9225	2,617	8,276	7,35	54,0225	2,711	8,573
6,36	40,4496	2,522	7,975	6,86	47,0596	2,619	8,283	7,36	54,1696	2,713	8,579
6,37	40,5769	2,524	7,981	6,87	47,1969	2,621	8,289	7,37	54,3169	2,715	8,585
6,38	40,7044	2,526	7,987	6,88	47,3344	2,623	8,295	7,38	54,4644	2,717	8,591
6,39	40,8321	2,528	7,994	6,89	47,4721	2,625	8,301	7,39	54,6121	2,718	8,597
6,40	40,9600	2,530	8,000	6,90	47,6100	2,627	8,307	7,40	54,7600	2,720	8,602
6,41	41,0881	2,532	8,006	6,91	47,7481	2,629	8,313	7,41	54,9081	2,722	8,608
6,42	41,2164	2,534	8,013	6,92	47,8864	2,631	8,319	7,42	55,0564	2,724	8,614
6,43	41,3449	2,536	8,019	6,93	48,0249	2,632	8,325				
6,44	41,4736	2,538	8,025	6,94	48,1636	2,634	8,331				
6,45	41,6025	2,540	8,031	6,95	48,3025	2,636	8,337	7,43	55,2049	2,726	8,620
6,46	41,7316	2,542	8,037	6,96	48,4416	2,638	8,343	7,44	55,3536	2,728	8,626
6,47	41,8609	2,544	8,044	6,97	48,5809	2,640	8,349	7,45	55,5025	2,729	8,631
6,48	41,9904	2,546	8,050	6,98	48,7204	2,642	8,355	7,46	55,6516	2,731	8,637
6,49	42,1201	2,548	8,056	6,99	48,8601	2,644	8,361	7,47	55,8009	2,733	8,643
6,50	42,2500	2,550	8,062	7,00	49,0000	2,646	8,367	7,48	55,9504	2,735	8,649
6,51	42,3801	2,551	8,068	7,01	49,1401	2,648	8,373	7,49	56,1001	2,737	8,654
6,52	42,5104	2,553	8,075	7,02	49,2804	2,650	8,379	7,50	56,2500	2,739	8,660
6,53	42,6409	2,555	8,081	7,03	49,4209	2,651	8,385	7,51	56,4001	2,740	8,666
6,54	42,7716	2,557	8,087	7,04	49,5616	2,653	8,390	7,52	56,5504	2,742	8,672
6,55	42,9025	2,559	8,093	7,05	49,7025	2,655	8,396	7,53	56,7009	2,744	8,678
6,56	43,0336	2,561	8,099	7,06	49,8436	2,657	8,402	7,54	56,8516	2,746	8,688
6,57	43,1649	2,563	8,106	7,07	49,9849	2,659	8,408	7,55	57,0025	2,748	8,689
6,58	43,2964	2,565	8,112	7,08	50,1264	2,661	8,414	7,56	57,1536	2,750	8,695
6,59	43,4281	2,567	8,118	7,09	50,2681	2,663	8,420	7,57	57,3049	2,751	8,701
6,60	43,5600	2,569	8,124	7,10	50,4100	2,665	8,426	7,58	57,4564	2,753	8,706
6,61	43,6921	2,571	8,130	7,11	50,5521	2,666	8,432	7,59	57,6081	2,755	8,712
6,62	43,8244	2,573	8,136	7,12	50,6944	2,668	8,438	7,60	57,7600	2,757	8,718
6,63	43,9569	2,575	8,142	7,13	50,8369	2,670	8,444	7,61	57,9121	2,759	8,724
6,64	44,0896	2,577	8,149	7,14	50,9796	2,672	8,450	7,62	58,0644	2,760	8,729
6,65	44,2225	2,579	8,155	7,15	51,1225	2,674	8,456	7,63	58,2169	2,762	8,735
6,66	44,3556	2,581	8,161	7,16	51,2656	2,676	8,462	7,64	58,3696	2,764	8,741
6,67	44,4889	2,583	8,167	7,17	51,4089	2,678	8,468	7,65	58,5225	2,766	8,746
6,68	44,6224	2,585	8,173	7,18	51,5524	2,680	8,473	7,66	58,6756	2,768	8,752
6,69	44,7561	2,587	8,179	7,19	51,6961	2,681	8,479	7,67	58,8289	2,769	8,758
6,70	44,8900	2,588	8,185	7,20	51,8400	2,683	8,485	7,68	58,9824	2,771	8,764
6,71	45,0241	2,590	8,191	7,21	51,9841	2,685	8,491	7,69	59,1361	2,773	8,769
6,72	45,1584	2,592	8,198	7,22	52,1284	2,687	8,497	7,70	59,2900	2,775	8,775
6,73	45,2929	2,594	8,204	7,23	52,2729	2,689	8,503	7,71	59,4441	2,777	8,781
6,74	45,4276	2,596	8,210	7,24	52,4176	2,691	8,509	7,72	59,5984	2,778	8,786
6,75	45,5625	2,598	8,216	7,25	52,5625	2,693	8,515	7,73	59,7529	2,780	8,792
6,76	45,6976	2,600	8,222	7,26	52,7076	2,694	8,521	7,74	59,9076	2,782	8,798
6,77	45,8329	2,602	8,228	7,27	52,8529	2,696	8,526	7,75	60,0625	2,784	8,803
6,78	45,9684	2,604	8,234	7,28	52,9984	2,698	8,532	7,76	60,2176	2,786	8,809
6,79	46,1041	2,606	8,240	7,29	53,1441	2,700	8,538	7,77	60,3729	2,787	8,815

Table 1 *(Suite)* CARRÉS ET RACINES CARRÉES

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$
7,78	60,5284	2,789	8,820	8,28	68,5584	2,878	9,010	8,78	77,0884	2,963	9,370
7,79	60,6841	2,791	8,826	8,29	68,7241	2,879	9,105	8,79	77,2641	2,965	9,376
7,80	60,8400	2,793	8,832	8,30	68,8900	2,881	9,110	8,80	77,4400	2,966	9,381
7,81	60,9961	2,795	8,837								
7,82	61,1524	2,796	8,843								
7,83	61,3089	2,798	8,849	8,31	69,0561	2,883	9,116	8,81	77,6161	2,968	9,386
7,84	61,4656	2,800	8,854	8,32	69,2224	2,884	9,121	8,82	77,7924	2,970	9,391
7,85	61,6225	2,802	8,860	8,33	69,3889	2,886	9,127	8,83	77,9689	2,972	9,397
7,86	61,7796	2,804	8,866	8,34	69,5556	2,888	9,132	8,84	78,1456	2,973	9,402
7,87	61,9369	2,805	8,871	8,35	69,7225	2,890	9,138	8,85	78,3225	2,975	9,407
7,88	62,0944	2,807	8,877	8,36	69,8896	2,891	9,143	8,86	78,4996	2,977	9,413
7,89	62,2521	2,809	8,883	8,37	70,0569	2,893	9,149	8,87	78,6769	2,978	9,418
7,90	62,4100	2,811	8,888	8,38	70,2244	2,895	9,154	8,88	78,8544	2,980	9,423
7,91	62,5681	2,812	8,894	8,39	70,3921	2,897	9,160	8,89	79,0321	2,982	9,429
7,92	62,7264	2,814	8,899	8,40	70,5600	2,898	9,165	8,90	79,2100	2,983	9,434
7,93	62,8849	2,816	8,905	8,41	70,7281	2,900	9,167	8,91	79,3881	2,985	9,439
7,94	63,0436	2,818	8,911	8,42	70,8964	2,902	9,176	8,92	79,5664	2,987	9,445
7,95	63,2025	2,820	8,916	8,43	71,0649	2,903	9,182	8,93	79,7449	2,988	9,450
7,96	63,3616	2,821	8,922	8,44	71,2336	2,905	9,187	8,94	79,9236	2,990	9,455
7,97	63,5209	2,823	8,927	8,45	71,4025	2,907	9,192	8,95	80,1025	2,992	9,460
7,98	63,6804	2,825	8,933	8,46	71,5716	2,909	9,198	8,96	80,2816	2,993	9,466
7,99	63,8401	2,827	8,939	8,47	71,7409	2,910	9,203	8,97	80,4609	2,995	9,471
8,00	64,0000	2,828	8,944	8,48	71,9104	2,912	9,209	8,98	80,6404	2,997	9,476
				8,49	72,0801	2,914	9,214	8,99	80,8201	2,998	9,482
				8,50	72,2500	2,915	9,220	9,00	81,0000	3,000	9,487
8,01	64,1601	2,830	8,950	8,51	72,4201	2,917	9,225	9,01	81,1801	3,001	9,492
8,02	64,3204	2,832	8,955	8,52	72,5904	2,919	9,230	9,02	81,3604	3,003	9,497
8,03	64,4809	2,834	8,961	8,53	72,7609	2,921	9,236	9,03	81,5409	3,005	9,503
8,04	64,6416	2,835	8,967	8,54	72,9316	2,922	9,241	9,04	81,7216	3,007	9,508
8,05	64,8025	2,837	8,972	8,55	73,1025	2,924	9,247	9,05	81,9025	3,008	9,513
8,06	64,9636	2,839	8,978	8,56	73,2736	2,926	9,252	9,06	82,0836	3,010	9,518
8,07	65,1249	2,841	8,983	8,57	73,4449	2,927	9,257	9,07	82,2649	3,012	9,524
8,08	65,2864	2,843	8,989	8,58	73,6164	2,929	9,263	9,08	82,4464	3,013	9,529
8,09	65,4481	2,844	8,994	8,59	73,7881	2,931	9,268	9,09	82,6281	3,015	9,534
8,10	65,6100	2,846	9,000	8,60	73,9600	2,933	9,274	9,10	82,8100	3,017	9,539
8,11	65,7721	2,848	9,006	8,61	74,1321	2,934	9,279				
8,12	65,9344	2,850	9,011	8,62	74,3044	2,936	9,284				
8,13	66,0969	2,851	9,017	8,63	74,4769	2,938	9,290				
8,14	66,2596	2,853	9,022	8,64	74,6496	2,939	9,295	9,11	82,9921	3,018	9,545
8,15	66,4225	2,855	9,028	8,65	74,8225	2,941	9,301	9,12	83,1744	3,020	9,550
8,16	66,5856	2,857	9,033	8,66	74,9956	2,943	9,306	9,13	83,3569	3,022	9,555
8,17	66,7489	2,858	9,039	8,67	75,1689	2,944	9,311	9,14	83,5396	3,023	9,560
8,18	66,9124	2,860	9,044	8,68	75,3424	2,946	9,317	9,15	83,7225	3,025	9,566
8,19	67,0761	2,862	9,050	8,69	75,5161	2,948	9,322	9,16	83,9056	3,027	9,571
8,20	67,2400	2,864	9,055	8,70	75,6900	2,950	9,327	9,17	84,0889	3,028	9,576
8,21	67,4041	2,865	9,061	8,71	75,8641	2,951	9,333	9,18	84,2724	3,030	9,581
8,22	67,5684	2,867	9,066	8,72	76,0384	2,953	9,338	9,19	84,4561	3,032	9,586
8,23	67,7329	2,869	9,072	8,73	76,2129	2,955	9,343	9,20	84,6400	3,033	9,592
8,24	67,8976	2,871	9,077	8,74	76,3876	2,956	9,349	9,21	84,8241	3,035	9,597
8,25	68,0625	2,872	9,083	8,75	76,5625	2,958	9,354	9,22	85,0084	3,036	9,602
8,26	68,2276	2,874	9,088	8,76	76,7376	2,960	9,359	9,23	85,1929	3,038	9,607
8,27	68,3929	2,876	9,094	8,77	76,9129	2,961	9,365	9,24	85,3776	3,040	9,612

Table 1 *(Suite)* CARRÉS ET RACINES CARRÉES

n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$	n	n^2	$\sqrt{n}$	$\sqrt{10n}$
9,25	85,5625	3,041	9,618								
9,26	85,7476	3,043	9,623								
9,27	85,9329	3,045	9,628								
9,28	86,1184	3,046	9,633	9,53	90,8209	3,087	9,762	9,75	95,0625	3,123	9,874
9,29	86,3041	3,048	9,638	9,54	91,0116	3,089	9,767	9,76	95,2576	3,124	9,879
9,30	86,4900	3,050	9,644	9,55	91,2025	3,090	9,772	9,77	95,4529	3,126	9,884
9,31	86,6761	3,051	9,649	9,56	91,3936	3,092	9,778	9,78	95,6484	3,127	9,889
9,32	86,8624	3,053	9,654	9,57	91,5849	3,093	9,783	9,79	95,8441	3,129	9,894
9,33	87,0489	3,055	9,659	9,58	91,7764	3,095	9,788	9,80	96,0400	3,131	9,899
9,34	87,2356	3,056	9,664	9,59	91,9681	3,097	9,793	9,81	96,2361	3,132	9,905
9,35	87,4225	3,058	9,670	9,60	92,1600	3,098	9,798	9,82	96,4324	3,134	9,910
9,36	87,6096	3,059	9,675	9,61	92,3521	3,100	9,803	9,83	96,6289	3,135	9,915
9,37	87,7969	3,061	9,680	9,62	92,5444	3,102	9,808	9,84	96,8256	3,137	9,920
9,38	87,9844	3,063	9,685					9,85	97,0225	3,138	9,925
9,39	88,1721	3,064	9,690					9,86	97,2196	3,140	9,930
9,40	88,3600	3,066	9,695					9,87	97,4169	3,142	9,935
9,41	88,5481	3,068	9,701	9,63	92,7369	3,103	9,813	9,88	97,6144	3,143	9,940
9,42	88,7364	3,069	9,706	9,64	92,9296	3,105	9,818	9,89	97,8121	3,145	9,945
9,43	88,9249	3,071	9,711	9,65	93,1225	3,106	9,823	9,90	98,0100	3,146	9,950
9,44	89,1136	3,072	9,716	9,66	93,3156	3,108	9,829	9,91	98,2081	3,148	9,955
9,45	89,3025	3,074	9,721	9,67	93,5089	3,110	9,834	9,92	98,4064	3,150	9,960
9,46	89,4916	3,076	9,726	9,68	93,7024	3,111	9,839	9,93	98,6049	3,151	9,965
9,47	89,6809	3,077	9,731	9,69	93,8961	3,113	9,844	9,94	98,8036	3,153	9,970
9,48	89,8704	3,079	9,737	9,70	94,0900	3,114	9,849	9,95	99,0025	3,154	9,975
9,49	90,0601	3,081	9,742	9,71	94,2841	3,116	9,854	9,96	99,2016	3,156	9,980
9,50	90,2500	3,082	9,747	9,72	94,4784	3,118	9,859	9,97	99,4009	3,157	9,985
9,51	90,4401	3,084	9,752	9,73	94,6729	3,119	9,864	9,98	99,6004	3,159	9,990
9,52	90,6304	3,085	9,757	9,74	94,8676	3,121	9,869	9,99	99,8001	3,161	9,995

TABLE 2 DISTRIBUTION BINÔMIALE

$${}_nC_x p^x q^{n-x}$$

n	X	p = 0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99
2	0	980	902	810	640	490	360	250	160	090	040	010	002	0+
	1	020	095	180	320	420	480	500	480	420	320	180	095	020
	2	0+	002	010	040	090	160	250	360	490	640	810	902	980
3	0	970	857	729	512	343	216	125	064	027	008	001	0+	0+
	1	029	135	243	384	441	432	375	288	189	096	027	007	0+
	2	0+	007	027	096	189	288	375	432	441	384	243	135	029
	3	0+	0+	001	008	027	064	125	216	343	512	729	857	970
4	0	961	815	656	410	240	130	062	026	008	002	0+	0+	0+
	1	039	171	292	410	412	346	250	154	076	026	004	0+	0+
	2	001	014	049	154	265	346	375	346	265	154	049	014	001
	3	0+	0+	004	026	076	154	250	346	412	410	292	171	039
	4	0+	0+	0+	002	008	026	062	130	240	410	656	815	961
5	0	951	774	590	328	168	078	031	010	002	0+	0+	0+	0+
	1	048	204	328	410	360	259	156	077	028	006	0+	0+	0+
	2	001	021	073	205	309	346	312	230	132	051	008	001	0+
	3	0+	001	008	051	132	230	312	346	309	205	073	021	001
	4	0+	0+	0+	006	028	077	156	259	360	410	328	204	048
	5	0+	0+	0+	0+	002	010	031	078	168	328	590	774	951
6	0	941	735	531	262	118	047	016	004	001	0+	0+	0+	0+
	1	057	232	354	393	303	187	094	037	010	002	0+	0+	0+
	2	001	031	098	246	324	311	234	138	060	015	001	0+	0+
	3	0+	002	015	082	185	276	312	276	185	082	015	002	0+
	4	0+	0+	001	015	060	138	234	311	324	246	098	031	001
	5	0+	0+	0+	002	010	037	094	187	303	393	354	232	057
	6	0+	0+	0+	0+	001	004	016	047	118	262	531	735	941
7	0	932	698	478	210	082	028	008	002	0+	0+	0+	0+	0+
	1	066	257	372	367	247	131	055	017	004	0+	0+	0+	0+
	2	002	041	124	275	318	261	164	077	025	004	0+	0+	0+
	3	0+	004	023	115	227	290	273	194	097	029	003	0+	0+
	4	0+	0+	003	029	097	194	273	290	227	115	023	004	0+
	5	0+	0+	0+	004	025	077	164	261	318	275	124	041	002
	6	0+	0+	0+	0+	004	017	055	131	247	367	372	257	066
	7	0+	0+	0+	0+	0+	002	008	028	082	210	478	698	932
8	0	923	663	430	168	058	017	004	001	0+	0+	0+	0+	0+
	1	075	279	383	336	198	090	031	008	001	0+	0+	0+	0+
	2	003	051	149	294	296	209	109	041	010	001	0+	0+	0+
	3	0+	005	033	147	254	279	219	124	047	009	0+	0+	0+
	4	0+	0+	005	046	136	232	273	232	136	046	005	0+	0+
	5	0+	0+	0+	009	047	124	219	279	254	147	033	005	0+
	6	0+	0+	0+	001	010	041	109	209	296	294	149	051	003
	7	0+	0+	0+	0+	001	008	031	090	198	336	383	279	075
	8	0+	0+	0+	0+	0+	001	004	017	058	168	430	663	923
9	0	914	630	387	134	040	010	002	0+	0+	0+	0+	0+	0+
	1	083	299	387	302	156	060	018	004	0+	0+	0+	0+	0+
	2	003	063	172	302	267	161	070	021	004	0+	0+	0+	0+
	3	0+	008	045	176	267	251	164	074	021	003	0+	0+	0+
	4	0+	001	007	066	172	251	246	167	074	017	001	0+	0+
	5	0+	0+	001	017	074	167	246	251	172	066	007	001	0+

TABLE 2 (*Suite*) DISTRIBUTION BINÔMIALE

		p												
n	*X*	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99
	6	0+	0+	0+	003	021	074	164	251	267	176	045	008	0+
	7	0+	0+	0+	0+	004	021	070	161	267	302	172	063	003
	8	0+	0+	0+	0+	0+	004	018	060	156	302	387	299	083
	9	0+	0+	0+	0+	0+	0+	002	010	040	134	387	630	914
10	0	904	599	349	107	028	006	001	0+	0+	0+	0+	0+	0+
	1	091	315	387	268	121	040	010	002	0+	0+	0+	0+	0+
	2	004	075	194	302	233	121	044	011	001	0+	0+	0+	0+
	3	0+	010	057	201	267	215	117	042	009	001	0+	0+	0+
	4	0+	001	011	088	200	251	205	111	037	006	0+	0+	0+
	5	0+	0+	001	026	103	201	246	201	103	026	001	0+	0+
	6	0+	0+	0+	006	037	111	205	251	200	088	011	001	0+
	7	0+	0+	0+	001	009	042	117	215	267	201	057	010	0+
	8	0+	0+	0+	0+	001	011	044	121	233	302	194	075	004
	9	0+	0+	0+	0+	0+	002	010	040	121	268	387	315	091
	10	0+	0+	0+	0+	0+	0+	001	006	028	107	349	599	904
11	0	895	569	314	086	020	004	0+	0+	0+	0+	0+	0+	0+
	1	099	329	384	236	093	027	005	001	0+	0+	0+	0+	0+
	2	005	087	213	295	200	089	027	005	001	0+	0+	0+	0+
	3	0+	014	071	221	257	177	081	023	004	0+	0+	0+	0+
	4	0+	001	016	111	220	236	161	070	017	002	0+	0+	0+
	5	0+	0+	002	039	132	221	226	147	057	010	0+	0+	0+
	6	0+	0+	0+	010	057	147	226	221	132	039	002	0+	0+
	7	0+	0+	0+	002	017	070	161	236	220	111	016	001	0+
	8	9+	0+	0+	0+	004	023	081	177	257	221	071	014	0+
	9	0+	0+	0+	0+	001	005	027	089	200	295	213	087	005
	10	0+	0+	0+	0+	0+	001	005	027	093	236	384	329	099
	11	0+	0+	0+	0+	0+	0+	0+	004	020	086	314	569	895
12	0	886	540	282	069	014	002	0+	0+	0+	0+	0+	0+	0+
	1	107	341	377	206	071	017	003	0+	0+	0+	0+	0+	0+
	2	006	099	230	283	168	064	016	002	0+	0+	0+	0+	0+
	3	0+	017	085	236	240	142	054	012	001	0+	0+	0+	0+
	4	0+	002	021	133	231	213	121	042	008	001	0+	0+	0+
	5	0+	0+	004	053	158	227	193	101	029	003	0+	0+	0+
	6	0+	0+	0+	016	079	177	226	177	079	016	0+	0+	0+
	7	0+	0+	0+	003	029	101	193	227	158	053	004	0+	0+
	8	0+	0+	0+	001	008	042	121	213	231	133	021	002	0+
	9	0+	0+	0+	0+	001	012	054	142	240	236	085	017	0+
	10	0+	0+	0+	0+	0+	002	016	064	168	283	230	099	006
	11	0+	0+	0+	0+	0+	0+	003	017	071	206	377	341	107
	12	0+	0+	0+	0+	0+	0+	0+	002	014	069	282	540	886
13	0	878	513	254	055	010	001	0+	0+	0+	0+	0+	0+	0+
	1	115	351	367	179	054	011	002	0+	0+	0+	0+	0+	0+
	2	007	111	245	268	139	045	010	001	0+	0+	0+	0+	0+
	3	0+	021	100	246	218	111	035	006	001	0+	0+	0+	0+
	4	0+	003	028	154	234	184	087	024	003	0+	0+	0+	0+
	5	0+	0+	006	069	180	221	157	066	014	001	0+	0+	0+
	6	0+	0+	001	023	103	197	209	131	044	006	0+	0+	0+
	7	0+	0+	0+	006	044	131	209	197	103	023	001	0+	0+
	8	0+	0+	0+	001	014	066	157	221	180	069	006	0+	0+
	9	9+	0+	0+	0+	003	024	087	184	234	154	028	003	0+

TABLE 2 (*Suite*) DISTRIBUTION BINÔMIALE

		p												
n	*X*	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99
	10	0+	0+	0+	0+	001	006	035	111	218	246	100	021	0+
	11	0+	0+	0+	0+	0+	001	010	045	139	268	245	111	007
	12	0+	0+	0+	0+	0+	0+	002	011	054	179	367	351	115
	13	0+	0+	0+	0+	0+	0+	0+	001	010	055	254	513	878
14	0	869	488	229	044	007	001	0+	0+	0+	0+	0+	0+	0+
	1	123	359	356	154	041	007	001	0+	0+	0+	0+	0+	0+
	2	008	123	257	250	113	032	006	001	0+	0+	0+	0+	0+
	3	0+	026	114	250	194	085	022	003	0+	0+	0+	0+	0+
	4	0+	004	035	172	229	155	061	014	001	0+	0+	0+	0+
	5	0+	0+	008	086	196	207	122	041	007	0+	0+	0+	0+
	6	0+	0+	001	032	126	207	183	092	023	002	0+	0+	0+
	7	0+	0+	0+	009	062	157	209	157	062	009	0+	0+	0+
	8	0+	0+	0+	002	023	092	183	207	126	032	001	0+	0+
	9	0+	0+	0+	0+	007	041	122	207	196	086	008	0+	0+
	10	0+	0+	0+	0+	001	014	061	155	229	172	035	004	0+
	11	0+	0+	0+	0+	0+	003	022	085	194	250	114	026	0+
	12	0+	0+	0+	0+	0+	001	006	032	113	250	257	123	008
	13	0+	0+	0+	0+	0+	0+	001	007	041	154	356	359	123
	14	0+	0+	0+	0+	0+	0+	0+	001	007	044	229	488	869
15	0	860	463	206	035	005	0+	0+	0+	0+	0+	0+	0+	0+
	1	130	366	343	132	031	005	0+	0+	0+	0+	0+	0+	0+
	2	009	135	267	231	092	022	003	0+	0+	0+	0+	0+	0+
	3	0+	031	129	250	170	063	014	002	0+	0+	0+	0+	0+
	4	0+	005	043	188	219	127	042	007	001	0+	0+	0+	0+
	5	0+	001	010	103	206	186	092	024	003	0+	0+	0+	0+
	6	0+	0+	002	043	147	207	153	061	012	001	0+	0+	0+
	7	0+	0+	0+	014	081	177	196	118	035	003	0+	0+	0+
	8	0+	0+	0+	003	035	118	196	177	081	014	0+	0+	0+
	9	0+	0+	0+	001	012	061	153	207	147	043	002	0+	0+
	10	0+	0+	0+	0+	003	024	092	186	206	103	010	001	0+
	11	0+	0+	0+	0+	001	007	042	127	219	188	043	005	0+
	12	0+	0+	0+	0+	0+	002	014	063	170	250	129	031	0+
	13	0+	0+	0+	0+	0+	0+	003	022	092	231	267	135	009
	14	0+	0+	0+	0+	0+	0+	0+	005	031	132	343	366	130
	15	0+	0+	0+	0+	0+	0+	0+	0+	005	035	206	463	860
16	0	851	440	185	028	003	0+	0+	0+	0+	0+	0+	0+	0+
	1	138	371	329	113	023	003	0+	0+	0+	0+	0+	0+	0+
	2	010	146	274	211	073	015	002	0+	0+	0+	0+	0+	0+
	3	0+	036	142	246	146	047	008	001	0+	0+	0+	0+	0+
	4	0+	006	051	200	204	101	028	004	0+	0+	0+	0+	0+
	5	0+	001	014	120	210	162	067	014	001	0+	0+	0+	0+
	6	0+	0+	003	055	165	198	122	039	006	0+	0+	0+	0+
	7	0+	0+	0+	020	101	189	175	084	018	001	0+	0+	0+
	8	0+	0+	0+	006	049	142	196	142	049	006	0+	0+	0+
	9	0+	0+	0+	001	018	084	175	189	101	020	0+	0+	0+
	10	0+	0+	0+	0+	006	039	122	198	165	055	003	0+	0+
	11	0+	0+	0+	0+	001	014	067	162	210	120	014	001	0+
	12	0+	0+	0+	0+	0+	004	028	101	204	200	051	006	0+
	13	0+	0+	0+	0+	0+	001	008	047	146	246	142	036	0+
	14	0+	0+	0+	0+	0+	0+	002	015	073	211	274	146	010
	15	0+	0+	0+	0+	0+	0+	0+	003	023	113	329	371	138
	16	0+	0+	0+	0+	0+	0+	0+	0+	003	028	185	440	851

TABLE 2 (*Suite*) DISTRIBUTION BINÔMIALE

		p												
n	X	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99
17	0	843	418	167	022	002	0+	0+	0+	0+	0+	0+	0+	0+
	1	145	374	315	096	017	002	0+	0+	0+	0+	0+	0+	0+
	2	012	158	280	191	058	010	001	0+	0+	0+	0+	0+	0+
	3	001	042	156	239	124	034	005	0+	0+	0+	0+	0+	0+
	4	0+	008	060	209	187	080	018	002	0+	0+	0+	0+	0+
	5	0+	001	018	136	208	138	047	008	001	0+	0+	0+	0+
	6	0+	0+	004	068	178	184	094	024	003	0+	0+	0+	0+
	7	0+	0+	001	027	120	193	148	057	010	0+	0+	0+	0+
	8	0+	0+	0+	008	064	161	186	107	028	002	0+	0+	0+
	9	0+	0+	0+	002	028	107	186	161	064	008	0+	0+	0+
	10	0+	0+	0+	0+	010	057	148	193	120	027	001	0+	0+
	11	0+	0+	0+	0+	003	024	094	184	178	068	004	0+	0+
	12	0+	0+	0+	0+	001	008	047	138	208	136	018	001	0+
	13	0+	0+	0+	0+	0+	002	018	080	187	209	060	008	0+
	14	0+	0+	0+	0+	0+	0+	005	034	124	239	156	042	001
	15	0+	0+	0+	0+	0+	0+	001	010	058	191	280	158	012
	16	0+	0+	0+	0+	0+	0+	0+	002	017	096	315	374	145
	17	0+	0+	0+	0+	0+	0+	0+	0+	002	022	167	418	843
18	0	835	397	150	018	002	0+	0+	0+	0+	0+	0+	0+	0+
	1	152	376	300	081	013	001	0+	0+	0+	0+	0+	0+	0+
	2	013	168	284	172	046	007	001	0+	0+	0+	0+	0+	0+
	3	001	047	168	230	105	025	003	0+	0+	0+	0+	0+	0+
	4	0+	009	070	215	168	061	012	001	0+	0+	0+	0+	0+
	5	0+	001	002	151	202	115	033	004	0+	0+	0+	0+	0+
	6	0+	0+	005	082	187	166	071	014	001	0+	0+	0+	0+
	7	0+	0+	001	035	138	189	121	037	005	0+	0+	0+	0+
	8	0+	0+	0+	012	081	173	167	077	015	001	0+	0+	0+
	9	0+	0+	0+	003	038	128	186	128	038	003	0+	0+	0+
	10	0+	0+	0+	001	015	077	167	173	081	012	0+	0+	0+
	11	0+	0+	0+	0+	005	037	121	189	138	035	001	0+	0+
	12	0+	0+	0+	0+	001	014	071	166	187	082	005	0+	0+
	13	0+	0+	0+	0+	0+	004	033	115	202	151	022	001	0+
	14	0+	0+	0+	0+	0+	001	012	061	168	215	070	009	0+
	15	0+	0+	0+	0+	0+	0+	003	025	105	230	168	097	001
	16	0+	0+	0+	0+	0+	0+	001	007	046	172	284	168	013
	17	0+	0+	0+	0+	0+	0+	0+	001	013	081	300	376	152
	18	0+	0+	0+	0+	0+	0+	0+	0+	002	018	150	397	835
19	0	826	377	135	014	001	0+	0+	0+	0+	0+	0+	0+	0+
	1	159	377	285	069	009	001	0+	0+	0+	0+	0+	0+	0+
	2	014	179	285	154	036	005	0+	0+	0+	0+	0+	0+	0+
	3	001	053	180	218	087	018	002	0+	0+	0+	0+	0+	0+
	4	0+	011	080	218	149	047	007	0+	0+	0+	0+	0+	0+
	5	0+	002	027	164	192	093	022	002	0+	0+	0+	0+	0+
	6	0+	0+	007	096	192	145	052	008	0+	0+	0+	0+	0+
	7	0+	0+	001	044	152	180	096	024	002	0+	0+	0+	0+
	8	0+	0+	0+	017	098	180	144	053	008	0+	0+	0+	0+
	9	0+	0+	0+	005	051	146	176	098	022	001	0+	0+	0+
	10	0+	0+	0+	001	022	098	176	146	051	005	0+	0+	0+
	11	0+	0+	0+	0+	008	053	144	180	098	017	0+	0+	0+
	12	0+	0+	0+	0+	002	024	096	180	152	044	001	0+	0+
	13	0+	0+	0+	0+	0+	008	052	145	192	096	007	0+	0+
	14	0+	0+	0+	0+	0+	002	022	093	192	164	027	002	0+
	15	0+	0+	0+	0+	0+	0+	007	047	149	218	080	011	0+
	16	0+	0+	0+	0+	0+	0+	002	018	087	218	180	053	001
	17	0+	0+	0+	0+	0+	0+	0+	005	036	154	285	179	014
	18	0+	0+	0+	0+	0+	0+	0+	001	009	069	285	377	159
	19	0+	0+	0+	0+	0+	0+	0+	0+	001	014	135	377	826
20	0	818	358	122	012	001	0+	0+	0+	0+	0+	0+	0+	0+

TABLE 2 (*Suite*) DISTRIBUTION BINÔMIALE

		p											
n	*X*	0,01	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95	0,99
	1	165	377	270	058	007	0+	0+	0+	0+	0+	0+	0+	0+
	2	016	189	285	137	028	003	0+	0+	0+	0+	0+	0+	0+
	3	001	060	190	205	072	012	001	0+	0+	0+	0+	0+	0+
	4	0+	013	090	218	130	035	005	0+	0+	0+	0+	0+	0+
	5	0+	002	032	175	179	075	015	001	0+	0+	0+	0+	0+
	6	0+	0+	009	109	192	124	037	005	0+	0+	0+	0+	0+
	7	0+	0+	002	054	164	166	074	015	001	0+	0+	0+	0+
	8	0+	0+	0+	022	114	180	120	036	004	0+	0+	0+	0+
	9	0+	0+	0+	007	065	160	160	071	012	0+	0+	0+	0+
	10	0+	0+	0+	002	031	117	176	117	031	002	0+	0+	0+
	11	0+	0+	0+	0+	012	071	160	160	065	007	0+	0+	0+
	12	0+	0+	0+	0+	004	036	120	180	114	022	0+	0+	0+
	13	0+	0+	0+	0+	001	015	074	166	164	054	002	0+	0+
	14	0+	0+	0+	0+	0+	005	037	124	192	109	009	0+	0+
	15	0+	0+	0+	0+	0+	001	015	075	179	175	032	002	0+
	16	0+	0+	0+	0+	0+	0+	005	035	130	218	090	013	0+
	17	0+	0+	0+	0+	0+	0+	001	012	072	205	190	060	001
	18	0+	0+	0+	0+	0+	0+	0+	003	028	137	285	189	016
	19	0+	0+	0+	0+	0+	0+	0+	0+	007	058	270	377	165
	20	0+	0+	0+	0+	0+	0+	0+	0+	001	012	122	358	818

TABLE 3 NOMBRES ALÉATOIRES

	1	2	3	4	5	6	7	8
1	04479	44211	81536	09686	26743	87001	62392	59946
2	87019	90503	16034	07862	19701	85949	85876	58188
3	54222	56179	09833	34227	43897	38517	11617	30338
4	17929	24021	50932	89349	08012	37925	59003	95503
5	56399	82269	69443	62020	03365	82164	01356	24871
6	79242	52682	36255	74168	28636	93043	65454	36152
7	88869	22489	50467	14964	93146	51852	32408	22545
8	83970	03473	42981	83127	98774	74392	12218	91841
9	52754	85751	92705	70949	24331	42672	04885	44251
10	27011	69215	90920	96218	81127	67792	08377	60773
11	67952	43155	65547	50055	97940	38833	08745	69207
12	93376	38289	89474	22350	84982	85224	29969	42745
13	13674	24899	60602	33203	91953	48635	43938	08285
14	19345	11394	09241	72723	09052	76987	89854	48849
15	91609	18375	16171	30692	37389	51879	29556	51315
16	68537	17630	70322	26128	15645	91691	81064	58083
17	10038	17181	93964	41122	13020	98243	46447	28675
18	57023	28928	73917	94774	62542	30536	14777	72360
19	70791	39030	11261	76783	31184	38669	95862	99067
20	88033	42447	17815	75351	40853	83513	88714	09887
21	71858	65129	60871	04586	90651	93207	85501	83600
22	40357	80097	82138	61279	70478	49731	94154	50436
23	03339	05350	61895	46420	81433	61995	16654	91274
24	43084	21898	98854	70139	31516	29990	40919	05125
25	77223	58612	93223	12495	12628	43715	88010	03080
26	87135	86620	56893	82220	33968	13380	38087	74056
27	82521	26025	67975	81512	85227	39786	82990	38936
28	63590	66694	35357	19452	67724	10912	58569	66929
29	38133	07569	71030	75769	89240	48888	27184	78014
30	70369	48709	65114	69725	42994	22584	18455	52022
31	16168	91235	17509	72148	34676	61011	03681	21135
32	38541	45056	27395	13139	57487	57389	10764	62267
33	35508	90052	94492	83678	11316	98396	20893	87494
34	27147	55333	29880	81775	05384	86224	43487	86643
35	95351	12900	12689	07330	29470	39802	79928	68896
36	48047	70852	63798	62452	83695	38200	17414	13151
37	86417	48099	72299	46033	88948	93459	89657	52339
38	10361	07412	48001	57271	13210	04328	23855	65719
39	29998	88220	63213	98976	78720	61138	90709	50003
40	13008	59213	55737	68130	74358	74687	79519	29409
41	92882	31482	05651	53952	00915	43967	62276	47818
42	57149	14046	02876	79221	76700	68078	67712	98230
43	87080	09985	68303	23068	73514	39328	56046	98785
44	33504	84019	91220	05463	19500	66509	87209	71293
45	44702	70429	73468	16316	87536	49921	00239	37743
46	40473	76124	12097	56736	84635	77172	47155	77306
47	31727	64165	28937	14805	22863	62154	87637	80982
48	98855	63471	83278	00131	90229	02976	49485	67541
49	56067	73922	05810	24125	09603	99539	04848	57223
50	87152	73758	86758	77787	47126	31822	72088	00927

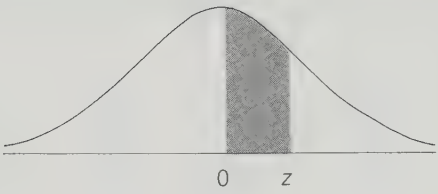


TABLE 4 AIRE SOUS LA COURBE NORMALE

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2518	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,34134	0,34375	0,34614	0,34850	0,35083	0,35314	0,35543	0,35769	0,35993	0,36214
1,1	0,36433	0,36650	0,36864	0,37076	0,37286	0,37493	0,37698	0,37900	0,38100	0,38298
1,2	0,38493	0,38686	0,38877	0,39065	0,39251	0,39435	0,39617	0,39796	0,39973	0,40147
1,3	0,40320	0,40490	0,40658	0,40824	0,40988	0,41149	0,41309	0,41466	0,41621	0,41774
1,4	0,41924	0,42073	0,42220	0,42364	0,42507	0,42647	0,42786	0,42922	0,43056	0,43189
1,5	0,43319	0,43448	0,43574	0,43699	0,43822	0,43943	0,44062	0,44179	0,42295	0,44408
1,6	0,44520	0,44630	0,44738	0,44845	0,44950	0,45053	0,45154	0,45254	0,45352	0,45449
1,7	0,45543	0,45637	0,45728	0,45818	0,45907	0,45994	0,46080	0,46164	0,46246	0,46327
1,8	0,46407	0,46485	0,46562	0,46638	0,46712	0,46784	0,46856	0,46926	0,46995	0,47062
1,9	0,47128	0,47193	0,47257	0,47320	0,47381	0,47441	0,47500	0,47558	0,47615	0,47670
2,0	0,47725	0,47778	0,47831	0,47882	0,47932	0,47982	0,48030	0,48077	0,48124	0,48169
2,1	0,48214	0,48257	0,48300	0,48341	0,48382	0,48422	0,48461	0,48500	0,48537	0,48574
2,2	0,48610	0,48645	0,48679	0,48713	0,48745	0,48778	0,48809	0,48840	0,48870	0,48899
2,3	0,48928	0,48956	0,48983	0,49010	0,49036	0,49061	0,49086	0,49111	0,49134	0,49158
2,4	0,49180	0,49202	0,49224	0,49245	0,49266	0,49286	0,49305	0,49324	0,49343	0,49361
2,5	0,49379	0,49396	0,49413	0,49430	0,49446	0,49461	0,49477	0,49492	0,49506	0,49520
2,6	0,49534	0,49547	0,49560	0,49573	0,49585	0,49598	0,49609	0,49621	0,49632	0,49643
2,7	0,49653	0,49664	0,49674	0,49683	0,49693	0,49702	0,49711	0,49720	0,49728	0,49736
2,8	0,49744	0,49752	0,49760	0,49767	0,49774	0,49781	0,49788	0,49795	0,49801	0,49807
2,9	0,49813	0,49819	0,49825	0,49831	0,49836	0,49841	0,49846	0,49851	0,49856	0,49861
3,0	0,49865	0,49869	0,49874	0,49878	0,49882	0,49886	0,49889	0,49893	0,49897	0,49900
3,1	0,49903	0,49906	0,49910	0,49913	0,49916	0,49918	0,49921	0,49924	0,49926	0,49929
3,2	0,49931	0,49934	0,49936	0,49938	0,49940	0,49942	0,49944	0,49946	0,49948	0,49950
3,3	0,49952									
3,4	0,49966									
3,5	0,49977									
3,6	0,49984									
3,7	0,49989									
3,8	0,49993									
3,9	0,49995									
4,0	0,49997									

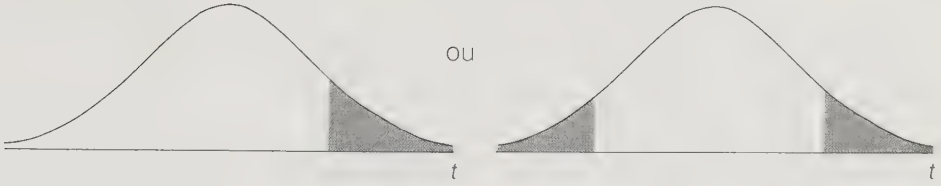


TABLE 5 DISTRIBUTION DU «*t*» DE STUDENT

Unilatéral / *Bilatéral*	0,10 / 0,20	0,05 / 0,10	0,025 / 0,05	0,01 / 0,02	0,005 / 0,01
D					
1	3,08	6,31	12,7	31,8	63,7
2	1,89	2,92	4,30	6,97	9,92
3	1,64	2,35	3,18	4,54	5,84
4	1,53	2,13	2,78	3,75	4,60
5	1,48	2,02	2,57	3,37	4,03
6	1,44	1,94	2,45	3,14	3,71
7	1,42	1,90	2,37	3,00	3,50
8	1,40	1,86	2,31	2,90	3,36
9	1,38	1,83	2,26	2,82	3,25
10	1,37	1,81	2,23	2,76	3,17
11	1,36	1,80	2,20	2,72	3,10
12	1,36	1,78	2,18	2,68	3,06
13	1,35	1,77	2,16	2,65	3,01
14	1,35	1,76	2,15	2,62	2,98
15	1,34	1,75	2,13	2,60	2,95
16	1,34	1,75	2,12	2,58	2,92
17	1,33	1,74	2,11	2,57	2,90
18	1,33	1,73	2,10	2,55	2,88
19	1,33	1,73	2,09	2,54	2,86
20	1,33	1,73	2,09	2,53	2,85
21	1,32	1,72	2,08	2,52	2,83
22	1,32	1,72	2,07	2,51	2,82
23	1,32	1,71	2,07	2,50	2,81
24	1,32	1,71	2,06	2,49	2,80
25	1,32	1,71	2,06	2,49	2,79
26	1,32	1,71	2,06	2,48	2,78
27	1,31	1,70	2,05	2,47	2,77
28	1,31	1,70	2,05	2,47	2,76
29	1,31	1,70	2,05	2,46	2,76
30	1,31	1,70	2,04	2,46	2,75
40	1,30	1,68	2,02	2,42	2,70
∞	1,28	1,65	1,96	2,33	2,58

Reproduit à partir de la Table 10 de *Biometrika Tables for Statisticians*, Vol. I, 1954, avec la permission de S. Pearson et des éditeurs de Biometrika.

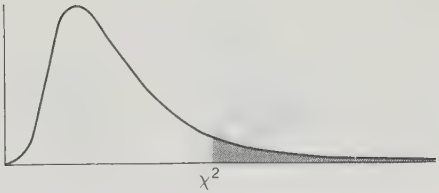


TABLE 6 DISTRIBUTION DU χ^2

D	0,10	0,05	0,025	0,01
1	2,71	3,84	5,02	6,63
2	4,61	5,99	7,38	9,21
3	6,25	7,81	9,35	11,3
4	7,78	9,49	11,1	13,3
5	9,24	11,1	12,8	15,1
6	10.6	12,6	14,4	16,8
7	12,0	14,1	16,0	18,5
8	13,4	15,5	17,5	20,1
9	14,7	16,9	19,0	21,7
10	16,0	18,3	20,5	23,2
11	17,3	19,7	21,9	24,7
12	18,5	21,0	23,3	26,2
13	19,8	22,4	24,7	27,7
14	21,1	23,7	26,1	29,1
15	22,3	25,0	27,5	30,6
16	23,5	26,3	28,8	32,0
17	24,8	27,6	30,2	33,4
18	26,0	28,9	31,5	34,8
19	27,2	30,1	32,9	36,2
20	28,4	31,4	34,2	37,6
21	29,6	32,7	35,5	38,9
22	30,8	33,9	36,8	40,3
23	32,0	35,2	38,1	41,6
24	33,2	36,4	39,4	43,0
25	34,4	37,7	40,6	44,3
26	35,6	38,9	41,9	45,6
27	36,7	40,1	43,2	47,0
28	37,9	41,3	44,5	48,3
29	39,1	42,6	45,7	49,6
30	40,3	43,8	47,0	50,9
40	51,8	55,8	59,3	63,7
50	63,2	67,5	71,4	76,2
60	74,4	79,1	83,3	88,4
70	85,5	90,5	95,0	100,4
80	96,6	101,9	106,6	112,3

Reproduit à partir de la Table 8 de *Biometrika Tables for Statisticians,* Vol. I, 1954, avec la permission de E.S. Pearson et des éditeurs de Biometrika.

TABLE 7a DISTRIBUTIONS F.
$\alpha = 0{,}05$

D_1 \ D_2	1	2	3	4	5	6	7	8	9
1	161	200	216	225	230	234	237	239	241
2	18,5	19,0	19,2	19,3	19,3	19,3	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,88	8,85	8,81
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04
80	3,96	3,11	2,72	2,48	2,32	2,21	2,12	2,05	1,99
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88

Reproduit à partir de la Table 18 de *Biometrika Tables for Statisticians*, Vol. I, 1954, avec la permission de E.S. Pearson et les éditeurs de Biometrika.

TABLE 7a DISTRIBUTION *F* (*Suite*)
$\alpha = 0{,}05$

D_y \ D_x	12	14	16	20	24	30	40	60	100
1	244	245	246	248	249	250	251	252	253
2	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	8,74	8,71	8,69	8,66	8,64	8,62	8,59	8,57	8,56
4	5,91	5,87	5,84	5,80	5,77	5,75	5,72	5,69	5,66
5	4,68	4,64	4,60	4,56	4,53	4,50	4,46	4,43	4,40
6	4,00	3,96	3,92	3,87	3,84	3,81	3,77	3,74	3,71
7	3,57	3,52	3,49	3,44	3,41	3,38	3,34	3,30	3,28
8	3,28	3,23	3,20	3,15	3,12	3,08	3,04	3,01	2,98
9	3,07	3,02	2,98	2,94	2,90	2,86	2,83	2,79	2,76
10	2,91	2,87	2,82	2,77	2,74	2,70	2,66	2,62	2,59
11	2,79	2,74	2,70	2,65	2,61	2,57	2,53	2,49	2,45
12	2,69	2,64	2,60	2,54	2,51	2,47	2,43	2,38	2,35
13	2,60	2,55	2,51	2,46	2,42	2,38	2,34	2,30	2,26
14	2,53	2,48	2,44	2,39	2,35	2,31	2,27	2,22	2,19
15	2,48	2,43	2,39	2,33	2,29	2,25	2,20	2,16	2,12
16	2,42	2,37	2,33	2,28	2,24	2,19	2,15	2,11	2,07
17	2,38	2,33	2,29	2,23	2,19	2,15	2,10	2,06	2,02
18	2,34	2,29	2,25	2,19	2,15	2,11	2,06	2,02	1,98
19	2,31	2,26	2,21	2,16	2,11	2,07	2,03	1,98	1,94
20	2,28	2,23	2,18	2,12	2,08	2,04	1,99	1,95	1,90
21	2,25	2,20	2,15	2,10	2,05	2,01	1,96	1,92	1,87
22	2,23	2,18	2,13	2,07	2,03	1,98	1,94	1,89	1,84
23	2,20	2,14	2,10	2,05	2,01	1,96	1,91	1,86	1,82
24	2,18	2,13	2,09	2,03	1,98	1,94	1,89	1,84	1,80
25	2,16	2,11	2,06	2,01	1,96	1,92	1,87	1,82	1,77
26	2,15	2,10	2,05	1,99	1,95	1,90	1,85	1,80	1,76
27	2,13	2,08	2,03	1,97	1,93	1,88	1,84	1,79	1,74
28	2,12	2,06	2,02	1,96	1,91	1,87	1,82	1,77	1,72
29	2,10	2,05	2,00	1,94	1,90	1,85	1,81	1,75	1,71
30	2,09	2,04	1,99	1,93	1,89	1,84	1,79	1,74	1,69
40	2,00	1,95	1,90	1,84	1,79	1,74	1,69	1,64	1,59
60	1,92	1,86	1,81	1,75	1,70	1,65	1,59	1,53	1,48
80	1,88	1,82	1,77	1,70	1,65	1,60	1,54	1,47	1,42
100	1,85	1,79	1,75	1,68	1,63	1,57	1,51	1,44	1,39
120	1,83	1,77	1,73	1,66	1,61	1,55	1,50	1,43	1,37
∞	1,75	1,69	1,64	1,57	1,52	1,46	1,39	1,32	1,24

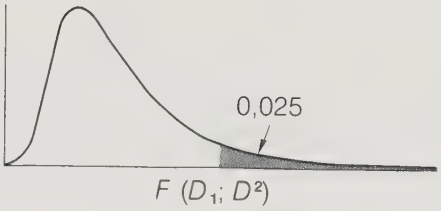


TABLE 7b DISTRIBUTION F
$\alpha = 0{,}025$

D_y \ D_x	1	2	3	4	5	6	7	8	9
1	648	800	864	900	922	937	948	957	963
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31
14	6,30	4,86	4,24	3,89	3,66	3,50	3,37	3,28	3,21
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68
26	5,66	4,26	3,66	3,32	3,10	2,94	2,82	2,72	2,65
27	5,63	4,23	3,64	3,30	3,08	2,92	2,80	2,70	2,63
28	5,61	4,21	3,62	3,28	3,06	2,90	2,78	2,68	2,61
29	5,59	4,19	3,60	3,26	3,04	2,88	2,76	2,66	2,59
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57
40	5,42	4,05	3,46	3,13	3,90	2,74	2,62	2,53	2,45
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33
80	5,20	3,85	3,28	2,94	2,72	2,57	2,45	2,37	2,25
100	5,17	3,82	3,24	2,90	2,68	2,53	2,41	2,31	2,21
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11

TABLE 7b DISTRIBUTION *F* (*Suite*)
$\alpha = 0{,}025$

D_y \ D_x	12	14	16	20	24	30	40	60	100
1	977	984	989	993	997	1000	1006	1010	1013
2	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5
3	14,3	14,3	14,3	14,2	14,1	14,1	14,0	14,0	13,9
4	8,75	8,69	8,64	8,56	8,51	8,46	8,41	8,36	8,32
5	6,52	6,45	6,41	6,33	6,28	6,23	6,18	6,12	6,08
6	5,37	5,28	5,25	5,17	5,12	5,07	5,01	4,96	4,92
7	4,67	4,59	4,55	4,47	4,42	4,36	4,31	4,25	4,21
8	4,20	4,13	4,08	4,00	3,95	3,89	3,84	3,78	3,74
9	3,87	3,80	3,75	3,67	3,62	3,56	3,51	3,45	3,40
10	3,62	3,55	3,50	3,42	3,37	3,31	3,26	3,20	3,15
11	3,43	3,36	3,31	3,23	3,17	3,12	3,06	3,00	2,95
12	3,28	3,20	3,15	3,07	3,02	2,96	2,91	2,85	2,80
13	3,15	3,08	3,03	2,95	2,89	2,84	2,78	2,72	2,67
14	3,05	2,98	2,92	2,84	2,79	2,73	2,67	2,61	2,56
15	2,96	2,89	2,84	2,76	2,70	2,64	2,59	2,52	2,47
16	2,89	2,82	2,76	2,68	2,63	2,57	2,51	2,45	2,40
17	2,82	2,75	2,70	2,62	2,56	2,50	2,44	2,38	2,33
18	2,77	2,70	2,64	2,56	2,50	2,44	2,38	2,32	2,27
19	2,72	2,65	2,59	2,51	2,45	2,39	2,33	2,27	2,22
20	2,68	2,60	2,54	2,46	2,41	2,35	2,29	2,22	2,17
21	2,64	2,57	2,50	2,42	2,37	2,31	2,25	2,18	2,13
22	2,60	2,53	2,47	2,39	2,33	2,27	2,21	2,14	2,09
23	2,57	2,50	2,44	2,36	2,30	2,24	2,18	2,11	2,06
24	2,54	2,47	2,41	2,33	2,27	2,21	2,15	2,08	2,03
25	2,51	2,44	2,38	2,30	2,24	2,18	2,12	2,05	2,00
26	2,48	2,41	2,35	2,27	2,21	2,15	2,09	2,02	1,97
27	2,46	2,39	2,33	2,25	2,19	2,13	2,07	2,00	1,95
28	2,44	2,37	2,31	2,23	2,17	2,11	2,05	1,98	1,93
29	2,42	2,35	2,29	2,21	2,15	2,09	2,03	1,96	1,91
30	2,41	2,34	2,28	2,20	2,14	2,07	2,01	1,94	1,89
40	2,29	2,21	2,15	2,07	2,01	1,94	1,88	1,80	1,76
60	2,17	2,09	2,03	1,94	1,88	1,82	1,74	1,67	1,61
80	2,10	2,03	1,96	1,87	1,81	1,75	1,67	1,59	1,52
100	2,07	1,99	1,93	1,84	1,78	1,71	1,63	1,55	1,48
120	2,05	1,98	1,92	1,82	1,76	1,69	1,61	1,53	1,46
∞	1,94	1,86	1,81	1,71	1,64	1,57	1,48	1,39	1,30

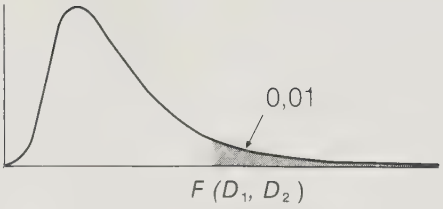


TABLE 7c DISTRIBUTIONS F
$\alpha = 0{,}01$

D_Y \ D_X	1	2	3	4	5	6	7	8	9
1	4052	4999	5403	5625	5764	5859	5928	5981	6022
2	98,5	99,0	99,2	99,3	99,3	99,3	99,3	99,4	99,4
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3
4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7
5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2
6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98
7	12,3	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71
8	11,3	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91
9	10,6	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35
10	10,0	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95
11	9,65	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68
18	8,28	6,01	5,09	4,58	4,25	4,01	3,85	3,71	3,60
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52
20	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45
21	8,02	5,78	4,87	4,37	4,04	3,81	3,65	3,51	3,40
22	7,94	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,40	3,30
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,25
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,21
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,17
27	7,68	5,49	4,60	4,11	3,79	3,56	3,39	3,26	3,14
28	7,64	5,45	4,57	4,08	3,76	3,53	3,36	3,23	3,11
29	7,60	5,42	4,54	4,05	3,73	3,50	3,33	3,20	3,08
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,06
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,88
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72
80	6,96	4,88	4,04	3,56	3,25	3,04	2,87	2,74	2,64
100	6,90	4,82	3,98	3,51	3,20	2,99	2,82	2,69	2,59
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56
∞	6,63	4,60	3,78	3,32	3,02	2,80	2,64	2,51	2,41

TABLE 7c DISTRIBUTION *F* (*Suite*)
$\alpha = 0{,}01$

D_Y	D_X 12	14	16	20	24	30	40	100
1	6110	6140	6170	6210	6230	6260	6290	6330
2	99,4	99,4	99,5	99,5	99,5	99,5	99,5	99,5
3	27,1	26,9	26,8	26,7	26,6	26,5	26,4	26,2
4	14,4	14,2	14,1	14,0	13,9	13,8	13,7	13,6
5	9,89	9,77	9,68	9,55	9,47	9,38	9,29	9,13
6	7,72	7,60	7,52	7,39	7,31	7,23	7,14	6,99
7	6,47	6,35	6,27	6,15	6,07	5,98	5,90	5,75
8	5,67	5,56	5,48	5,36	6,28	5,20	5,11	4,90
9	5,11	5,00	4,92	4,80	4,73	4,64	4,56	4,41
10	4,71	4,60	4,52	4,40	4,33	4,25	4,17	4,01
11	4,40	4,29	4,21	4,10	4,02	3,94	3,86	3,70
12	4,16	4,05	3,98	3,86	3,78	3,70	3,61	3,46
13	3,96	3,85	3,78	3,67	3,59	3,51	3,42	3,27
14	3,80	3,70	3,62	3,51	3,43	3,34	3,26	3,11
15	3,67	3,56	3,48	3,36	3,29	3,20	3,12	2,97
16	3,55	3,45	3,37	3,25	3,18	3,10	3,01	2,86
17	3,45	3,35	3,27	3,15	3,08	3,00	2,92	2,76
18	3,37	3,27	3,19	3,07	3,00	2,92	2,83	2,68
19	3,30	3,19	3,12	3,00	2,92	2,84	2,76	2,60
20	3,23	3,13	3,05	2,94	2,86	2,77	2,69	2,53
21	3,17	3,07	2,99	2,88	2,80	2,72	2,63	2,47
22	3,12	3,02	2,94	2,83	2,75	2,67	2,58	2,42
23	3,07	2,97	2,89	2,78	2,70	2,62	2,53	2,37
24	3,03	2,93	2,85	2,74	2,66	2,58	2,49	2,33
25	2,99	2,89	2,81	2,70	2,62	2,54	2,45	2,29
26	2,96	2,86	2,77	2,66	2,58	2,50	2,41	2,25
27	2,93	2,83	2,74	2,63	2,55	2,47	2,38	2,21
28	2,90	2,80	2,71	2,60	2,52	2,44	2,35	2,18
29	2,87	2,77	2,68	2,57	2,49	2,41	2,32	2,15
30	2,84	2,74	2,66	2,55	2,46	2,38	2,29	2,13
40	2,66	2,56	2,49	2,37	2,29	2,20	2,11	1,97
60	2,50	2,40	2,32	2,20	2,12	2,03	1,93	1,74
80	2,41	2,32	2,24	2,11	2,03	1,94	1,84	1,65
100	2,36	2,26	2,19	2,06	1,98	1,89	1,79	1,59
120	2,34	2,24	2,16	2,03	1,95	1,86	1,76	1,56
∞	2,18	2,07	1,99	1,87	1,79	1,70	1,59	

TABLE 8 TRANSFORMATION DE **r à Z** ET DE ρ **à** μ_z

r	Z	r	Z	r	Z	r	Z
0,00	0,000						
0,01	0,010	0,26	0,266	0,51	0,563	0,76	0,996
0,02	0,020	0,27	0,277	0,52	0,576	0,77	1,020
0,03	0,030	0,28	0,288	0,53	0,590	0,78	1,045
0,04	0,040	0,29	0,299	0,54	0,604	0,79	1,071
0,05	0,050	0,30	0,310	0,55	0,618	0,80	1,099
0,06	0,060	0,31	0,321	0,56	0,633	0,81	1,127
0,07	0,070	0,32	0,332	0,57	0,648	0,82	1,157
0,08	0,080	0,33	0,343	0,58	0,663	0,83	1,188
0,09	0,090	0,34	0,354	0,59	0,678	0,84	1,221
0,10	0,100	0,35	0,365	0,60	0,693	0,85	1,256
0,11	0,110	0,36	0,377	0,61	0,709	0,86	1,293
0,12	0,121	0,37	0,388	0,62	0,725	0,87	1,333
0,13	0,131	0,38	0,400	0,63	0,741	0,88	1,376
0,14	0,141	0,39	0,412	0,64	0,758	0,89	1,422
0,15	0,151	0,40	0,424	0,65	0,775	0,90	1,472
0,16	0,161	0,41	0,436	0,66	0,793	0,91	1,528
0,17	0,172	0,42	0,448	0,67	0,811	0,92	1,589
0,18	0,182	0,43	0,460	0,68	0,829	0,93	1,658
0,19	0,192	0,44	0,472	0,69	0,848	0,94	1,738
0,20	0,203	0,45	0,485	0,70	0,867	0,95	1,832
0,21	0,213	0,46	0,497	0,71	0,887	0,96	1,946
0,22	0,224	0,47	0,510	0,72	0,908	0,97	2,092
0,23	0,234	0,48	0,523	0,73	0,929	0,98	2,298
0,24	0,245	0,49	0,536	0,74	0,950	0,99	2,647
0,25	0,255	0,50	0,549	0,75	0,973		

TABLE 9 DISTRIBUTION DE *R* DANS UN TEST DE SÉQUENCES (RUNS).

La table *A* donne la borne inférieure et la table *B* la borne supérieur de la valeur critique de *R* avec un niveau de signification de 0,05.

A

n_Y \ n_X	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											2	2	2	2	2	2	2	2	2
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
4				2	2	2	3	3	3	3	3	3	3	3	4	4	4	4	4
5			2	2	3	3	3	3	3	4	4	4	4	4	4	4	5	5	5
6		2	2	3	3	3	3	4	4	4	4	5	5	5	5	5	5	6	6
7		2	2	3	3	3	4	4	5	5	5	5	5	6	6	6	6	6	6
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	7	7	7	7
9		2	3	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8
10		2	3	3	4	5	5	5	6	6	7	7	7	7	8	8	8	8	9
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13
18	2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	13	13
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13
20	2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14

B

n_Y \ n_X	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2											6	6	6	6	6	6	6	6	6
3					8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
4				9	9	10	10	10	10	10	10	10	10	10	10	10	10	10	10
5			9	10	10	11	11	12	12	12	12	12	12	12	12	12	12	12	12
6		8	9	10	11	12	12	13	13	13	13	14	14	14	14	14	14	14	14
7		8	10	11	12	13	13	14	14	14	14	15	15	15	16	16	16	16	16
8		8	10	11	12	13	14	14	15	15	16	16	16	16	17	17	17	17	17
9		8	10	12	13	14	14	15	16	16	16	17	17	18	18	18	18	18	18
10		8	10	12	13	14	15	16	16	17	17	18	18	18	19	19	19	20	20
11		8	10	12	13	14	15	16	17	17	18	19	19	19	20	20	20	21	21
12	6	8	10	12	13	14	16	16	17	18	19	19	20	20	21	21	21	22	22
13	6	8	10	12	14	15	16	17	18	19	19	20	20	21	21	22	22	23	23
14	6	8	10	12	14	15	16	17	18	19	20	20	21	22	22	23	23	23	24
15	6	8	10	12	14	15	16	18	18	19	20	21	22	22	23	23	24	24	25
16	6	8	10	12	14	16	17	18	19	20	21	21	22	23	23	24	25	25	25
17	6	8	10	12	14	16	17	18	19	20	21	22	23	23	24	25	25	26	26
18	6	8	10	12	14	16	17	18	19	20	21	22	23	24	25	25	26	26	27
19	6	8	10	12	14	16	17	18	20	21	22	23	23	24	25	26	26	27	27
20	6	8	10	12	14	16	17	18	20	21	22	23	24	25	25	26	27	27	28

TABLE 10 DISTRIBUTION DE U DANS UN TEST DE MANN-WHITNEY. $\alpha = 0{,}05$

Toute valeur de U qui est égale ou inférieure à la valeur critique de la table est significative au niveau de 0,05 pour un test bilatéral.

n du grand échantillon	n du petit échantillon																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4	–	–	0																
5	–	0	1	2															
6	–	1	2	3	5														
7	–	1	3	5	6	8													
8	0	2	4	6	8	10	13												
9	0	2	4	7	10	12	15	17											
10	0	3	5	8	11	14	17	20	23										
11	0	3	6	9	13	16	19	23	26	30									
12	1	4	7	11	14	18	22	26	29	33	37								
13	1	4	8	12	16	20	24	28	33	37	41	45							
14	1	5	9	13	17	22	26	31	36	40	45	50	55						
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64					
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75				
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87			
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99		
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	3	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138	147	156
25	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145	154	163
26	4	11	19	28	37	46	55	64	74	83	93	102	112	122	132	141	151	161	171
27	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158	168	178
28	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164	175	186
29	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171	182	193
30	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177	189	200
31	5	14	24	34	45	56	67	78	90	101	113	125	136	148	160	172	184	196	208
32	5	14	24	35	46	58	69	81	93	105	117	129	141	153	166	178	190	203	215
33	5	15	25	37	48	60	72	84	96	108	121	133	146	159	171	184	197	210	222
34	5	15	26	38	50	62	74	87	99	112	125	138	151	164	177	190	203	217	230
35	6	16	27	39	51	64	77	89	103	116	129	142	156	169	183	196	210	224	237
36	6	16	28	40	53	66	79	92	106	119	133	147	161	174	188	202	216	231	245
37	6	17	29	41	55	68	81	95	109	123	137	151	165	180	194	209	223	238	252
38	6	17	30	43	56	70	84	98	112	127	141	156	170	185	200	215	230	245	259
39	7	18	31	44	58	72	86	101	115	130	145	160	175	190	206	221	236	252	267
40	7	18	31	45	59	74	89	103	119	134	149	165	180	196	211	227	243	258	274

INDEX

A

Analyse de variance, 261
Arrangement, 90
Au hasard, 67

B

Bimodale, 31

C

Carré, 8
Centiles, 38
Classe, 18
  Bornes, 19
  Largeur, 18
  Limite inférieure, 19
  Limite supérieure, 19
  Ouverte, 19
  Point milieu, 19
Coefficient, 297
  De corrélation, 292
  Calcul de r, 297
  De détermination, 297
Combinaison, 93
Correction de Yates, 248
Corrélation de rang de Spearman, 336
Courbe normale, 109, 123
  Aire sous, 123
  Centrée, réduite, 109, 110
    Aire sous, 110
    Moyenne, 109

D

Décision, 189
  Règle, 189
Degré de liberté, 219, 244
  Tableau de contingences, 244
Diagramme (s), 21, 286
  De Venn, 65
  Dispersion, 286
  En bâtonnets, 21
Dissymétrie, 23, 56
Distribution (s), 87, 88, 158, 215, 217, 237, 255
  Binômiale, 95, 98
    Écart-type, 98
    Histogramme, 105
    Moyenne, 98
  D'échantillonnage, 140, 141, 163, 168
    Différences de moyennes, 163
    Différences de proportions, 168
    Moyenne (s), 140, 141
  De fréquence (s), 18, 87
  De probabilité, 87, 88
    Écart-type de la population, 88
    Moyenne de la population, 88
  Échantillonnage de proportions, 158
  F, 255, 258
    Tests de variance, 258
  Fréquences relatives, 88
  Normale (s), 108, 110, 122, 126
    Écart-type, 110, 126
    Moyenne, 126
  t de Student, 215, 233
  Table, 219
  χ^2, 237, 238
Données, 15
  À caractères spécifiques (nominales), 15
  Ordinales, 16
  Quantitatives (cardinales), 16
Droite, 280, 286
  Équation, 280
  Graphique, 280
  Régression, 286

E

Écart-type, 49, 53, 116, 158, 162, 165, 168
  Distribution d'échantillonnage des différences de moyennes, 165
  Données groupées, 53
  Population, 116
  Proportions, 158
Échantillon (s), 14
  Aléatoires, 135
  Avec proportions, 173
  Différence de moyennes, 175
  Différence de proportions, 175
  Taille, 171, 172, 173, 175
Échantillonnage, 136, 139
  Amas, 136
  Sans remise, 139
  Stratifié, 136
  Systématique, 136
Éléments, 65

Ensemble, 65
De référence, 65
Universel, 65
Erreur (s), 189, 198
D'estimation, 288
De variation, 294
Première espèce, 189, 198
Deuxième espèce, 189, 198
Erreur-type, 141
Moyenne, 141
Espace échantillonnal, 65
Estimateurs, 150
Biaisés, 150
Efficaces, 150
Estimation, 146
Ponctuelle, 147
Étendue, 47
Événement (s), 63, 65, 70, 71, 76
Certain, 63
Dépendants, 76, 77
Impossible, 63
Indépendants, 76, 77
Mutuellement exclusifs, 70, 71

F

Factoriel, 92
Fréquence (s), 18
Marginales, 243
Observée, 236
Relative, 64
Théorique, 236

H

Histogramme, 22, 105
Hypothèse, 185, 186, 188
Nulle, 186
Rivale, 188

I

Intervalle (s), 48, 314
De confiance, 147, 149, 151, 161
Moyenne (s), 147, 149
Pour une proportion, 161
Semi-Interquartile, 48
Prédiction, 314
Inférence (s), 311
Régression, 311
Statistique, 146

M

Médiane, 31
Mode, 30
Moindres carrés, 288
Méthode, 288
Moyenne, 32, 35, 116, 165, 168
Distribution d'échantillonnage des différences de moyennes, 165
Distribution d'échantillonnage de différences de proportions, 168
Données regroupées, 35
D'un échantillon, 32
D'une population, 172

N

Niveau de confiance, 148
De signification, 189, 191

O

Opération, 3

P

Permutation, 90
Polygone de fréquences, 22
Population, 14
Probabilité (s), 63, 65, 68, 73
Axiomes de, 73
Définition, 63
D'un événement, 68
Subjective, 65

Q

Quartiles, 38
Écart de, 48

R

Racines carrées, 8
Rangs centiles, 38

S

Sigma, 4
Sommation, 5
Sous-ensemble, 65
Statistique (s), 13
Descriptive, 13
Symétrie, 23
Symétrique, 109

T

Table F, 256
Tableaux de contingences, 241
Tableau de données, 17
Brutes, 17
Test, 192, 195, 196, 269, 329, 330, 331, 333
Bilatéral, 192, 195, 196
D'efficacité, 239
Ajustement, 239
Non-paramétriques, 329
Mann-Whitney, 333
Runs, 331
Séquences, 331
Signe, 330
Wilcoxon, 333
Puissance, 335
Scheffé, 269

Unilatéral, 195, 196
Théorème central limite, 144

V

Valeurs critiques, 192
Variables, 24
Continues, 24
Discrètes, 24
Variance, 49, 53
Données groupées, 53
Variation, 46
Rapport de, 46
χ^2, 235, 253
Distribution, 238

Achevé d'imprimer sur les presses
de Laflamme et Charrier,
lithographes

Imprimé au Québec



29 7902 30
$28.45 $29.95

Les Éditions HRW 0-03-929343-2