



MODULO

3^e édition

NOTIONS DE STATISTIQUE

Christiane Simard

VERSION MAÎTRE

NOTIONS DE STATISTIQUE

Christiane Simard

Conception et rédaction
du matériel complémentaire Web
Marc-Antoine Nadeau – Cégep Garneau
Annie Turcotte – Cégep de Lévis-Lauzon

VERSION MAÎTRE

Notions de statistique3^e édition**Version maître**

Christiane Simard

© 2015, 2009 **Groupe Modulo Inc.**

© 2001 Éditions Le Griffon d'Argile Inc.

Conception éditoriale: Éric Mauras*Édition*: Suzanne Champagne*Coordination*: Olivier Rolko*Révision linguistique*: Nicole Blanchette*Correction d'épreuves*: Katie Delisle*Adaptation de la conception graphique originale*: Christian Campana*Conception de la couverture*: Micheline Roy*Impression*: TC Imprimeries Transcontinental**Catalogage avant publication****de Bibliothèque et Archives nationales du Québec****et Bibliothèque et Archives Canada**

Simard, Christiane, 1949-

Notions de statistique. **Version maître**3^e édition.

Comprend des références bibliographiques et un index.

Pour les étudiants du niveau collégial.

ISBN 978-2-89732-017-1

ISBN 978-2-89710-813-7

1. Statistique. 2. Probabilités. 3. Statistique mathématique. 4. Statistique – Problèmes et exercices. i. Titre.

1. Statistique – Étude et enseignement (Collégial). 2. Probabilités – Étude et enseignement (Collégial). 3. Statistique mathématique – Étude et enseignement (Collégial). 4. Statistique – Problèmes et exercices.

i. Titre.

QA273.S55 2014

519.2

C2014-942016-1

QA273.S55 2014 Suppl.

519.2

C2014-942017-X

MODULO

5800, rue Saint-Denis, bureau 900

Montréal (Québec) H2S 3L5 Canada

Téléphone : 514 273-1066

Télécopieur : 514 276-0324 ou 1 800 814-0324

info.modulo@tc.tc

TOUS DROITS RÉSERVÉS.

Toute reproduction du présent ouvrage, en totalité ou en partie, par tous les moyens présentement connus ou à être découverts, est interdite sans l'autorisation préalable de Groupe Modulo Inc.

Toute utilisation non expressément autorisée constitue une contrefaçon pouvant donner lieu à une poursuite en justice contre l'individu ou l'établissement qui effectue la reproduction non autorisée.

ISBN 978-2-89732-017-1**ISBN 978-2-89710-813-7**Dépôt légal: 1^{er} trimestre 2015

Bibliothèque et Archives nationales du Québec

Bibliothèque et Archives Canada

Imprimé au Canada

1 2 3 4 5 ITIB 18 17 16 15 14

Nous reconnaissons l'aide financière du gouvernement du Canada par l'entremise du Fonds du livre du Canada (FLC) pour nos activités d'édition.

Gouvernement du Québec – Programme de crédit d'impôt pour l'édition de livres – Gestion SODEC.

Sources iconographiques**Couverture**: Christian Beirle González/Getty Images;**Ouvertures de chapitres**:

© MACIEJ NOSKOWSKI/iStockphoto;

p. 157: Wikipedia Commons;

p. 163: A. Wittmann/Wikipedia Commons;

p. 212: Wikipedia Commons;

p. 275: © TopFoto/Fotomas/The Image Works.

Des marques de commerce sont mentionnées ou illustrées dans cet ouvrage. L'Éditeur tient à préciser qu'il n'a reçu aucun revenu ni avantage conséquemment à la présence de ces marques. Celles-ci sont reproduites à la demande de l'auteur en vue d'appuyer le propos pédagogique ou scientifique de l'ouvrage.

Le matériel complémentaire mis en ligne dans notre site Web est réservé aux résidants du Canada, et ce, à des fins d'enseignement uniquement.

L'achat en ligne est réservé aux résidants du Canada.

À Lise, Francine, Chantale et Bruno

REMERCIEMENTS

Je tiens à remercier toutes les personnes qui ont collaboré à cette troisième édition de *Notions de statistique*.

J'exprime tout particulièrement ma gratitude à Suzanne Champagne, Éric Mauras et Olivier Rolko, qui m'ont accompagnée tout au long du processus d'édition. Ce fut un réel plaisir de travailler avec vous ! J'aimerais également souligner l'excellent travail de Nicole Blanchette à la révision linguistique et de Katie Delisle à la correction d'épreuves.

Je suis aussi reconnaissante aux professeurs suivants pour leurs commentaires ; ils m'ont permis de peaufiner l'ouvrage.

Les évaluateurs

Félix Baaden (Cégep Garneau), Jonathan Cantin (Université Laval), Ann-Sophie Pépin (Collège Montmorency), Audrey Samson (Collège Lionel-Groulx) et Matthieu Willems (Cégep André-Laurendeau).

Les consultants

Étienne Dauphin (Collège de Rosemont) et Maxime Savary (Collège Laflèche).

Les réviseurs scientifiques

Marc-Antoine Nadeau (Cégep Garneau) et Annie Turcotte (Cégep de Lévis-Lauzon).

Enfin, je suis redevable à mes étudiants qui, en me montrant qu'ils appréciaient mon approche de la matière, m'ont encouragée à raffiner, édition après édition, l'approche pédagogique qui particularise cet ouvrage.

Christiane Simard

AVANT-PROPOS

Le manuel *Notions de statistique* expose les diverses méthodes statistiques employées pour analyser des données, tester une hypothèse, effectuer un sondage ou encore contrôler la qualité d'un produit usiné. Dans un texte clair et accessible, l'auteure y présente les notions de base de la statistique descriptive, de la théorie des probabilités et de l'inférence statistique.

La **statistique descriptive** (souvent appelée, dans le langage populaire, «les statistiques») permet, comme son nom l'indique, de décrire un phénomène à l'aide de données portant sur ce même phénomène (les statistiques financières, économiques et sociales en sont des exemples). De façon plus précise, la statistique descriptive a pour objet l'application de diverses méthodes de présentation et d'analyse de données. C'est grâce à elle que l'on peut transformer des masses de données et de faits en une information concise facilement utilisable. Le chapitre 1 est consacré à l'étude de la statistique descriptive. À cette étude peut se greffer l'analyse des séries chronologiques présentée au chapitre 8.

La **théorie des probabilités** permet de calculer les risques lors d'une prise de décision dans une situation où le hasard intervient. La notion de probabilité est présentée au chapitre 2, alors que le chapitre 3 est consacré à l'étude des lois de probabilité.

L'**inférence statistique** permet pour sa part d'étudier une population au moyen des données d'un échantillon aléatoire tiré de cette population. Dans ce troisième volet de la statistique, on apprend comment estimer un paramètre d'une population, tester une hypothèse de recherche et étudier la relation entre deux variables à partir de données échantillonnelles. Par exemple, on voit comment, à partir d'un échantillon aléatoire de 1 000 électeurs, on peut prévoir les résultats d'une élection avec une certaine précision, ou encore comment on peut tester l'hypothèse voulant qu'il y ait un lien entre l'âge d'un conducteur et le risque d'accident. L'inférence statistique est étudiée en détail dans les chapitres 4, 5, 6 et 7.

Tout au long de l'ouvrage, les notions sont présentées selon une approche pédagogique originale caractérisée par une **présentation visuelle** des différents concepts. Afin de donner du sens à ces derniers, on privilégie le recours à une **approche intuitive**, notamment grâce à des **mises en situation**, avant la formalisation. En outre, des **exercices de compréhension** viennent soutenir l'apprentissage des étudiants en leur permettant de vérifier au fur et à mesure, en classe, leur compréhension de la matière. À cela s'ajoutent des exemples et des exercices diversifiés, la plupart du temps conçus à partir de **données réelles**. Des **exercices récapitulatifs**, un **résumé** et une liste des compétences à acquérir facilitent la révision des notions abordées dans chaque chapitre.

Il est reconnu que l'apprentissage est meilleur et plus intéressant lorsqu'on fait plutôt que lorsqu'on regarde faire ; aussi le manuel propose-t-il des **exemples à compléter** qui invitent les étudiants, par des questionnements, à participer activement à la construction de leurs connaissances.

CARACTÉRISTIQUES DE L'OUVRAGE



Objectifs du chapitre et du laboratoire

Au début de chaque chapitre sont énoncés les objectifs de formation visés ainsi que les objectifs du laboratoire associé au chapitre.

Mise en situation

La mise en situation est un problème qui permet d'apprivoiser graduellement une notion avant de passer à sa formalisation.

L'icône figurant dans certaines mises en situation annonce une question dont la réponse permettra de guider les étudiants dans la recherche d'une solution.

Exemple

Pour éviter un apprentissage par mimétisme sans réelle compréhension des notions abordées, nous n'avons pas misé sur la quantité des exemples, mais plutôt sur leur qualité et leur efficacité pédagogique. Pour créer un cours plus dynamique, nous avons choisi de ne pas écrire la solution de certains exemples afin d'inviter les étudiants à participer activement à la construction de la solution.

Exercices

Chaque série d'exercices comporte des problèmes de compréhension, des applications (dont un grand nombre sont basées sur des données réelles) et des questions portant sur des notions vues dans les chapitres antérieurs.

Afin d'éviter les problèmes répétitifs qui conduisent trop souvent à un apprentissage irréfléchi, chaque problème d'une série d'exercices aborde la matière sous un angle différent du problème précédent.

Exercices de compréhension

Ces exercices permettent de faire une pause dans le déroulement du cours afin que les étudiants et l'enseignant puissent mesurer le degré de compréhension de la matière.

Exercices de révision

Des exercices de révision d'application sont proposés au fur et à mesure des thématiques traitées.

1. **Les propriétés de l'hydrogène**

Étudiez les propriétés de l'hydrogène et démontrez que l'hydrogène est un élément très particulier.

 - a) Écrivez la liste des propriétés de l'hydrogène.
 - b) Expliquez pourquoi l'hydrogène possède ces propriétés.

EXERCICES RECAPITULATIFS

1. **LA CHIMIE DES ALCOOLS**

Le 22/02/2013, le Comité de Bioéthanol a publié une notice à destination des professionnels de la vente au détail et des consommateurs sur la vente et la consommation de boissons contenant de l'alcool.

Quelques extraits de cette notice :

 - a) Quels sont les alcools autorisés dans les boissons vendues au détail ?
 - b) Quels sont les types de boissons où l'alcool peut être ajouté ?
 - c) Quels sont les types de boissons où l'alcool ne peut pas être ajouté ?
 - d) Quels sont les produits que l'alcool ne peut pas remplacer ?
 - e) Quels sont les produits que l'alcool ne peut pas remplacer ?
 - f) Quels sont les produits que l'alcool ne peut pas remplacer ?
 - g) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - h) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - i) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - j) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - k) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - l) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - m) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - n) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - o) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - p) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - q) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - r) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - s) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - t) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - u) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - v) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - w) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - x) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - y) La consommation d'alcool peut-elle entraîner des maladies chroniques ?
 - z) La consommation d'alcool peut-elle entraîner des maladies chroniques ?

Exercices récapitulatifs

À la fin de chaque chapitre figurent des problèmes, présentés dans un ordre aléatoire, qui portent sur l'ensemble des notions vues dans le chapitre.

RÉSUMÉ DU CHAPITRE 4

Calcul de la probabilité d'un événement à l'aide de la définition

- Chaque $P(A) = \frac{n_A}{n}$, où n_A est le nombre d'éléments de l'échantillonage favorable à A .
- $P(A) = P(A^c) = 1 - P(A)$: somme des probabilités d'événements complements.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$: lorsque les deux événements sont mutuellement exclusifs.

Calcul de la probabilité d'un événement particulier

- $P(A) = \frac{n_A}{n}$
- $P(A) = \frac{1}{n}$
- $P(A) = \frac{\text{Nombre d'éléments dans } A}{\text{Nombre total d'éléments}}$
 - Utilisez la définition de probabilité si tous les éléments sont égaux.
 - $P(A) = \frac{n_A}{n} = \frac{n_A}{N}$ ou $P(A) = \frac{n_A}{N} = \frac{N_A}{N}$
- Si tous les résultats d'un échantillonage sont égaux, on utilise la règle d'échantillonage.
- $P(A) = \frac{\text{Nombre d'éléments dans } A}{\text{Nombre total d'éléments}}$
- $P(A) = \frac{\text{Nombre d'éléments dans } A}{\text{Nombre total d'éléments}}$
- $P(A) = \frac{\text{Nombre d'éléments dans } A}{\text{Nombre total d'éléments}}$
- $P(A) = \frac{\text{Nombre d'éléments dans } A}{\text{Nombre total d'éléments}}$

Calcul d'une probabilité conditionnelle

Probabilité d'événement B sachant que l'événement A a déjà eu lieu : $P(B|A) = \frac{P(A \cap B)}{P(A)}$ ou $P(B|A) = \frac{P(A \cap B)}{n_A}$

Complémentaire indépendance

• Ainsi si deux événements A et B ($A \cap B$) sont indépendants, alors la probabilité de l'événement A avec la conditionnalité B est égale à la probabilité de l'événement A sans la conditionnalité B .

Exercices de révision

1. Exposons une situation où l'événement A est indépendant de l'événement B .
2. Si deux événements sont complémentaires, sont-ils indépendants ?
3. Si deux événements sont mutuellement exclusifs, sont-ils indépendants ?
4. Si deux événements sont indépendants, sont-ils mutuellement exclusifs ?
5. Si deux événements sont indépendants, sont-ils complémentaires ?
6. Si deux événements sont indépendants, sont-ils toujours égaux ?
7. Si deux événements sont indépendants, sont-ils toujours égaux ?
8. Si deux événements sont indépendants, sont-ils toujours égaux ?

Résumé du chapitre

On trouve à la fin de chaque chapitre un résumé des notions étudiées.

Préparation à l'examen

Sous cette rubrique apparaît la liste des compétences attendues à la fin d'un chapitre, présentée sous la forme d'une liste de contrôle. Les étudiants peuvent s'en servir pour guider leur révision en vue d'une évaluation.

Laboratoires Excel

En complément au manuel *Notions de statistique*, les utilisateurs peuvent se procurer *Notions de statistique. Laboratoires Excel, 4^e édition*. Conçu pour accompagner le manuel *Notions de statistique*, ce document initie les étudiants, par la méthode du pas à pas, à l'utilisation d'Excel pour le traitement statistique de données. Chaque laboratoire est présenté sous la forme d'un travail personnel, qui peut être fait au collège ou à la maison, en utilisant l'une ou l'autre des versions suivantes du logiciel: Excel 2010, 2007 ou 2003 sous Windows et Excel 2011, 2008 ou 2004 sous Mac OS. Le tableau ci-dessous donne les sujets abordés ainsi qu'une estimation du temps de travail par laboratoire.

Laboratoire	Sujet	Durée approximative
1	Statistique descriptive	130 minutes ¹
2	Probabilités	60 minutes
3	Lois de probabilité	45 minutes
4	Inférence statistique	60 minutes
5	Test du khi-deux	75 minutes
6	Corrélation et régression linéaire	45 minutes

1. La durée inclut les 15 minutes requises pour la saisie des données.

TABLE DES MATIÈRES

Chapitre 1	
La statistique descriptive	1
1.1 La terminologie, les variables et les échelles de mesure	2
1.1.1 La terminologie	2
1.1.2 La classification des variables	3
1.1.3 Les échelles de mesure	6
Exercices 1.1	10
1.2 Les tableaux de distribution et les représentations graphiques	11
1.2.1 La présentation d'une variable qualitative	13
1.2.2 La présentation d'une variable quantitative discrète	18
1.2.3 La présentation d'une variable quantitative continue	21
1.2.4 L'ogive ou la courbe de fréquences cumulées	32
1.2.5 Quel graphique faut-il construire?	34
Exercices 1.2	35
1.3 Les mesures de tendance centrale	39
1.3.1 La moyenne	39
1.3.2 Le mode et la classe modale	47
1.3.3 La médiane	50
1.4 Les mesures de position (quantiles)	56
Exercices 1.3	58
1.5 Les mesures de dispersion	61
1.5.1 L'étendue	61
1.5.2 La variance et l'écart type	62
1.5.3 Le coefficient de variation	68
1.5.4 L'utilisation du mode statistique de la calculatrice	70
Guide d'utilisation de la calculatrice scientifique de base	70
Guide d'utilisation de la calculatrice graphique	71
Exercices 1.4	72
1.6 Les mesures de position (cote z)	74
Exercices 1.5	79
Résumé du chapitre 1	81
Exercices récapitulatifs	83
Préparation à l'examen	86
Chapitre 2	
Les probabilités	87
2.1 Le lien entre les probabilités et l'inférence statistique	88
2.2 La terminologie	89
2.3 Le calcul d'une probabilité	89
2.4 Les événements particuliers	92
2.5 Les propriétés des probabilités	94
Exercices 2.1	98
2.6 La probabilité conditionnelle	100
2.7 Les événements indépendants	103
2.8 La règle de multiplication	108
Exercices 2.2	115
2.9 L'analyse combinatoire	118
2.9.1 Le principe de multiplication	118
2.9.2 Les permutations et les arrangements	119
2.9.3 Les combinaisons	121
Exercices 2.3	124
Résumé du chapitre 2	126
Exercices récapitulatifs	127
Préparation à l'examen	129
Chapitre 3	
Les lois de probabilité	131
3.1 Les variables aléatoires	132
3.1.1 Le concept de variable aléatoire	133
3.1.2 La distribution de probabilité	134
3.1.3 L'espérance et l'écart type d'une variable aléatoire	140
3.1.4 L'espérance et les jeux de hasard	142
Exercices 3.1	144

3.2 La loi binomiale	145	Exercices 4.2	201
3.2.1 Le contexte d'une expérience aléatoire binomiale	145	4.3 L'estimation de la moyenne d'une population	203
3.2.2 La fonction de probabilité d'une loi binomiale	147	4.3.1 L'échantillon de grande taille ($n \geq 30$)	203
Exercices 3.2	151	4.3.2 L'échantillon de petite taille ($n < 30$)	211
3.2.3 L'espérance et l'écart type d'une loi binomiale	151	4.3.3 Le choix de la taille de l'échantillon	213
Exercices 3.3	155	Exercices 4.3	216
3.3 La loi de Poisson	157	4.4 L'estimation d'un pourcentage d'une population	219
3.3.1 Le contexte d'une expérience aléatoire de Poisson	157	4.4.1 La distribution d'échantillonnage d'un pourcentage	219
3.3.2 L'approximation de la loi binomiale par la loi de Poisson	160	4.4.2 L'estimation d'un pourcentage par intervalle de confiance	224
Exercices 3.4	161	4.4.3 Le choix de la taille de l'échantillon	228
3.4 La loi normale	162	4.4.4 La répartition des indécis	230
3.4.1 Un rappel	163	Exercices 4.4	232
3.4.2 La courbe normale ou courbe de Gauss	163	Résumé du chapitre 4	235
3.4.3 Le calcul d'une probabilité pour une loi normale	165	Exercices récapitulatifs	237
3.4.4 La loi normale centrée réduite $N(0; 1)$	168	Préparation à l'examen	238
Exercices 3.5	171		
3.4.5 La loi normale comme modèle mathématique	172		
3.4.6 L'approximation de la loi binomiale par la loi normale	175		
Exercices 3.6	179		
Résumé du chapitre 3	181		
Exercices récapitulatifs	182		
Préparation à l'examen	184		
Chapitre 4			
La distribution d'échantillonnage et l'estimation	185		
4.1 L'échantillonnage	186	Les tests paramétriques	241
4.1.1 Pourquoi faire des sondages?	186	5.1 Le test d'hypothèse sur une moyenne	242
4.1.2 L'historique du sondage	186	Exercices 5.1	250
4.1.3 Comment choisir un échantillon?	187	5.2 Le test d'hypothèse sur un pourcentage	251
Méthodes pour générer des nombres aléatoires	187	Exercices 5.2	254
Exercices 4.1	191	5.3 Le test d'hypothèse sur l'égalité de deux paramètres	255
4.2 La distribution d'échantillonnage d'une moyenne	191	5.3.1 Le test sur l'égalité de deux moyennes	256
		5.3.2 Le test sur l'égalité de deux pourcentages	260
		Exercices 5.3	262
		Résumé du chapitre 5	264
		Exercices récapitulatifs	265
		Préparation à l'examen	267
Chapitre 6			
Les tests du khi-deux	269		
6.1 Le test d'ajustement du khi-deux	270		

6.1.1	La construction d'un test d'ajustement du khi-deux	271
Utilisation des touches de mémoire pour calculer le khi-deux		277
6.1.2	La représentativité d'un échantillon	279
6.1.3	La population est-elle conforme au modèle normal ?	280
Exercices 6.1		284
6.2	Le test d'indépendance du khi-deux	286
6.2.1	La construction d'un test d'indépendance du khi-deux	288
Exercices 6.2		293
Résumé du chapitre 6		296
Exercices récapitulatifs		297
Préparation à l'examen		298

Chapitre 7

La corrélation et la régression linéaire		299
7.1	La corrélation linéaire	300
7.1.1	Le diagramme de dispersion (ou nuage de points)	300
7.1.2	La corrélation	301
7.1.3	Le coefficient de corrélation linéaire	302
7.2	La régression linéaire	305
7.2.1	La droite de régression	305
7.2.2	Le coefficient de détermination	307
7.2.3	L'utilisation du mode statistique de la calculatrice (deux variables) ...	309
Guide d'utilisation de la calculatrice scientifique de base (deux variables)		309
Guide d'utilisation de la calculatrice graphique (deux variables)		310

Exercices 7.1		310
Résumé du chapitre 7		313
Exercice récapitulatif		313
Préparation à l'examen		314

Chapitre 8		
Les séries chronologiques		315
8.1	La définition et la représentation d'une série chronologique	316
8.2	Les composantes d'une série chronologique	318
8.3	Le lissage d'une série chronologique	321
8.3.1	Le lissage par moyenne mobile	323
8.3.2	Le lissage exponentiel	325
8.4	Les séries désaisonnalisées	327
Exercices 8.1		329
Résumé du chapitre 8		333
Exercices récapitulatifs		334
Préparation à l'examen		336

Annexe 1	Table de la loi binomiale	337
Annexe 2	Table de la loi de Poisson	342
Annexe 3	Table de la loi normale centrée réduite	347
Annexe 4	Table de la loi de Student	348
Annexe 5	Table de la loi du khi-deux	349

Réponses aux exercices		350
Bibliographie		386
Index		387

1

Chapitre

La statistique descriptive



OBJECTIFS DU CHAPITRE

Appliquer la démarche servant à traiter et à analyser une série de données :

- déterminer le type de variable et l'échelle de mesure ;
- construire un tableau de distribution ;
- représenter graphiquement une distribution ;
- calculer et interpréter différentes mesures.

OBJECTIFS DU LABORATOIRE

Le laboratoire 1 vise à utiliser Excel pour traiter et analyser les données d'une variable qualitative et d'une variable quantitative continue.

La statistique est le champ d'études des mathématiques consacré à l'analyse de données. Elle se divise en trois branches : la statistique descriptive, qui s'intéresse à la présentation et à l'analyse de données ; les lois de probabilité, qui servent à modéliser des situations où le hasard intervient ; et l'inférence statistique, qui permet d'étudier une population au moyen des données d'un échantillon.

La statistique descriptive a pour objet l'application de méthodes de traitement de données visant à dégager les caractéristiques d'une masse de données. Dans le présent chapitre, nous présentons les règles à suivre pour construire des tableaux et des graphiques qui permettent d'effectuer une première analyse des données. Cette analyse est ensuite approfondie par le calcul de mesures statistiques prises sur les données.

1.1 La terminologie, les variables et les échelles de mesure

1.1.1 La terminologie

Comme toute science, la statistique possède son propre vocabulaire, qu'il convient de définir avant d'en aborder l'étude.

Population et recensement

Une **population** est l'ensemble de toutes les personnes, de tous les objets ou de tous les faits sur lesquels porte une étude. Chaque élément d'une population est appelé **unité statistique**.

Un **recensement** est une étude réalisée sur toutes les unités statistiques d'une population.

Échantillon et sondage

Un **échantillon** est un sous-ensemble d'unités de la population sur lesquelles on effectue une étude. Si l'échantillon est choisi au hasard, on peut généraliser certains résultats à l'ensemble de la population. Dans le cas contraire, on ne peut pas le faire.

Un **sondage** est une enquête menée auprès d'un échantillon de la population que l'on désire étudier.

EXEMPLE 1

Dans un sondage, on interroge 1 000 électeurs, choisis au hasard parmi tous les électeurs du Québec, afin de connaître leur intention de vote.

Population étudiée : Tous les électeurs du Québec.

Unité statistique : Un électeur.

Échantillon : Les 1 000 électeurs choisis au hasard.

EXEMPLE 2

On veut vérifier si le volume moyen de jus versé par une machine dans des contenants durant la dernière heure de production est conforme au volume désiré. Pour ce faire, on prélève au hasard 20 contenants parmi ceux qui ont été remplis durant cette heure et l'on mesure le volume de jus versé dans chacun.

Population étudiée : Tous les contenants de jus remplis durant la dernière heure de production.

Unité statistique : Un contenant de jus.

Échantillon : Les 20 contenants de jus prélevés au hasard parmi ceux qui ont été remplis durant la dernière heure de production.

EXEMPLE 3

Chaque année, le ministère de la Sécurité publique dresse le bilan statistique des infractions criminelles commises au Québec en utilisant les données inscrites sur un formulaire par le policier qui rapporte l'infraction.

Population étudiée : Les infractions criminelles commises au Québec pour l'année considérée.

Unité statistique : Une infraction criminelle.

Échantillon : Comme il s'agit d'un recensement (on étudie toutes les infractions criminelles rapportées), il n'y a pas d'échantillon.

Variable

Une **variable** est une caractéristique de l'unité statistique que l'on désire étudier. Sa valeur peut différer d'une unité statistique à l'autre. Il est possible d'associer plus d'une variable à une même unité statistique. On emploie une lettre majuscule (X , Y ou Z) pour désigner une variable, et une lettre minuscule (x , y ou z) pour désigner sa valeur.

EXEMPLE

- Si l'unité statistique considérée est un travailleur, la variable pourrait être l'âge, le sexe ou le salaire.
- Si l'unité statistique est un cégep, la variable pourrait être le nombre d'étudiants, la langue d'enseignement ou le nombre de programmes de la formation technique offerts.

1.1.2 La classification des variables

On distingue deux types de variables : les variables quantitatives et les variables qualitatives.

Une variable est dite **quantitative** lorsque ses valeurs possibles sont des nombres (quantité). Sinon, elle est dite **qualitative** (qualité) et ses valeurs sont alors des **catégories** ou des **modalités**.

NOTE

Si, dans une question portant sur une variable précise, on désire dresser la liste des valeurs ou des catégories de la variable, cette liste doit avoir les caractéristiques suivantes :

- Elle doit être **exhaustive**: cela signifie qu'elle doit inclure tous les cas. Il faut souvent recourir à une catégorie «autres» pour les réponses que l'on ne peut pas prévoir.
- Elle doit être **exclusive**: cela signifie qu'elle doit exclure toute possibilité de chevauchement ou de recouplement entre les valeurs ou les catégories.

EXEMPLE

Pour chacune des questions suivantes, nommer la variable, donner ses valeurs ou ses catégories et dire si elle est quantitative ou qualitative.

a) «Depuis combien d'années êtes-vous mariés?»

Variable : Le nombre d'années de mariage.
Valeurs : Un nombre compris dans l'un des intervalles suivants :
moins de 10 ans [10 ans ; 20 ans[[20 ans ; 30 ans[30 ans et plus.
Type de variable : Variable quantitative.

b) «Quel est votre sexe?»

Variable : Le sexe du répondant.
Catégories : Féminin, masculin.
Type de variable : Variable qualitative.

c) «Combien d'enfants avez-vous?»

Variable : Le nombre d'enfants.
Valeurs : 0, 1, 2, 3, 4, 5 et plus.
Type de variable : Variable quantitative.

d) «Vivez-vous dans une famille monoparentale, biparentale traditionnelle ou reconstituée?»

Variable : Le type de famille.
Catégories : Monoparentale, biparentale traditionnelle, reconstituée.
Type de variable : Variable qualitative.

NOTE

Certaines variables peuvent être considérées comme quantitatives ou qualitatives. Par exemple, la taille d'un chandail est une variable quantitative si elle prend l'une des valeurs 8 ans, 10 ans, 12 ans..., mais c'est une variable qualitative si elle prend l'une des catégories petit, moyen, grand.

Types de variables qualitatives

Une variable qualitative est **ordinale** si l'on peut établir une relation d'ordre entre les catégories de la variable ; sinon, elle est **nominale**.

EXEMPLE

a) Les catégories suggérées pour la question «En général, quel est votre état de santé?» sont:

Excellent Très bon Bon Passable Mauvais

La variable «état de santé» est une variable qualitative ordinale.

b) Les catégories suggérées pour la question «Quel est votre état civil?» sont:

Célibataire Séparé(e) Marié(e) Veuf (veuve) Divorcé(e) Conjoint(e) de fait

La variable «état civil» est une variable qualitative nominale.

c) Les catégories suggérées pour la question «À quelle fréquence avez-vous ressenti des moments de solitude au cours de la dernière semaine?» sont:

Jamais De temps en temps Assez souvent Très souvent

La variable «fréquence des moments de solitude au cours de la dernière semaine» est une variable qualitative ordinale.

Types de variables quantitatives

Une variable quantitative est dite **continue** si elle peut en théorie prendre n'importe laquelle des valeurs contenues dans un intervalle donné de nombres réels; sinon, elle est dite **discrète**. En d'autres termes, il est possible d'augmenter la précision de la mesure d'une variable quantitative continue, mais pas celle d'une variable quantitative discrète.

EXEMPLE

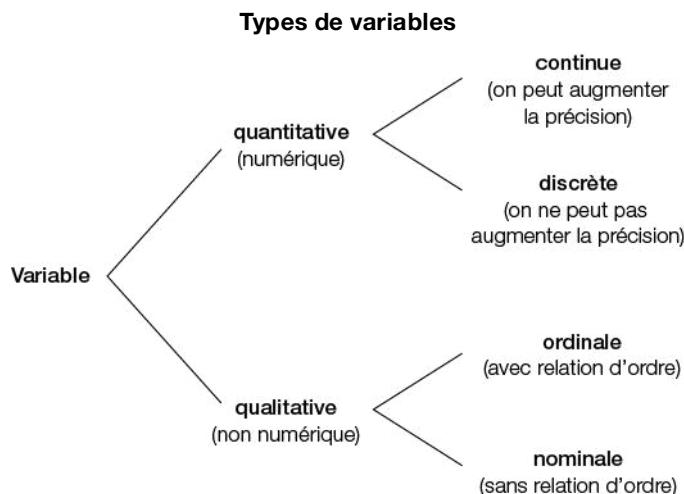
a) Les variables suivantes sont des variables quantitatives continues:

- L'âge d'une personne : quand une personne déclare être âgée de 18 ans, en fait son âge réel est un nombre entre 18 ans et 0 jour et 18 ans et 364 jours, soit un nombre situé quelque part dans l'intervalle [18 ans ; 19 ans[. En réalité, cette personne a peut-être 18,43 ans.
- La taille d'un individu : quand on dit qu'un individu mesure 172 cm, ce résultat est approximatif; sa vraie taille est en fait un nombre entre 171,5 cm et 172,5 cm, donc compris dans l'intervalle [171,5 cm ; 172,5 cm[. Pour obtenir la taille exacte, il faudrait utiliser un instrument de mesure beaucoup plus précis qu'un ruban gradué en centimètres.
- Le temps nécessaire à un traversier pour faire la navette entre Québec et Lévis; par exemple, 12,32 minutes.
- Le nombre moyen d'enfants par famille; par exemple, 9 familles ayant au total 15 enfants donne une moyenne de 1,7 enfant par famille.
- Le pourcentage de filles dans un programme du cégep; par exemple, 60 filles dans un groupe de 90 étudiants donne un pourcentage de 66,7 % de filles.

b) Les variables suivantes sont des variables quantitatives discrètes:

- Le nombre d'enfants dans une famille : 0, 1, 2, ...
- Le nombre de téléviseurs dans un ménage : 0, 1, 2, ...
- Le nombre de passagers dans une automobile : 1, 2, 3, ...
- La pointure des souliers d'un adulte : 6, 6½, 7, 7½, ...

Le schéma suivant illustre la subdivision des variables selon leurs types.



1.1.3 Les échelles de mesure

On distingue plusieurs types d'échelles de mesure, soit les échelles nominale, ordinale, d'intervalle et de rapports.

Échelle nominale

On associe souvent un code numérique à chaque catégorie d'une variable qualitative nominale afin de faciliter la manipulation des données. On dit alors que l'on emploie une **échelle nominale** pour coder les catégories. Par exemple, on peut attribuer le code 0 à la catégorie «masculin» et le code 1 à la catégorie «féminin» de la variable «sex». Ces codes servent uniquement à différencier les catégories, on ne peut pas les utiliser pour établir une relation d'ordre entre celles-ci : dire que 0 est plus petit que 1 ($0 < 1$) serait faux dans ce contexte. On ne peut pas non plus effectuer d'opérations mathématiques sur les codes. À titre d'exemple, si la série 0, 1, 1, 1 correspond au sexe de quatre personnes, dire que $(0 + 1 + 1 + 1) \div 4 = 0,75$ donne la moyenne du sexe des quatre personnes n'a aucun sens.

À RETENIR

Une échelle nominale sert uniquement à différencier les catégories d'une variable au moyen de codes. Elle ne permet pas d'établir une relation d'ordre entre les codes, ni d'effectuer des opérations arithmétiques (+, -, ×, ÷) sur ceux-ci.

Échelle ordinale

Si l'on assigne un code à chaque catégorie d'une variable qualitative ordinale, on dit que l'on emploie une **échelle ordinale** pour le codage. Par exemple, à la question «Aimez-vous les études?», les catégories pourraient être codées ainsi :

1. Pas du tout
2. Un peu
3. Moyennement
4. Beaucoup

Dans ce cas, on peut établir une relation d'ordre entre les codes et affirmer que $1 < 2$ ou que $4 > 3$.

On utilise aussi une échelle ordinaire dans le cas d'une variable quantitative dont les valeurs sont présentées sous forme de catégories codées. Par exemple, à la question «Quel est votre salaire?», on a codé les réponses ainsi :

1. 29 999 \$ et moins
2. Entre 30 000 \$ et 59 999 \$
3. 60 000 \$ et plus

Les codes permettent d'affirmer que les revenus de code 1 sont inférieurs à ceux de code 2, mais on ne peut pas faire de calculs avec ces codes. À titre d'exemple, si la série 1, 2, 2, 4 correspond au salaire de quatre personnes, l'opération $(1 + 2 + 2 + 4) \div 4 = 2,25$ ne peut pas correspondre à la moyenne des salaires des quatre personnes.

À RETENIR

Une échelle ordinaire sert à différencier les catégories d'une variable et à les ordonner selon des codes. Elle ne permet pas d'effectuer des opérations arithmétiques (+, -, ×, ÷) sur les codes.

Échelle d'intervalle

On se sert d'une **échelle d'intervalle** uniquement pour des variables quantitatives. Les échelles de température graduées en degrés Celsius ou Fahrenheit emploient une échelle d'intervalle. Ce type d'échelle est caractérisé par le fait que le zéro est fixé par convention ; il sert uniquement de point de repère fixe à partir duquel on prend des mesures. Ainsi, la valeur 0 n'indique pas l'absence de la caractéristique mesurée. Par exemple, pour une température mesurée en degrés Celsius, 0 ne signifie pas une absence de température, mais correspond au point de congélation de l'eau. L'heure est un autre exemple d'échelle d'intervalle : 0 h n'indique pas un moment qui n'existe pas dans la journée ; il indique plutôt le début de la journée. L'heure 0 est un point de repère fixé par convention pour mesurer le temps.

En plus de servir à différencier et à ordonner les valeurs, une échelle d'intervalle permet de mesurer l'écart entre deux valeurs, mais pas d'établir de rapport entre elles. Par exemple, pour des températures extérieures, si l'on compare 2 °C et 6 °C, on peut dire qu'il y a un écart de 4 °C entre ces deux températures. Par contre, si l'on construit le rapport 6 °C/2 °C, le quotient 3 obtenu ne permet pas d'affirmer qu'à 6 °C, la température extérieure est 3 fois plus chaude qu'à 2 °C. Il suffit d'exprimer les mêmes températures en Fahrenheit pour s'en convaincre. On sait que 6 °C correspond à 43 °F et que 2 °C correspond à 36 °F. En construisant le rapport 43 °F/36 °F, on obtient 1,2 et non 3, ce qui contredit l'affirmation précédente.

Le calendrier grégorien utilise une échelle d'intervalle pour mesurer le temps. L'an zéro, qui ne figure pas dans ce calendrier, mais qui est utilisé dans d'autres calendriers, est fixé par convention et ne signifie pas l'absence de temps. On peut dire qu'une personne née en 1970 est de 10 ans plus âgée qu'une personne née en 1980, mais le rapport 1980/1970 n'a pas de sens.

À RETENIR

Une échelle d'intervalle sert à différencier les valeurs de la variable, à les ordonner et à mesurer les écarts entre celles-ci, mais elle ne permet pas d'établir de rapport entre les valeurs. L'addition et la soustraction (+, -) sont permises.

Échelle de rapport

On emploie une **échelle de rapport** avec des variables quantitatives. Dans cette échelle, le zéro est absolu, c'est-à-dire qu'il n'est pas fixé par convention : la valeur 0 signifie l'absence de la caractéristique mesurée. De plus, on peut comparer les valeurs en mesurant l'écart ou en établissant un rapport entre celles-ci. Par exemple, on utilise une échelle de rapport pour mesurer le nombre d'échecs d'un étudiant. La valeur 0 indique qu'un étudiant n'a aucun échec. Un étudiant qui a 6 échecs en a 4 de plus que celui qui en a 2 ; on peut aussi dire qu'il en a 3 fois plus (6/2).

À RETENIR

Une échelle de rapport sert à différencier et à ordonner les valeurs d'une variable. Elle permet aussi de comparer les valeurs en mesurant l'écart ou en établissant un rapport entre elles. Toutes les opérations arithmétiques ($+$, $-$, \times , \div) sont permises. C'est l'échelle qui offre le plus de possibilités.

Utilité des échelles de mesure

Le choix d'une échelle de mesure est déterminant pour l'analyse d'une variable : plus une mesure est précise, plus l'analyse des résultats pourra être raffinée. Par exemple, une échelle de rapport permet une analyse plus détaillée des données qu'une échelle nominale ou ordinale. Il faut donc toujours tenir compte de la précision recherchée pour la caractéristique mesurée lors du choix d'une échelle de mesure.

EXEMPLE

Supposons qu'un chercheur désire obtenir de l'information sur la consommation de cigarettes chez les jeunes. Pour ce faire, il peut utiliser l'une ou l'autre des questions suivantes :

Q1: Fumez-vous la cigarette ?

1. Oui 2. Non

On utilise ici une échelle nominale. À l'analyse des données, ce type de mesure permet uniquement de calculer le pourcentage de fumeurs ou de non-fumeurs parmi les répondants.

Q2: À quelle fréquence fumez-vous la cigarette ?

1. Jamais 2. Rarement 3. Occasionnellement 4. Régulièrement

Dans cette question, on utilise une échelle ordinale qui permettra une analyse un peu plus détaillée que l'échelle choisie à la question Q1. En effet, on peut calculer le pourcentage de fumeurs parmi les répondants et répartir ces fumeurs en trois types.

Q3: Généralement, combien de cigarettes fumez-vous par jour ?

1. 0 2. De 1 à 10 3. De 11 à 25 4. 26 et plus

On emploie ici aussi une échelle ordinale qui permettra de faire le même genre d'analyse qu'à la question Q2. Toutefois, l'utilisation d'une variable quantitative permet de mieux distinguer les différents types de fumeurs.

Q4: Généralement, combien de cigarettes fumez-vous par jour ? _____

Pour cette question, on emploie une échelle de rapport. Ce type d'échelle permettra une analyse très raffinée des réponses obtenues. On pourra calculer le pourcentage de fumeurs parmi les répondants, le pourcentage de fumeurs selon les différents types de fumeurs qu'il nous plaira de définir, le nombre moyen de cigarettes fumées par jour, etc.

Cet exemple illustre bien comment le choix de réponses offert à une question est déterminant pour l'analyse d'une variable.

Il ne faut pas conclure de ce qui précède que l'échelle de rapport est toujours le meilleur choix : il peut arriver qu'un répondant soit incapable de donner une réponse précise à la question «Combien de minutes avez-vous consacrées à répondre à vos courriels hier?». Un autre pourrait juger indiscret qu'on lui demande son âge, mais accepter d'indiquer le groupe d'âge auquel il appartient.

Le tableau suivant présente une comparaison des échelles de mesure.

Comparaison des échelles de mesure

Échelle	Caractéristiques spécifiques	Opérations permises
Nominale	<ul style="list-style-type: none"> Requiert une variable qualitative nominale. Permet uniquement de différencier les catégories ($=$, \neq). 	Aucune
Ordinal	<ul style="list-style-type: none"> La variable peut être qualitative ou quantitative. Permet de différencier et d'ordonner les catégories ou les valeurs ($=$, \neq, $<$, $>$). 	Aucune
Intervalle	<ul style="list-style-type: none"> Requiert une variable quantitative. Le zéro est fixé par convention: il n'indique pas l'absence de la caractéristique. Permet de différencier et d'ordonner les valeurs ($=$, \neq, $<$, $>$). L'écart entre deux valeurs a un sens, mais pas le rapport entre celles-ci. 	$+, -$
De rapport	<ul style="list-style-type: none"> Requiert une variable quantitative. Le zéro est absolu: il indique l'absence de la caractéristique. Permet de différencier et d'ordonner les valeurs ($=$, \neq, $<$, $>$). L'écart et le rapport entre deux valeurs ont un sens. 	$+, -, \times, \div$

EXERCICES DE COMPRÉHENSION | 1.1

Les réponses figurent en fin d'ouvrage.

1. Pour chacune des questions suivantes, donner le type de la variable étudiée et l'échelle de mesure.

a) Êtes-vous marié(e)? 1. Oui 2. Non

Type: _____ . Échelle _____ .

b) En quelle année vous êtes-vous marié(e)?

Type: _____ . Échelle _____ .

c) Depuis combien de temps êtes-vous marié(e)?

Type: _____ . Échelle _____ .

d) Dans quelle mesure êtes-vous satisfait(e) de votre relation de couple?

1. Insatisfait(e) 2. Satisfait(e) 3. Très satisfait(e)

Type: _____ . Échelle _____ .

e) Combien de partenaires sexuels autres que votre conjoint(e) avez-vous eus depuis deux ans?

1. 0 2. 1 3. 2 ou 3 4. 4 et plus

Type: _____ . Échelle _____ .

➤ 2. Après avoir visité quatre Centres de la petite enfance (CPE), un inspecteur a noté les informations suivantes.

		CPE 1	CPE 2	CPE 3	CPE 4
A	Heure d'arrivée sur les lieux	8 h	10 h	13 h	15 h
B	Nombre d'éducatrices	6	3	6	8
C	Évaluation du programme d'activités 1. Très bon 2. Bon 3. Insuffisant	2	3	1	1
D	Durée de la visite (en minutes)	50	100	75	75

Vrai ou faux ?

- a) On peut comparer les données de la ligne A en mesurant leurs écarts ou en établissant un rapport entre elles. _____
- b) Il serait insensé de calculer la moyenne des quatre données de la ligne C. _____
- c) On ne peut pas établir un rapport pour comparer les données de la ligne B. _____
- d) Seule la variable de la ligne D est quantitative continue. _____

EXERCICES 1.1

Les réponses figurent en fin d'ouvrage.

1. Pour chacune des quatre études suivantes :

- a) Décrire la population étudiée.
- b) Décrire l'échantillon.
- c) Décrire l'unité statistique.
- d) Nommer la variable étudiée.
- e) Décrire l'ensemble des catégories ou des valeurs de la variable.
- f) Donner le type de variable étudiée.
 - i) Un sondage est effectué auprès de 200 citoyens de la ville de Québec afin de connaître leur chaîne de télévision favorite.
 - ii) Dans une étude portant sur l'évolution de la situation économique au Québec de 2000 à 2010, on s'intéresse au taux de chômage annuel de cette décennie.
 - iii) Afin de déterminer le profil socioéconomique des ménages d'un quartier de Montréal, on a noté le nombre d'enfants par ménage pour un échantillon de 380 ménages.
 - iv) Selon les données du recensement de 2011, 78,1 % de Québécois ont le français seulement comme langue maternelle, 7,7 % ont l'anglais

seulement, 12,3 % n'ont ni le français ni l'anglais comme langue maternelle et 2 % ont déclaré avoir plusieurs langues maternelles.

Source: Statistique Canada. Recensement 2011.

2. Donner le type de la variable.

- a) La superficie des lacs du Québec.
- b) Le pays d'origine des immigrants.
- c) Le nombre d'employés dans une entreprise.
- d) Le diamètre d'une tige.
- e) Possédez-vous une automobile ?

1. Oui 2. Non

f) Ressentez-vous du stress avant un examen ?

1. Toujours	3. Parfois	5. Jamais
2. Souvent	4. Rarement	

3. Donner le type de variable et l'échelle de mesure.

- a) Avez-vous échoué à des cours à votre premier trimestre au cégep ?
- 1. Non 2. Oui
- b) À combien de cours avez-vous échoué à votre premier trimestre au cégep ?
- 1. 0 2. 1 3. 2 ou 3 4. 4 et plus

- c) À combien de cours avez-vous échoué à votre premier trimestre au cégep? _____
- d) Parmi les taux suivants, lequel correspond à votre taux d'échec à votre premier trimestre au cégep?
- Taux d'échec: $\frac{n^{\text{bre}} \text{ de cours non réussis}}{n^{\text{bre}} \text{ de cours suivis}}$
1. 0 %
 2. De 1 % à 15,9 %
 3. De 16 % à 49,9 %
 4. 50 % et plus
- e) Indiquez votre taux d'échec à votre premier trimestre au cégep. _____

- f) Indiquez votre degré d'accord avec l'affirmation suivante: «Les étudiants qui ont échoué à plus de la moitié de leurs cours ne devraient pas être admis au trimestre suivant.»
1. Fortement en désaccord
 2. En désaccord
 3. D'accord
 4. Fortement d'accord
- g) Quelle est votre année de naissance ?
4. Sur le site Internet de MétéoMédia, pour chaque jour de la semaine, on peut trouver les informations suivantes: heure du lever du soleil et vitesse des vents. Pour chacune de ces deux variables, donner le type de variable et l'échelle de mesure.

1.2 Les tableaux de distribution et les représentations graphiques

Afin de faciliter l'analyse des données recueillies lors d'une étude, on les groupe par valeur, par catégorie ou par classe, puis on les présente sous forme de tableaux ou de graphiques. Dans la présente section, nous apprendrons à construire les tableaux et les graphiques appropriés à chaque type de variable.

La mise en situation suivante servira à présenter les notions abordées dans les sections 1.2.1 et 1.2.2.

MISE EN

SITUATION

En 2010, on a colligé les informations suivantes pour chacun des 48 collèges publics du Québec :

- Le nombre d'étudiants (y compris les étudiants en formation continue)
- Le nombre de programmes en techniques administratives¹ offerts
- Le pourcentage d'étudiants en formation technique
- La formation prédominante, que l'on définit ainsi :
 1. Préuniversitaire (au moins 55 % des étudiants sont en formation préuniversitaire)
 2. Technique (au moins 55 % des étudiants sont en formation technique)
 3. Aucune (de 45 % à 55 % des étudiants sont en formation technique ou préuniversitaire)

Parmi les variables retenues, on compte deux variables quantitatives discrètes (nombre d'étudiants et nombre de programmes techniques offerts), une variable quantitative continue (pourcentage d'étudiants en formation technique) et une variable qualitative nominale (formation prédominante).

On recourt à une échelle de rapport pour les trois premières variables et à une échelle nominale pour la dernière.

1. Il y a six programmes en techniques administratives : Logistique du transport, Comptabilité et gestion, Conseil en assurances et services financiers, Gestion de commerces, Techniques de bureautique et Techniques de l'informatique.

Série statistique

On donne le nom de **série statistique** à l'ensemble des données brutes recueillies pour chacune des variables étudiées. Voici les séries statistiques pour les quatre variables de la mise en situation.

Nom du collège	Nombre d'étudiants	Nombre de programmes en techniques administratives	Pourcentage d'étudiants en formation technique	Formation prédominante
1. Abitibi-Témiscamingue	2 884	2	63,8 %	2
2. Ahuntsic	9 036	3	71,6 %	2
3. Alma	1 427	2	55,9 %	2
4. André-Laurendeau	3 581	5	59,1 %	2
5. Baie-Comeau	850	1	64,8 %	2
6. Beauce-Appalaches	2 003	2	60,2 %	2
7. Bois-de-Boulogne	3 321	2	42,1 %	1
8. Champlain	5 172	4	23,1 %	1
9. Chicoutimi	3 180	3	67,1 %	2
10. Dawson	10 096	4	44,5 %	1
11. Drummondville	2 259	5	54,1 %	3
12. Édouard-Montpetit	7 605	4	50,0 %	3
13. François-Xavier-Garneau	6 243	5	56,6 %	2
14. Gaspésie et des Îles	1 230	3	63,4 %	2
15. Gérald-Godin	1 482	2	42,6 %	1
16. Granby-Haute-Yamaska	2 232	3	59,5 %	2
17. Héritage	1 110	3	46,6 %	3
18. John Abbott	6 706	4	34,6 %	1
19. Jonquière	3 775	4	77,8 %	2
20. Lanaudière	6 328	6	49,3 %	3
21. La Pocatière	1 227	2	75,6 %	2
22. Lévis-Lauzon	3 321	5	62,4 %	2
23. Limoilou	6 189	4	56,0 %	2
24. Lionel-Groulx	5 751	5	41,1 %	1
25. Maisonneuve	6 758	4	51,1 %	3
26. Marie-Victorin	5 551	2	72,8 %	2
27. Matane	735	2	78,6 %	2
28. Montmorency	7 164	5	56,7 %	2
29. Outaouais	5 618	4	52,9 %	3
30. Rimouski	3 332	5	72,1 %	2
31. Rivière-du-Loup	1 420	2	74,8 %	2
32. Rosemont	6 504	4	66,1 %	2
33. Sainte-Foy	7 360	4	52,2 %	3
34. Saint-Félicien	1 316	2	67,1 %	2
35. Saint-Hyacinthe	4 186	4	61,1 %	2
36. Saint-Jean-sur-Richelieu	3 850	5	55,2 %	2
37. Saint-Jérôme	4 699	3	63,8 %	2
38. Saint-Laurent	3 768	0	62,4 %	2
39. Sept-Îles	803	3	65,0 %	2
40. Shawinigan	1 396	3	64,7 %	2
41. Sherbrooke	7 074	4	59,1 %	2
42. Sorel-Tracy	1 314	3	68,3 %	2
43. Thetford	946	4	69,7 %	2
44. Trois-Rivières	4 796	4	56,3 %	2
45. Valleyfield	2 533	4	63,4 %	2
46. Vanier	7 918	5	41,0 %	1
47. Victoriaville	1 658	2	59,8 %	2
48. Vieux-Montréal	7 792	4	67,2 %	2

Source: Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

1.2.1 La présentation d'une variable qualitative

Dans la mise en situation, considérons la série statistique de la variable qualitative « formation prédominante » dont les trois catégories sont codées ainsi :

1. Préuniversitaire
2. Technique
3. Aucune

Tableau de distribution d'une variable qualitative

Pour faciliter l'analyse de la variable « formation prédominante », il faut grouper les données par catégorie dans un **tableau de distribution** (ou **tableau de fréquences**).

Pour construire un tel tableau, on énumère les catégories de la variable, puis on fait correspondre à chacune d'entre elles le nombre ou le pourcentage de données de la série statistique qui en font partie. La première ligne du tableau contient les titres de colonnes et la dernière, le total. On dit que le tableau construit donne la distribution de la variable étudiée.

Voici le tableau de distribution de la variable « formation prédominante ».

Répartition des 48 collèges publics selon la formation prédominante, Québec, 2010

Formation prédominante	Nombre de collèges	Pourcentage de collèges
Préuniversitaire	7	14,6 %
Technique	34	70,8 %
Aucune	7	14,6 %
Total	48	100,0 %

Source: Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

À l'aide de ce tableau, il est maintenant beaucoup plus facile d'analyser les données recueillies. Cette analyse consiste à attirer l'attention du lecteur sur deux ou trois faits saillants (ou marquants) et non à énumérer tous les résultats du tableau de distribution.

Analyse des données

En 2010, on compte 48 collèges publics au Québec. Pour 71 % d'entre eux, la formation technique prédomine (au moins 55 % des étudiants y sont inscrits) alors que les études préuniversitaires priment dans seulement 15 % des cégeps.

NOTES

- Dans un tableau de distribution, la deuxième ou la troisième colonne peut être absente.
- Dans cet ouvrage, nous conviendrons de conserver une seule décimale dans le calcul des pourcentages du tableau de distribution. Nous nous permettrons toutefois d'arrondir les pourcentages à l'entier dans le texte d'analyse des données.
- Pour arrondir les pourcentages, nous appliquerons la règle suivante : si la deuxième décimale est 5 ou plus, on ajoute 1 à la première décimale. Par exemple, 6,38 % sera arrondi à 6,4 %, alors que 10,63 % sera arrondi à 10,6 %.

Règles de présentation d'un tableau de distribution

Voici quelques règles à respecter dans la présentation d'un tableau de distribution.

1. Le tableau doit porter un titre qui évoque son contenu. On doit aussi y préciser le lieu et l'année de réalisation de l'étude si ces informations sont connues. Nous suggérons la formulation suivante pour le titre : « Répartition des (unités statistiques) selon (nom de la variable), (lieu), (année de réalisation) ».
2. La première colonne (ou ligne) du tableau indique les catégories (ou les valeurs) de la variable étudiée et a pour titre le nom de cette variable.
3. La deuxième colonne (ou ligne) donne le nombre ou le pourcentage d'unités statistiques pour chaque catégorie (ou valeur) de la variable. Le titre de cette colonne revêt la forme suivante : « Nombre (d'unités statistiques) » ou « Pourcentage (d'unités statistiques) ».
4. La dernière ligne (ou colonne) du tableau donne le nombre ou le pourcentage total des unités statistiques.
5. Lorsque les données présentées sont tirées d'une recherche, on indique la source et l'année de publication sous le tableau.

En règle générale, lorsqu'une recherche compte un grand nombre de tableaux, on les numérote, puis on en dresse la liste au début ou à la fin du rapport.

Effectif ou fréquence absolue

Un **effectif**, ou **fréquence absolue**, est le nombre de données d'une catégorie (ou d'une valeur).

À titre d'exemple, dans le tableau précédent, l'effectif de la catégorie « Préuniversitaire » est 7.

Fréquence relative

La **fréquence relative** est la proportion de données d'une catégorie (ou d'une valeur). Elle est généralement exprimée en pourcentage. Elle est particulièrement utile lorsqu'on veut comparer deux séries statistiques dont l'effectif total est différent.

À titre d'exemple, la fréquence relative de la catégorie « Préuniversitaire » est 14,6 %. On obtient ce pourcentage ainsi :

$$\frac{7}{48} \times 100 \% = 14,6 \%$$

Représentations graphiques d'une variable qualitative

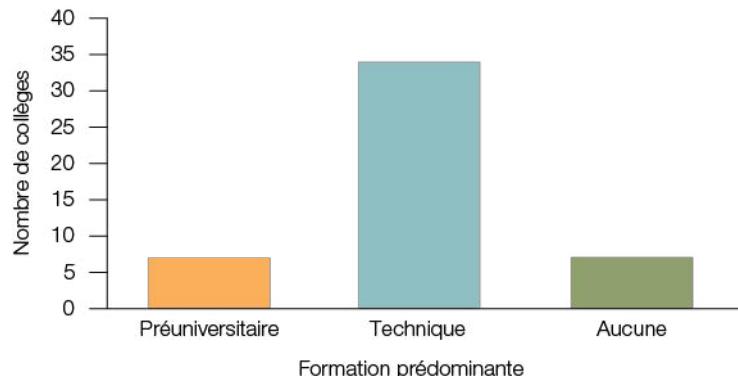
La représentation graphique d'une distribution permet, en un coup d'œil, de se faire une idée de la répartition des données entre les catégories de la variable. Pour une variable qualitative, il existe quatre types de graphiques : le diagramme à rectangles verticaux, le diagramme à rectangles horizontaux, le diagramme circulaire et le diagramme linéaire.

Diagramme à rectangles verticaux

On construit le diagramme à rectangles verticaux de la façon suivante : après avoir inscrit chaque catégorie de la variable sous l'axe horizontal d'un système d'axes, on érige des rectangles, non adjacents, de hauteur égale à l'effectif ou à la fréquence relative (en pourcentage) au-dessus de chaque catégorie. On désigne l'axe horizontal par le nom de la variable et l'axe vertical par le nombre ou le pourcentage des unités statistiques. Par la suite, on donne au graphique un titre décrivant la distribution représentée.

Diagramme à rectangles verticaux

Répartition des 48 collèges publics selon la formation prédominante, Québec, 2010



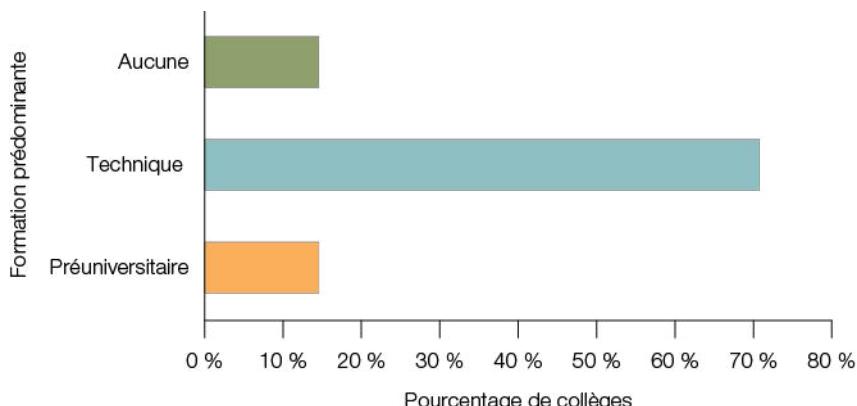
Source: Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

Diagramme à rectangles horizontaux

Contrairement au diagramme à rectangles verticaux, on réserve ici l'axe vertical aux catégories de la variable et l'axe horizontal au nombre ou au pourcentage des unités statistiques. Les rectangles construits se trouvent alors en position horizontale.

Diagramme à rectangles horizontaux

Répartition des 48 collèges publics selon la formation prédominante, Québec, 2010



Source: Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

Caractéristiques des diagrammes à rectangles

Les diagrammes à rectangles verticaux et horizontaux permettent de comparer visuellement les catégories entre elles et l'on peut les construire en utilisant le nombre ou le pourcentage de données de la distribution de la variable.

Diagramme circulaire

Pour construire un diagramme circulaire, on divise un cercle en autant de secteurs circulaires qu'il y a de catégories pour la variable. L'angle de chaque secteur doit être proportionnel à la fréquence relative de

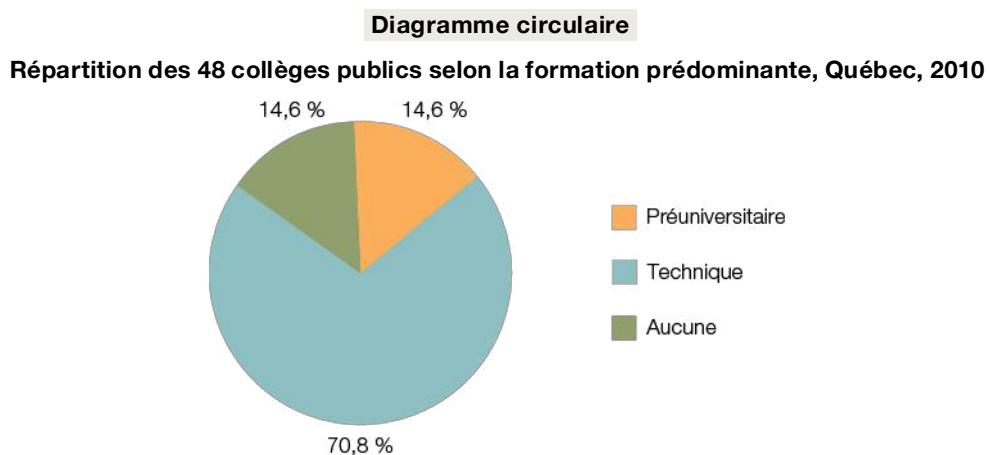
la catégorie qu'il représente. Un titre et une légende désignant chaque secteur doivent accompagner le graphique. Ce diagramme est surtout utilisé pour une variable qualitative nominale.

Il est plus facile de construire ce type de graphique à l'aide d'un ordinateur; si on le trace à la main, on utilise la formule suivante pour déterminer les angles des secteurs circulaires :

$$\text{Angle du secteur} = 360^\circ \times \text{fréquence relative de la catégorie}$$

A titre d'exemple, on calcule ainsi l'angle du secteur de la catégorie préuniversitaire :

$$360^\circ \times 14,6\% \approx 53^\circ$$



Caractéristiques du diagramme circulaire

L'avantage du diagramme circulaire par rapport au diagramme à rectangles est que l'ensemble des données y est représenté par la surface du cercle, ce qui permet de visualiser la part du total des données attribuée à chaque catégorie. Ses inconvénients sont que l'on ne peut pas le construire à partir des effectifs et qu'il ne peut être utilisé si la variable comporte plus de sept catégories, car sa lecture devient alors trop complexe.

Diagramme linéaire

Dans un diagramme linéaire, on utilise la surface d'un rectangle au lieu de celle d'un cercle pour représenter l'ensemble des données d'une variable qualitative. La surface de chaque section du rectangle doit être proportionnelle à la fréquence relative des catégories nommées dans la légende.

Si l'on doit construire un diagramme linéaire à la main, on fixe la longueur et la hauteur que l'on veut donner au rectangle, on trace sous celui-ci un axe gradué de 0 % à 100 %, puis on détermine la longueur de chaque section en fonction de la catégorie représentée à l'aide de la formule ci-dessous. Pour terminer, on ajoute une légende et un titre.

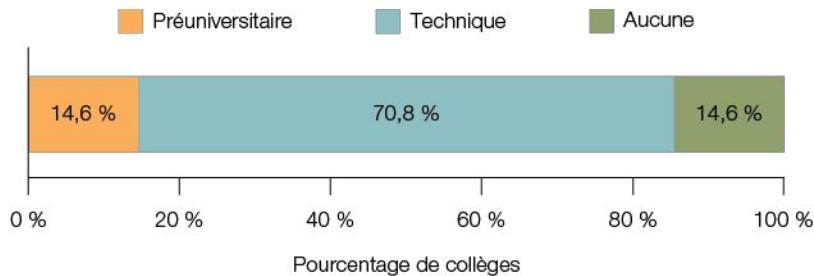
$$\text{Longueur d'une section} = \text{longueur du rectangle} \times \text{fréquence relative de la catégorie}$$

À titre d'exemple, si l'on utilise un rectangle de 10 cm de long sur 1 cm de haut pour représenter la distribution de la variable «formation prédominante», la section attribuée à la catégorie «Préuniversitaire» aura la longueur suivante :

$$\text{Longueur de la section «Préuniversitaire»} = 10 \text{ cm} \times 14,6\% \approx 1,5 \text{ cm}$$

Diagramme linéaire

Répartition des 48 collèges publics selon la formation prédominante, Québec, 2010



Source: Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

NOTE

Un diagramme linéaire peut aussi être en position verticale. Dans ce cas, on indique les pourcentages sur l'axe vertical. Le diagramme linéaire ne peut être utilisé s'il y a plus de sept catégories.

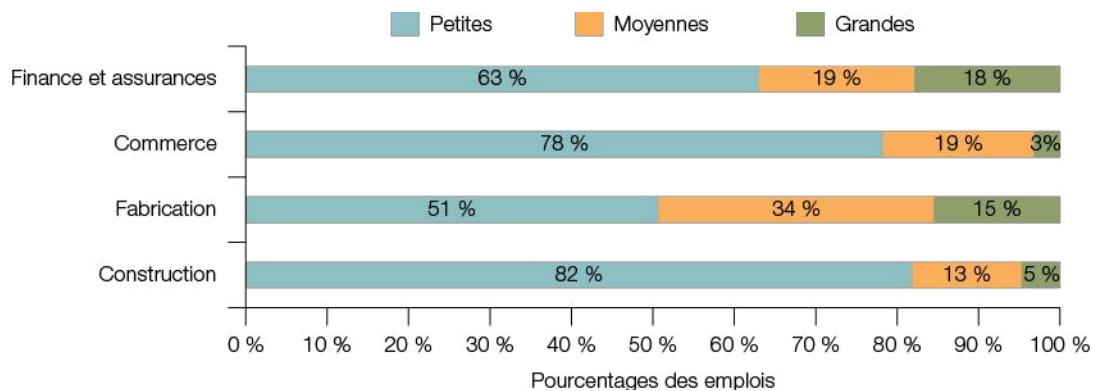
Caractéristiques du diagramme linéaire

Tout comme un diagramme circulaire, le diagramme linéaire permet de visualiser l'importance d'une catégorie par rapport à l'ensemble des données représentées par la surface du rectangle, mais il a l'avantage d'être beaucoup plus facile à tracer à la main. C'est aussi le graphique à privilégier pour comparer plus de deux distributions d'une même variable, comme l'illustre l'exemple ci-dessous.

EXEMPLE

On divise les entreprises en trois groupes : les petites entreprises (moins de 100 employés), les moyennes entreprises (de 101 à 499 employés) et les grandes entreprises (500 employés et plus). On dit que ce sont les petites entreprises qui créent le plus d'emplois. Est-ce vrai ? Pour répondre à cette question, analyser les statistiques suivantes.

Répartition des emplois selon la taille des entreprises pour certains secteurs d'activité, Québec, 2010



Source: Industrie Canada. Principales statistiques relatives aux petites entreprises – Août 2013.

Analyse des données

Les petites entreprises sont à l'origine de plus de 50 % des emplois dans les quatre secteurs étudiés et de plus de 75 % des emplois pour deux des secteurs, soit le commerce (78 %) et la construction (82 %).

Règles de présentation d'un graphique

- Le graphique doit porter un titre qui évoque son contenu. On doit aussi y préciser le lieu et l'année de réalisation de l'étude si ces informations sont connues. Nous suggérons la formulation suivante pour le titre : « Répartition des (unités statistiques) selon (nom de la variable), (lieu), (année de réalisation) ».
- Pour un graphique qui comporte des axes, il faut nommer ceux-ci.
- Pour un diagramme circulaire ou linéaire, une légende doit identifier chaque secteur.
- Lorsque les données sont tirées d'une recherche, on en indique la source et l'année de publication sous le graphique.

En règle générale, lorsqu'une recherche compte un grand nombre de graphiques, on les numérote, puis on en dresse la liste au début ou à la fin du rapport.

1.2.2 La présentation d'une variable quantitative discrète

MISE EN

SITUATION (suite)

Pour les 48 collèges publics du Québec, on a obtenu la série statistique suivante pour la variable quantitative discrète « nombre de programmes en techniques administratives offerts ».

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

Cette variable comporte sept valeurs différentes : 0, 1, 2, 3, 4, 5 et 6.

Tableau de distribution d'une variable quantitative discrète

Pour construire le tableau de distribution d'une variable quantitative discrète, on énumère les valeurs de la variable, puis on fait correspondre à chaque valeur le nombre ou le pourcentage de données de la série statistique ayant cette valeur.

Répartition des 48 collèges publics selon le nombre de programmes en techniques administratives offerts, Québec, 2010

Dénombrement (voir la note à la page suivante)	Nombre de programmes en techniques administratives	Nombre de collèges	Pourcentage de collèges
I	0	1	2,1 %
I	1	1	2,1 %
	2	11	22,9 %
	3	9	18,8 %
	4	16	33,3 %
	5	9	18,8 %
I	6	1	2,1 %
Total		48	100,1 %¹

1. Le total est différent de 100 % en raison des arrondis.

Source: Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

Analyse des données

En 2010, 42 % des collèges publics offrent 2 ou 3 des 6 programmes en techniques administratives et 52 % en offrent 4 ou 5. On remarque qu'un collège n'en offre aucun (cégep de Saint-Laurent) et qu'un seul collège offre les 6 programmes (cégep de Lanaudière).

NOTE

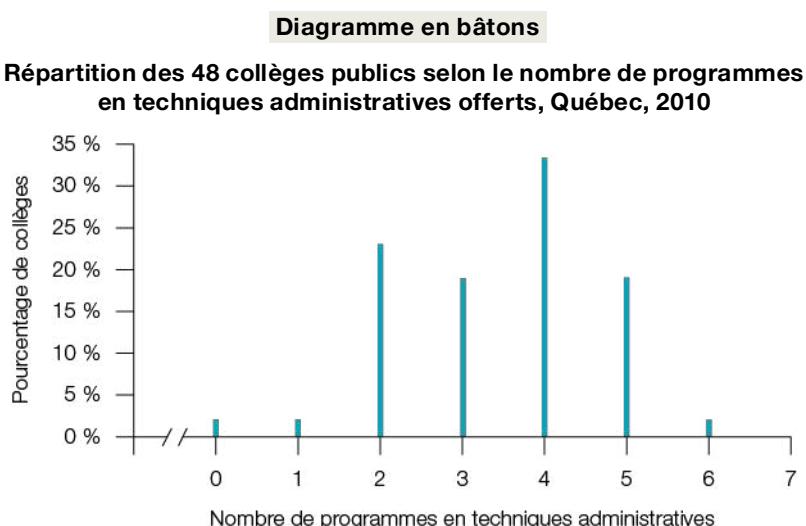
La partie intitulée « Dénombrement » indique une technique rapide pour compiler manuellement les effectifs de chaque valeur. Au lieu de parcourir sept fois la liste des données pour compter le nombre de 0, de 1, de 2, etc., on le fait une seule fois. On lit la valeur de la première donnée de la série, puis on trace un trait à gauche de cette valeur dans le tableau. On fait de même pour les autres données, mais, lorsqu'on lit une valeur pour la cinquième fois, on indique ce fait en traçant un trait horizontal sur les quatre traits déjà inscrits à gauche de la valeur.

Représentation graphique d'une variable quantitative discrète

On représente la distribution d'une variable quantitative discrète au moyen d'un diagramme en bâtons.

Diagramme en bâtons

Pour construire un diagramme en bâtons, on porte sur l'axe horizontal les différentes valeurs de la variable selon une échelle choisie arbitrairement, puis on élève sur chaque valeur un bâton de longueur proportionnelle à l'effectif ou à la fréquence relative (en pourcentage). On donne ensuite un titre au graphique et on nomme les axes.



Source: Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

Le choix d'ériger un bâton, et non un rectangle, au-dessus de chaque valeur permet de visualiser le fait qu'une variable discrète n'est pas une approximation, mais bien un nombre précis : il n'y a rien avant ou après la valeur considérée. Par exemple, si l'on élevait au-dessus de la valeur 4 un rectangle de hauteur 33,3 % dont la base irait de 3,9 à 4,1, on aurait visuellement l'impression que le nombre de programmes en techniques administratives peut se situer entre 3,9 et 4,1, ce qui est faux : le nombre de programmes en techniques administratives n'est pas une variable quantitative continue.

EXERCICE DE COMPRÉHENSION | 1.2

Une étude indique que 200 accidents sont survenus sur une autoroute durant l'année : 160 n'ont fait aucune victime, 32 ont fait une victime, 6 ont fait deux victimes et 2 ont fait trois victimes. Parmi les 50 victimes, 30 ont eu des blessures légères, 16 ont eu des blessures graves et 4 sont décédées.

a) Les tableaux suivants présentent les résultats de l'étude. Donner un titre aux tableaux.

Tableau 1

Titre: _____

Nombre de victimes	Nombre d'accidents	Pourcentage d'accidents
0	160	80 %
1	32	16 %
2	6	3 %
3	2	1 %
Total	200	100 %

Tableau 2

Titre: _____

Gravité des blessures	Nombre de victimes	Pourcentage de victimes
Légères	30	60 %
Graves	16	32 %
Mortelles	4	8 %
Total	50	100 %

b) Vrai ou faux ? Si faux, donner la bonne réponse.

i) 16 % des accidents ont fait au moins une victime.

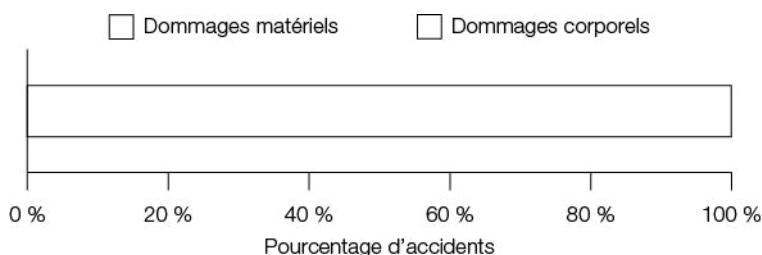
ii) 40 % des accidents ont causé des blessures graves ou mortelles.

➤ iii) On peut représenter la distribution du tableau 1 par un diagramme à rectangles horizontaux.

iv) On peut représenter la distribution du tableau 2 par un diagramme à rectangles verticaux.

c) Un accident qui ne fait aucune victime est dit «accident avec dommages matériels», sinon c'est un accident avec dommages corporels. Utiliser ces définitions et les statistiques de la page précédente pour compléter le diagramme linéaire.

Titre: _____



1.2.3 La présentation d'une variable quantitative continue

La mise en situation suivante sera utilisée pour présenter la façon de traiter les données d'une variable quantitative continue.

NOTE

On traite une variable quantitative discrète qui comporte un grand nombre de valeurs différentes, comme le nombre d'étudiants inscrits dans chaque cégep du Québec, de la même façon qu'une variable quantitative continue.

MISE EN SITUATION

Un café offre sur demande un code d'accès à un réseau Wi-Fi à ses clients. Toutefois, afin de s'assurer qu'il y ait toujours des tables disponibles pour la clientèle, la durée du branchement au réseau est limitée à 60 minutes. Pour vérifier si, malgré cette contrainte, le temps alloué répond aux besoins des clients, on note la durée (en minutes) du branchement au réseau d'un échantillon de 40 clients. Voici la série statistique obtenue:

48	35	29	44	42	52	43	38	40	47
30	56	32	49	40	59	37	39	40	46
53	37	48	45	46	42	43	35	33	51
26	45	41	41	34	38	43	41	38	35

Tableau de distribution d'une variable quantitative continue

La durée du branchement est une variable quantitative continue. Pour construire un tableau de distribution pour une variable quantitative continue, on groupe d'abord les données en **classes**, puis on associe à chaque classe le nombre ou le pourcentage de données de la série qui en font partie.

On pourrait, par exemple, grouper les temps de branchement des 40 clients de la mise en situation en 7 classes ; on obtiendrait alors le tableau de distribution suivant.

Répartition des 40 clients de l'échantillon selon la durée du branchement au réseau Wi-Fi

Durée du branchement (en minutes)	Nombre de clients	Pourcentage de clients
$25 \leq X < 30$	2	5,0 %
$30 \leq X < 35$	4	10,0 %
$35 \leq X < 40$	9	22,5 %
$40 \leq X < 45$	12	30,0 %
$45 \leq X < 50$	8	20,0 %
$50 \leq X < 55$	3	7,5 %
$55 \leq X < 60$	2	5,0 %
Total	40	100,0 %

Analyse des données

Les 60 minutes de branchement allouées semblent convenir ; en effet, pour 73 % des clients de l'échantillon, la durée du branchement se situe entre 35 et 50 minutes, et seulement 5 % des clients sont encore branchés 5 minutes avant le délai limite.

Notation des classes

Pour décrire la 1^{re} classe, on utilise la notation « $25 \leq X < 30$ », qui signifie que la durée du branchement, que l'on symbolise par la lettre X , des clients dans cette classe va de 25 minutes à moins de 30 minutes. On peut aussi utiliser la notation « $[25 ; 30[$ » pour décrire cette classe. On dit que 25 est la **limite inférieure** de la classe, et 30 sa **limite supérieure**.

NOTE

Lorsque la variable étudiée est l'âge, on rencontre aussi la notation 25-29, 30-34, etc.

Amplitude d'une classe

L'amplitude d'une classe est égale à la différence entre sa limite supérieure et sa limite inférieure. Dans le tableau ci-dessus, chaque classe a une amplitude de cinq minutes.

Démarche pour construire des classes de même amplitude

Comment détermine-t-on l'amplitude et le nombre de classes nécessaire au regroupement des données d'une série statistique ?

Nous utiliserons les 40 données de la série statistique de la mise en situation pour illustrer la procédure qui a permis de créer les classes du tableau de distribution.

1. Fixer temporairement le nombre de classes

Le nombre de classes nécessaire pour grouper les données d'une série statistique dépend du nombre de données. Nous utiliserons la table de Sturges pour fixer temporairement le nombre de classes de la distribution de la durée du branchement.

Table de Sturges

Nombre de données	Nombre approximatif de classes ¹
Entre 10 et 22	5
Entre 23 et 44	6
Entre 45 et 90	7
Entre 91 et 180	8
Entre 181 et 360	9
Entre 361 et 720	10

1. Ce nombre est déterminé par la formule $1 + 3,322 \log n$, où n est le nombre de données.

Repérons, dans la première colonne du tableau, l'intervalle dans lequel se situe le nombre de données de la série statistiques.

❓ Avec 40 données, quel est le nombre approximatif de classes suggéré par la table de Sturges? _____

Nous retiendrons donc, temporairement, ce nombre de classes pour la suite de la démarche.

2. Calculer l'étendue de la série

L'amplitude des classes dépend de l'étendue des données. L'étendue, que l'on note E , est égale à la différence entre la plus grande et la plus petite valeur de la série statistique.

$$E = x_{\max} - x_{\min}$$

❓ Calculer l'étendue de la série statistique étudiée: _____

3. Déterminer l'amplitude

Amplitude calculée

Nous voulons savoir quelle sera l'amplitude de chaque classe si nous prenons une étendue de 33 minutes et que nous la divisons en 6 classes. Il apparaît logique de penser que le calcul suivant donne la réponse à cette question.

❓ Amplitude calculée = $\frac{\text{étendue}}{\text{nombre de classes}} = \text{_____ minutes}$

Amplitude choisie

À cette étape de la démarche, il faut se servir de son jugement pour choisir une amplitude qui facilitera la lecture du tableau de distribution. Les éléments suivants peuvent vous guider dans votre choix :

- Il est préférable de choisir comme amplitude un multiple de 5 ou un nombre pair. Pour atteindre cet objectif, on peut s'éloigner de l'amplitude calculée en faisant toutefois preuve de mesure (choisir une amplitude qui serait le double de l'amplitude calculée serait grandement exagéré).
- Nous conviendrons de choisir un nombre entier comme amplitude si les données à grouper sont des entiers. Si les données sont précises au dixième près ou au centième près, nous utiliserons la même précision pour l'amplitude choisie.

❓ Quelle amplitude choisissez-vous? _____ minutes

4. Choisir la limite inférieure de la première classe

La limite inférieure de la 1^{re} classe est déterminante dans la construction des classes. Nous choisirons donc un nombre qui, en considérant l'amplitude choisie, permettra de produire un tableau de distribution agréable à lire. Il va de soi que la limite inférieure de la 1^{re} classe devra être plus petite ou égale à la plus petite donnée de la série statistique.

❓ Que devrions-nous prendre pour limite inférieure de la 1^{re} classe? _____

Quelle est la 1^{re} classe de la distribution? _____

Par la suite, on construit les autres classes en s'assurant bien que la plus grande donnée de la série statistique soit comprise dans la dernière classe de la distribution.

NOTE

Les choix que nous avons faits pour l'amplitude et la limite inférieure de la 1^{re} classe nous donneront finalement 7 classes au lieu des 6 que nous avions prévues au début; cela n'a pas d'importance, l'essentiel étant que le tableau se lise bien.

Démarche pour construire des classes de même amplitude

1. Fixer temporairement le nombre de classes avec la table de Sturges.
2. Calculer l'étendue de la série: $E = x_{\max} - x_{\min}$
3. Déterminer l'amplitude:
 - Amplitude calculée = $\frac{\text{étendue}}{\text{nombre de classes}}$
 - Amplitude choisie
4. Choisir la limite inférieure de la 1^{re} classe, puis déterminer les classes en utilisant l'amplitude choisie.

NOTE

Il peut arriver dans certains milieux que l'on ne suive pas la démarche proposée pour créer des classes parce qu'une convention, propre au type de données à traiter, a déjà été établie pour répondre aux besoins. Par exemple, dans le milieu scolaire, il est habituel de grouper les notes d'un examen par tranches de 10 points: [40 ; 50[, [50 ; 60[, [60 ; 70[, etc.

EXEMPLE

Donner la première des classes qui permettrait de grouper une série statistique de 36 données, précises au centième près, sachant que la plus petite donnée est 2,65 \$, et la plus grande 18,45 \$.

Solution

EXERCICE DE COMPRÉHENSION | 1.3

On désire grouper en classes les revenus hebdomadaires d'un groupe de 80 personnes. Ces revenus sont exprimés en nombres entiers ; le plus petit revenu est 252 \$ et le plus grand, 937 \$. Donner la 1^{re} classe de la distribution du revenu.

Solution

Représentation graphique d'une variable quantitative continue

On représente la distribution d'une variable continue par un histogramme ou un polygone de fréquences.

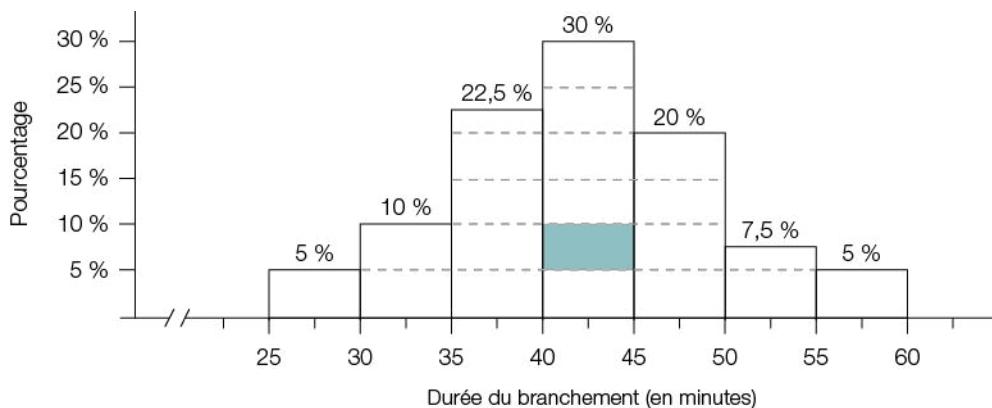
Histogramme (classes de même amplitude)

L'histogramme est formé de rectangles adjacents ; ceux-ci sont accolés afin d'indiquer que l'étude porte sur une variable continue. La base de ces rectangles correspond aux classes de la distribution, alors que la surface est proportionnelle aux fréquences (relatives ou absolues) des classes.

SITUATION (suite)

Voici l'histogramme représentant la distribution du temps de branchement des 40 clients du café (voir la page 22) :

Répartition des 40 clients de l'échantillon selon la durée du branchement au réseau Wi-Fi



Assurons-nous de bien comprendre le principe de proportionnalité, car la technique utilisée pour déterminer les quantiles à la section 1.4 et les calculs de probabilités d'une distribution normale à la section 3.4 est basée sur cette notion de proportionnalité.

Concrètement, dire qu'il y a proportionnalité entre la surface des rectangles et les pourcentages de données des classes signifie que les rapports observés entre les pourcentages s'appliquent aussi aux surfaces des rectangles. Pour illustrer ceci, considérons les deux observations suivantes :

- Le plus grand pourcentage de données est dans la 4^e classe.
Le rectangle qui a la plus grande surface est dans la 4^e classe.
- Le pourcentage de données dans la 4^e classe égale six fois celui de la 1^{re} classe :

$$30 \% = 6 \times 5 \%$$

De même, la surface du 4^e rectangle (S_4) égale 6 fois celle du 1^{er} rectangle (S_1). En effet, si l'on prend le rectangle bleu comme unité de mesure, on a :

$$S_4 = 6 \times S_1$$

Voyons maintenant comment, grâce à la notion de proportionnalité, on peut utiliser un rapport de surfaces pour déterminer le pourcentage de données d'une classe.

Puisque 5 % de toutes les données sont dans la 1^{re} classe, en vertu du principe de proportionnalité, on peut affirmer que la surface du 1^{er} rectangle occupe 5 % de la surface totale de l'histogramme. Vérifions cette affirmation :

Aire du 1^{er} rectangle = 1 × aire du rectangle bleu
Aire de l'histogramme = 20 × aire du rectangle bleu

$$\frac{\text{Aire du 1^{er} rectangle}}{\text{Aire totale de l'histogramme}} \times 100 \% = \frac{1}{20} \times 100 \% = 5 \%$$

À partir de l'histogramme, le pourcentage de données dans une classe peut donc se calculer ainsi :

Pourcentage de données dans une classe

$$\frac{\text{Aire du rectangle de la classe}}{\text{Aire totale de l'histogramme}} \times 100 \%$$

EXEMPLE

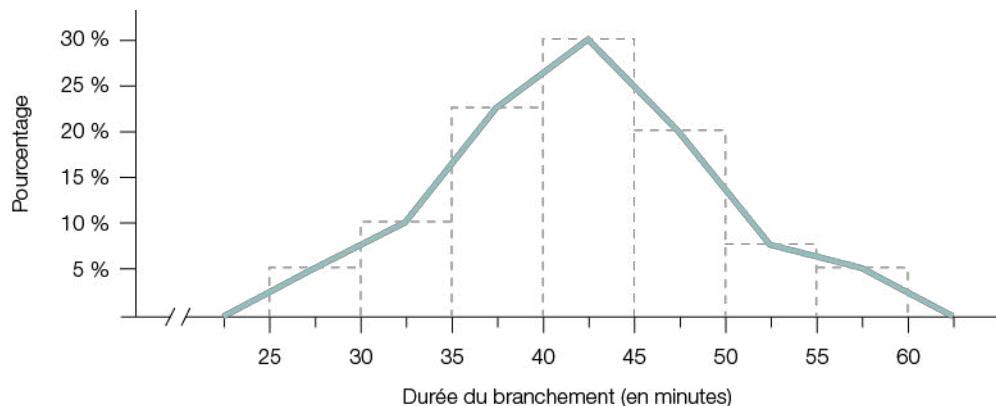
Imaginons que, dans la mise en situation qui précède, on donne l'histogramme de la distribution de la durée du branchement, mais sans aucun pourcentage au-dessus des rectangles et sans axe vertical. Dans ces conditions, déterminer, à l'aide d'un rapport de surfaces et en prenant le rectangle bleu comme unité de mesure, le pourcentage de clients de l'échantillon dont la durée du branchement au réseau a été de 35 minutes à moins de 40 minutes.

Solution

Polygone de fréquences

Pour construire un polygone de fréquences, il suffit de joindre par un segment de droite le point milieu du sommet de chacun des rectangles et de fermer la base de la figure ainsi construite en ajoutant, au début et à la fin de l'histogramme, une classe de fréquence nulle. Il est important de fermer le polygone afin de s'assurer que son aire soit égale à celle de l'histogramme.

Répartition des 40 clients de l'échantillon selon la durée du branchement au réseau Wi-Fi



On recourt au polygone de fréquences, de préférence à l'histogramme, lorsqu'on désire comparer deux distributions. On s'en sert aussi pour trouver le modèle mathématique qui pourrait s'appliquer à la distribution ; par exemple, si le polygone de fréquences a la forme d'une cloche, comme à la page précédente, on dira que l'on a une distribution normale.

EXEMPLE

Utiliser les données du tableau ci-dessous pour comparer la distribution de l'âge des Québécois en 2011 avec celle qui a été obtenue en 1665 lors du premier recensement québécois.

Répartition de la population du Québec en 1665 et en 2011 selon l'âge

Âge (en ans)	1665		2011	
	Nombre de Québécois	Pourcentage de Québécois	Nombre de Québécois	Pourcentage de Québécois
]0; 10[1 014	32,3 %	834 593	10,5 %
[10; 20[416	13,3 %	895 389	11,2 %
[20; 30[802	25,6 %	1 038 904	13,0 %
[30; 40[501	16,0 %	1 079 285	13,5 %
[40; 50[233	7,4 %	1 162 840	14,6 %
[50; 60[108	3,4 %	1 214 445	15,2 %
[60; 70[45	1,4 %	903 755	11,3 %
[70; 80[13	0,4 %	517 899	6,5 %
[80; 90[4	0,1 %	277 859	3,5 %
90 et plus	0	0,0 %	54 694	0,7 %
Total	3 136	99,9 %¹	7 979 663	100,0 %

1. Le total est différent de 100 % en raison des arrondis.

Sources: Statistique Canada. Recensement 1665 et Recensement 2011.

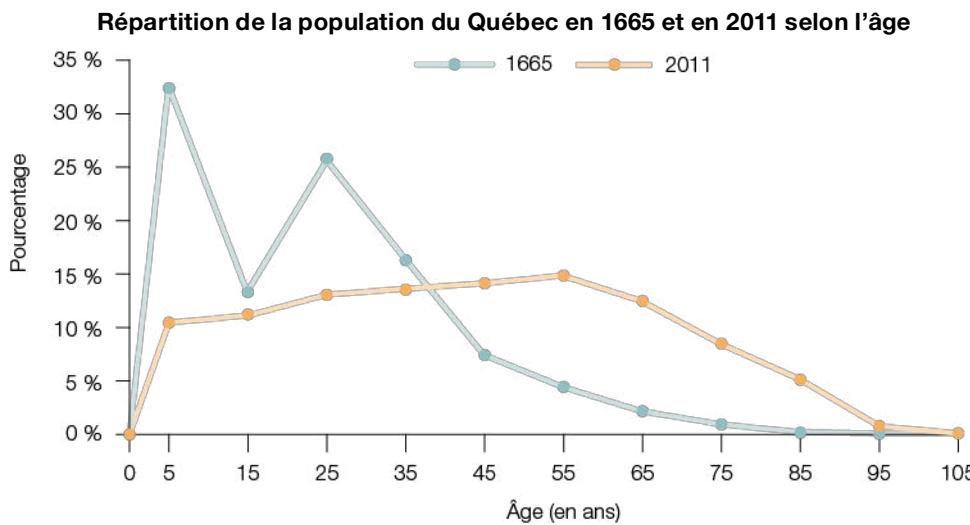
Une représentation graphique mettrait beaucoup plus en évidence la différence entre ces deux distributions. Le polygone de fréquences est le graphique approprié : en effet, la construction de deux histogrammes superposés produirait un graphique tout à fait illisible. Pour comparer les distributions, nous utiliserons les pourcentages et non les effectifs de chaque recensement puisque l'effectif total des distributions est différent. Il faut noter qu'il n'est pas nécessaire de passer par l'histogramme pour construire un polygone de fréquences.

Classe ouverte

Lorsque la limite inférieure ou supérieure d'une classe n'est pas clairement définie, comme pour la classe «90 et plus», on dit que la distribution a une **classe ouverte**. Pour construire l'histogramme ou le polygone de fréquences, on ferme généralement cette classe en lui donnant la même amplitude que celle des autres classes, ce qui donnera la classe [90 ; 100[.

NOTE

Normalement, les polygones devraient être fermés à gauche au centre de l'intervalle [-10 ; 0[, mais comme l'âge ne peut être négatif, nous fermerons les polygones au point (0 ; 0).



Sources: Statistique Canada. Recensement 1665 et Recensement 2011.

Analyse des données

La population du Québec est passée de quelque 3 000 habitants en 1665 à près de 8 millions en 2011.

En 1665, la population du Québec était très jeune (on observe que la surface sous le polygone est beaucoup plus grande avant 35 ans qu'après cet âge). Il est intéressant de souligner que, à cette époque, seulement 5 % des habitants avaient 50 ans et plus, alors qu'aujourd'hui, le pourcentage est environ 7 fois plus grand, soit 37 %.

En 1665, le groupe des moins de 10 ans avait un poids démographique très important, avec 32 % de la population (le polygone atteint un sommet dans cet intervalle). La tranche des 20 à 30 ans arrive au deuxième rang avec le quart de la population.

En 2011, les personnes âgées de 40 à 50 ans et celles âgées de 50 à 60 ans sont les 2 groupes les plus populaires avec 15 % de la population chacun.

Histogramme (classes d'amplitude inégale)

Il peut arriver que l'on doive grouper des données dans des classes d'amplitudes inégales lorsque, avec des classes d'amplitude égale, on a peu ou pas de données dans une ou plusieurs classes de la distribution.

Nous ne vous demanderons pas de construire de telles classes à partir d'une série de données, mais par contre vous devrez être en mesure de représenter graphiquement une distribution composée de classes inégales. Comment faire pour construire un histogramme ou un polygone de fréquences constitué de classes d'amplitude inégale ? La mise en situation suivante montre la façon de procéder dans un tel cas.

MISE EN SITUATION

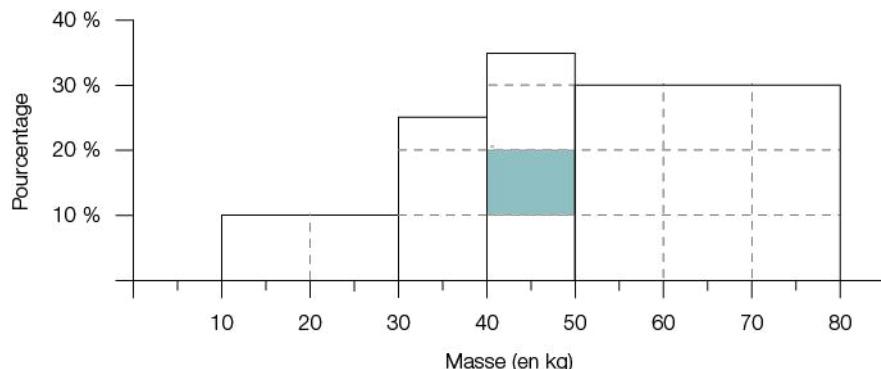
Voici le tableau de distribution de la masse, en kilogrammes, des différentes roches que l'on trouve dans une rocaille de fleurs. On désire construire l'histogramme de cette distribution.

Répartition des roches de la rocaille selon la masse

Masse (en kg)	Pourcentage de roches
$10 \leq X < 30$	10 %
$30 \leq X < 40$	25 %
$40 \leq X < 50$	35 %
$50 \leq X < 80$	30 %
Total	100 %

Dans ce tableau de distribution, les classes ont les amplitudes suivantes : 20, 10, 10 et 30.

Voici l'histogramme que l'on obtiendrait en appliquant la même technique que celle qui a été utilisée pour construire un histogramme composé de classes d'amplitude égale.



Ce graphique respecte-t-il le principe de proportionnalité entre surface et pourcentage ?

- On trouve le plus grand pourcentage dans la 3^e classe du tableau de distribution et pourtant le rectangle de l'histogramme qui présente la plus grande surface se trouve dans la 4^e classe du tableau !
- La surface du 1^{er} rectangle occupe-t-elle 10 % de la surface totale de l'histogramme ?

$$\frac{\text{Aire du 1er rectangle}}{\text{Aire totale de l'histogramme}} \times 100 \% = \frac{2}{17} \times 100 \% = 11,8 \%$$

La preuve est faite. La technique utilisée pour construire des histogrammes composés de classes égales ne permet pas de respecter le principe de proportionnalité lorsque les classes sont inégales. Il faut donc choisir une autre technique de construction.

Nous utiliserons la technique suivante pour construire un histogramme comportant des classes inégales :

- On choisit une amplitude de base, généralement celle de la majorité des classes de la distribution.
- Une classe dont l'amplitude est égale à deux fois l'amplitude de base sera considérée comme le regroupement de deux classes standard et, par conséquent, la hauteur du rectangle de cette classe devra être égale à la moitié de sa fréquence. On applique ce même principe si une classe correspond à plus de deux fois l'amplitude standard. On dit alors que l'on fait une **rectification de fréquences**.

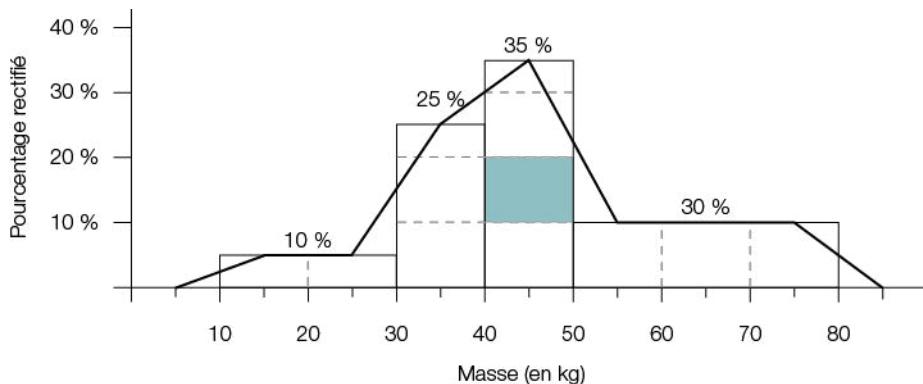
Effectuer la rectification de fréquences de la distribution suivante :

Répartition des roches de la rocallie selon la masse

Amplitude	Masse (en kg)	Pourcentage de roches	Pourcentage rectifié
20	$10 \leq X < 30$	10 %	
10	$30 \leq X < 40$	25 %	
10	$40 \leq X < 50$	35 %	
30	$50 \leq X < 80$	30 %	
	Total	100 %	

L'histogramme et le polygone de fréquences représentant la distribution obtenue sont présentés à la page suivante.

Répartition des roches de la rocallie selon la masse



NOTES

- On assigne l'étiquette **pourcentage rectifié** à l'axe vertical.
- On indique le pourcentage (ou l'effectif) de données de chaque classe au-dessus des rectangles.
- On ferme le polygone de fréquences en ajoutant, au début et à la fin, une classe d'amplitude égale à l'amplitude de base, et de fréquence nulle, afin de s'assurer que l'aire du polygone soit égale à celle de l'histogramme.



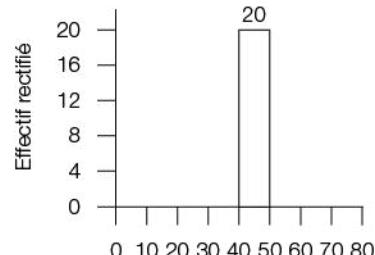
Vérifier qu'avec cet histogramme la surface du 1^{er} rectangle correspond bien à 10 % de la surface totale de l'histogramme, respectant ainsi le principe de proportionnalité.

$$\frac{\text{Aire du 1er rectangle}}{\text{Aire totale de l'histogramme}} \times 100 \% = \underline{\hspace{2cm}} \times 100 \% = \underline{\hspace{2cm}}$$

EXERCICES DE COMPRÉHENSION | 1.4

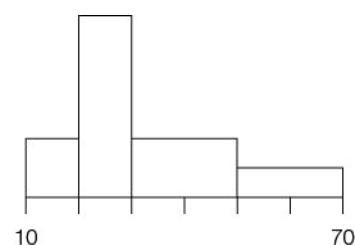
1. Compléter le tableau de distribution et l'histogramme suivants.

Amplitude	Classe	Effectif	
	[10; 40[12	
	[40; 50[20	
	[50; 60[18	
	[60; 80[10	
	Total	60	

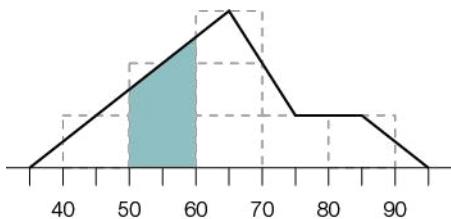


2. Compléter le tableau de distribution suivant en utilisant l'information donnée par l'histogramme.

Classe	Pourcentage
[10 ; 10[
[10 ; 20[
[20 ; 30[
[30 ; 70[
Total	100,0 %



3. Le polygone de fréquences suivant représente une distribution. À quel pourcentage de l'aire totale du polygone l'aire de la portion en bleu est-elle approximativement égale ?



1.2.4 L'ogive ou la courbe de fréquences cumulées

Distribution de fréquences cumulées

Pour connaître le pourcentage de données d'une distribution inférieures à une certaine valeur, on construit un tableau de distribution de fréquences cumulées. On construit ce tableau en ajoutant, à droite du tableau de distribution, une colonne qui indique les pourcentages cumulés des données.

MISE EN SITUATION

En 2011, 72 % des infractions entraînant l'inscription de points d'inaptitude concernent l'excès de vitesse, ce qui représente 644 578 infractions. Quel âge ont les conducteurs qui ont commis ces excès de vitesse ? Voici des statistiques à ce sujet :

Répartition des infractions pour excès de vitesse selon l'âge du conducteur, Québec, 2011

Âge du conducteur	Pourcentage d'infractions	Pourcentage cumulé
Moins de 25 ans	16,6 %	16,6 %
[25 ans ; 35 ans[21,3 %	37,9 %
[35 ans ; 45 ans[21,0 %	58,9 %
[45 ans ; 55 ans[21,0 %	79,9 %
[55 ans ; 65 ans[13,1 %	93,0 %
65 ans et plus	7,0 %	100,0 %
Total	100,0 %	

Source: Société de l'assurance automobile du Québec. *Les infractions et les sanctions reliées à la conduite d'un véhicule routier, 2002-2011, 2012.*

Analyse des données

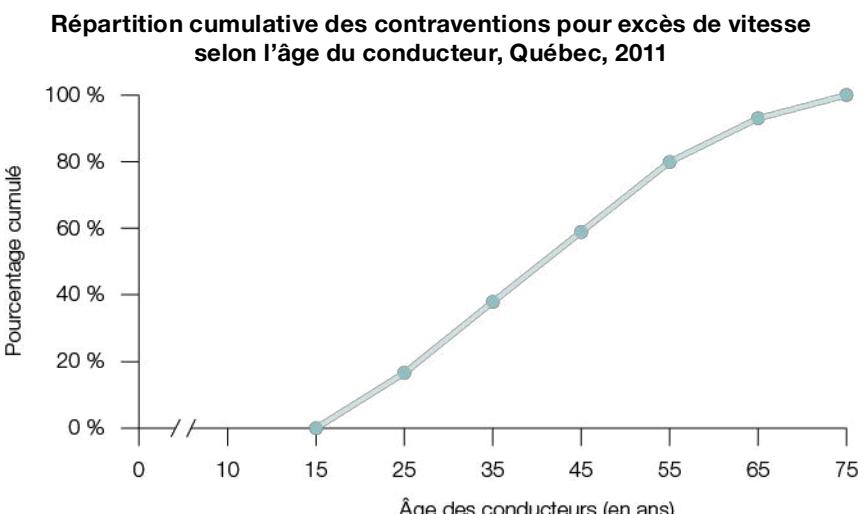
En 2011, près de 60 % des infractions pour excès de vitesse sont commises par des conducteurs de moins de 45 ans. Si l'on ajoute celles commises par les conducteurs de 45 à 55 ans, on atteint 80 % des infractions. Seulement 7 % des infractions pour excès de vitesse sont commises par des conducteurs de 65 ans et plus.

Ogive ou courbe de fréquences cumulées

Pour représenter la distribution de fréquences cumulées d'une variable quantitative continue, on trace une ogive, ou courbe de fréquences cumulées. On construit cette courbe de la façon suivante :

- Pour chaque classe, on détermine un point (x, y) dans le plan cartésien où x est la limite supérieure de la classe et y , la fréquence cumulée de cette classe.
- En partant du point correspondant à la limite inférieure de la 1^{re} classe sur l'axe horizontal, on relie ensuite tous ces points par des segments de droite.
- Enfin, on nomme les axes et on donne un titre au graphique.

En appliquant cette technique à la mise en situation, on obtient l'ogive en joignant les points $(15 ; 0 \%)$, $(25 ; 16,6 \%)$, $(35 ; 37,9 \%)$, $(45 ; 58,9 \%)$, $(55 ; 79,9 \%)$, $(65 ; 93,0 \%)$ et $(75 ; 100,0 \%)$, ce qui donne la courbe suivante.



Source: Société de l'assurance automobile du Québec. *Les infractions et les sanctions reliées à la conduite d'un véhicule routier, 2002-2011*, 2012.

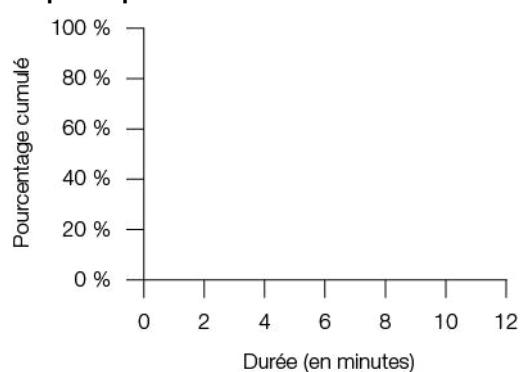
EXERCICE DE COMPRÉHENSION | 1.5

On a étudié la durée d'un échantillon d'appels téléphoniques. Compléter le tableau de distribution de fréquences cumulées et construire l'ogive correspondante.

Répartition des appels téléphoniques de l'échantillon selon la durée

Durée (en min)	Pourcentage des appels	Pourcentage cumulé
[2; 4[20 %	
[4; 6[10 %	
[6; 8[40 %	
[8; 10[30 %	
Total	100 %	

Répartition cumulative des appels téléphoniques de l'échantillon selon la durée

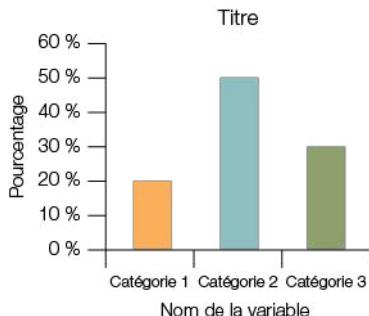


1.2.5 Quel graphique faut-il construire ?

Le graphique à construire pour représenter une distribution dépend du type de variable. Nous énumérons ci-dessous les différents choix possibles pour chaque type de variable.

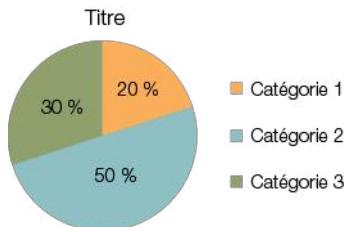
- Pour représenter la distribution d'une variable qualitative**

Diagramme à rectangles
(verticaux ou horizontaux)



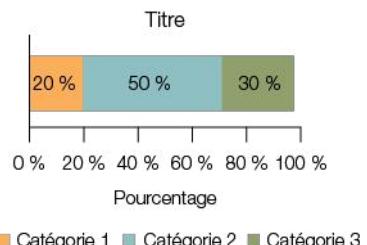
Favorise la comparaison des catégories entre elles.

Diagramme circulaire



Favorise la comparaison de chaque catégorie par rapport à l'ensemble des données.

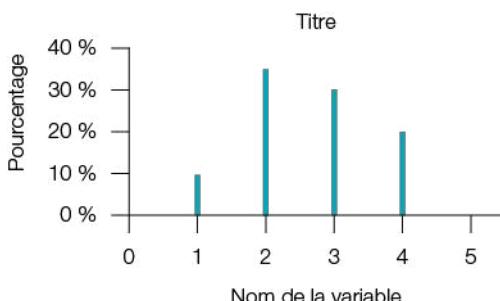
Diagramme linéaire



Favorise la comparaison de chaque catégorie par rapport à l'ensemble des données. Facilite la comparaison de plusieurs distributions.

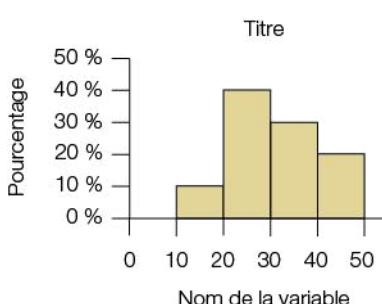
- Pour représenter la distribution d'une variable quantitative discrète**

Diagramme en bâtons



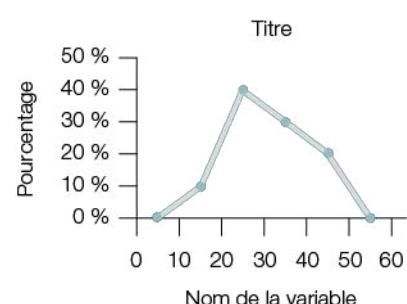
- Pour représenter la distribution d'une variable quantitative continue**

Histogramme



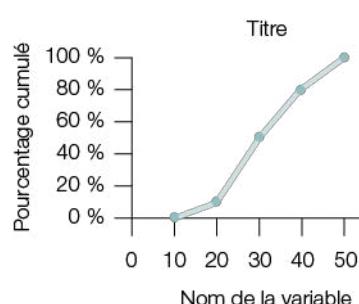
Attention, si les classes n'ont pas la même amplitude, il faut effectuer une rectification de fréquences.

Polygone de fréquences



Facilite la comparaison de plusieurs distributions ayant les mêmes classes.

Ogive ou courbe de fréquences cumulées

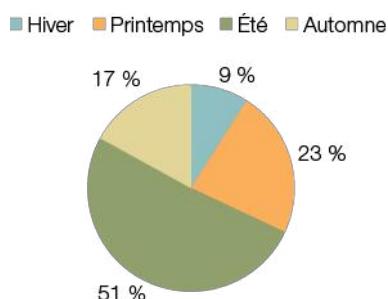


Pour représenter une distribution de fréquences cumulées.

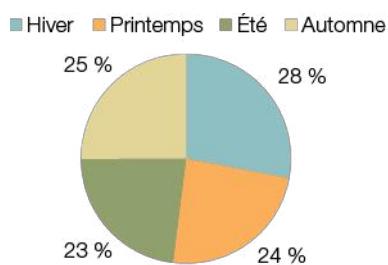
EXERCICES 1.2

1. Dans le *Bilan démographique du Québec, édition 2012*, on trouve les statistiques suivantes.

Graphique 1
Répartition des mariages selon la saison, Québec, 2011



Graphique 2
Répartition des décès selon la saison, Québec, 2011



- a) Compte tenu de ces statistiques, laquelle des interprétations suivantes est vraie ?
- En 2011, un peu plus de la moitié des Québécois se marient en été et le quart décèdent à l'automne.
 - En 2011, un peu plus de la moitié des mariages ont lieu en été et un Québécois sur quatre meurt à l'automne.
 - En 2011, un peu plus de la moitié des mariages se font en été et le quart des décès ont lieu à l'automne.
 - En 2011, un peu plus d'un Québécois sur deux se marie en été et un décès sur quatre se produit à l'automne.
- b) En se basant sur les données de 2011, peut-on dire que les mariages et les décès sont influencés par les saisons ? Écrire un court texte pour répondre à cette question en appuyant l'argumentation de quelques données statistiques.

2. Les familles recomposées² sont-elles nombreuses au Québec ?

Les données de 2011 indiquent que 146 144 des 1 273 240 familles avec enfants à la maison sont recomposées, que 761 581 sont intactes³ et que 365 515 sont monoparentales.

Source: Statistique Canada. *Recensement 2011*.

- Présenter ces statistiques dans un tableau de distribution.
- Construire un diagramme linéaire.
- Vrai ou faux ?
 - En 2011, 11,5 % des familles du Québec sont recomposées.
 - En 1996, 67 % des familles avec enfants à la maison étaient intactes. Les données de 2011 indiquent une diminution de cette proportion de 7,2 points de pourcentage en 15 ans.

3. Comment se porte l'industrie du cinéma québécois ?

En 2012, on a produit 116 longs métrages québécois : 30 pour le marché du cinéma, 37 pour celui de la télévision et 49 pour d'autres marchés (DVD, Internet). Des 116 films produits, 45 étaient des fictions et 71 étaient des documentaires.

Source: Institut de la statistique du Québec. *Statistiques sur l'industrie du film et de la production télévisuelle, édition 2013*, tome 2.

- Présenter les statistiques sur les marchés ciblés par les longs métrages québécois dans un tableau, puis analyser les données.
 - Construire, avec les effectifs, le graphique approprié pour présenter les statistiques sur le genre de longs métrages produits.
4. Répondre aux questions en se reportant à la distribution suivante.

Répartition des familles selon leur nombre de téléviseurs

Nombre de téléviseurs	0	1	2	3	4
Nombre de familles	1	13	11	3	2

-
- Famille dont au moins un enfant d'une union antérieure de l'un des conjoints vit sous le même toit que les conjoints.
 - Famille dont tous les enfants du ménage sont les enfants biologiques ou adoptifs des deux membres du couple.

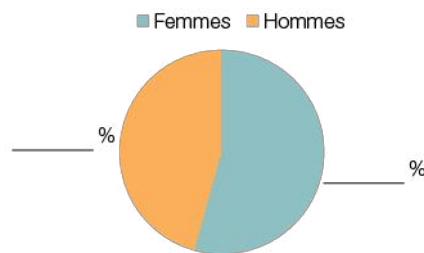
- a) Nommer la variable étudiée et donner ses valeurs.
 - b) Donner le type de la variable.
 - c) Compléter le tableau en indiquant le nombre total de données de la série statistique.
 - d) Énumérer les données de la série statistique à l'origine de la distribution.
 - e) Construire le graphique approprié pour représenter la distribution.
5. Qu'est-ce qui caractérise la clientèle des navires de croisières qui sillonnent le Saint-Laurent?

Un sondage effectué auprès d'un échantillon de 2 330 voyageurs pendant qu'ils effectuaient une croisière sur le Saint-Laurent révèle les statistiques suivantes.

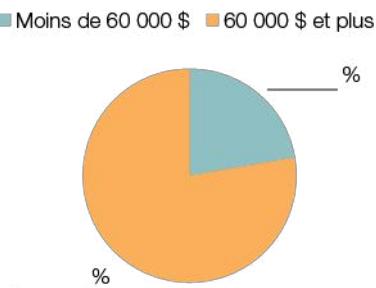
Source: Tourisme Québec. *Étude auprès des croisiéristes et des membres d'équipage des navires de croisières dans les ports du Saint-Laurent*, juin 2013.

- a) Parmi les répondants, il y a 1 265 femmes et 1 810 personnes dont le revenu familial est de 60 000 \$ et plus. Utiliser ces statistiques pour compléter les diagrammes ci-dessous.

Répartition des croisiéristes de l'échantillon selon le sexe, Québec, 2013



Répartition des croisiéristes de l'échantillon selon le revenu familial, Québec, 2013



- b) À la question «À quel groupe d'âge appartenez-vous?», on a obtenu les réponses suivantes : 124 personnes ont moins de 45 ans, 254 ont de 45 à moins de 55 ans, 692 ont de 55 à moins de 65 ans et 1 260 ont 65 ans et plus. Construire le tableau de distribution de la variable «âge» et en faire l'analyse.

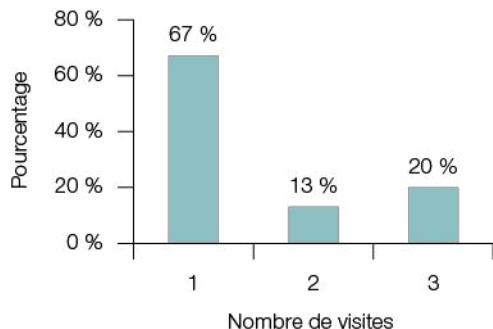
- c) i) À la question «Quelle est la capacité du navire (en nombre de passagers) sur lequel vous voyagez?», on a obtenu les réponses suivantes : 4 % voyagent sur un navire de moins de 1 000 passagers, 32 % sur un navire de 1 000 à moins de 2 000 passagers, 21 % sur un navire de 2 000 à moins de 3 000 passagers et 43 % sur un navire de 3 000 passagers et plus. Construire l'histogramme de cette distribution.

- ii) Vrai ou faux ? 64 % des navires de croisières ont une capacité de 2 000 passagers et plus.

- d) Parmi les répondants, 15 % sont déjà venus au Québec. On a posé la question suivante à ces derniers : «En excluant la présente visite, combien de fois êtes-vous venu au Québec?» On a obtenu les réponses suivantes : 67 % sont déjà venus 1 fois, 13 % sont venus 2 fois et 20 % sont venus 3 fois.

- i) Le graphique suivant présente ces statistiques, mais, malheureusement, il contient deux erreurs. Les trouver.

Répartition des croisiéristes de l'échantillon selon le nombre de visites au Québec avant la croisière, Québec, 2013



- ii) Combien de croisiéristes de l'échantillon sont déjà venus au Québec avant la croisière ?

- iii) Interpréter le pourcentage de 20 % obtenu à cette question du sondage.

6. Quelle amplitude devrait-on choisir pour grouper en classes les séries statistiques suivantes ? Dans chaque cas, indiquer la démarche justifiant ce choix et la 1^{re} classe.

- a) La série statistique compte 150 données dont la plus petite est 0,1 et la plus grande, 11,6.

- b) La série statistique compte 74 données dont la plus petite est 142 et la plus grande, 206.

7. Quel âge ont les professeurs qui vous enseignent ?

Voici des statistiques à ce sujet.

Répartition des professeurs de cégep selon l'âge, Québec, 2009-2010

Âge (en ans)	Pourcentage de professeurs
Moins de 30	10 %
$30 \leq X < 40$	27 %
$40 \leq X < 50$	27 %
$50 \leq X < 60$	28 %
60 et plus	8 %
Total	100 %

Source: Ministère de l'Éducation, du Loisir et du Sport, et Ministère de l'Enseignement supérieur. *Statistiques de l'éducation, édition 2011*, 2013.

- Donner quelques faits saillants de la distribution de l'âge des professeurs de cégep et formuler une hypothèse sur le pourcentage du corps professoral qu'il faudra renouveler sur une période de 10 ans, à partir de 2010.
- Construire l'histogramme de cette distribution et y superposer le polygone de fréquences.
- Construire l'ogive (courbe de fréquences cumulées).

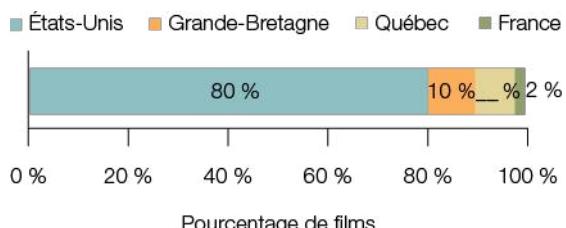
8. Combien de spectateurs un film à succès attire-t-il dans les salles de cinéma au Québec ? Y a-t-il des films québécois parmi les films les plus populaires ? Voyons ce qu'il en est.

Répartition des 50 films les plus populaires selon l'assistance, Québec, 2009 à 2011

Assistance (en milliers de spectateurs)	Nombre de films
$250 \leq X < 350$	11
$350 \leq X < 450$	14
$450 \leq X < 550$	10
$550 \leq X < 850$	11
$850 \leq X < 1\ 850$	4
Total	50

Source: Observatoire de la culture et des communications du Québec. Juin 2012.

Répartition des 50 films les plus populaires selon la provenance, Québec, 2009 à 2011



Source: Observatoire de la culture et des communications du Québec. Juin 2012.

- On note que les classes du tableau de distribution n'ont pas toutes la même amplitude. Construire l'histogramme qui conviendrait à cette distribution.

- Compléter le diagramme linéaire ainsi que l'analyse des données.

Analyse des données

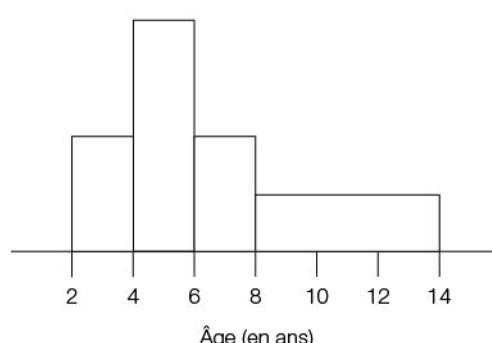
Parmi les 50 films ayant eu le plus de succès au Québec de 2009 à 2011, il y a _____ films québécois, soit _____ % des films. La proportion des 50 films à succès ayant attiré plus de 550 000 spectateurs au cinéma a été de _____ %.

(Le film québécois *De père en flic* se classe en deuxième position, après *Avatar*, avec une assistance de 1 242 370 spectateurs.)

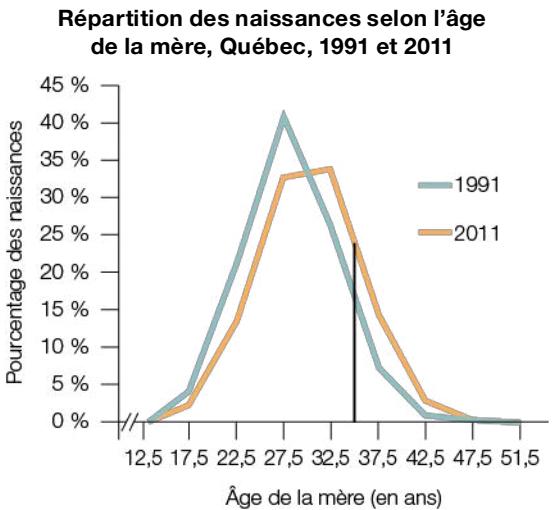
- Construire le tableau de distribution (en pourcentages) correspondant à l'histogramme ci-dessous.

- Superposer le polygone de fréquences sur cet histogramme.

Répartition des enfants selon l'âge



10. On dit qu'il n'est pas rare de nos jours de voir des femmes de plus de 35 ans donner naissance à un enfant. Mythe ou réalité ? Pour répondre à cette question, on a construit les polygones de fréquences ci-dessous. Qu'est-ce qui, visuellement, permet de conclure que le pourcentage de nouveau-nés dont la mère a plus de 35 ans est plus élevé en 2011 qu'en 1991 ?



Source: Institut de la statistique du Québec. *Naissances et fécondité*, 2012.

11. La mise en situation à la page 11 présentait la série statistique suivante pour le pourcentage d'étudiants en formation technique pour chacun des 48 collèges publics du Québec en 2010.

63,8 %	67,1 %	46,6 %	51,1 %	52,2 %	59,1 %
71,6 %	44,5 %	34,6 %	72,8 %	67,1 %	68,3 %
55,9 %	54,1 %	77,8 %	<u>78,6 %</u>	61,1 %	69,7 %
59,1 %	50,0 %	49,3 %	56,7 %	55,2 %	56,3 %
64,8 %	56,6 %	75,6 %	52,9 %	63,8 %	63,4 %
60,2 %	63,4 %	62,4 %	72,1 %	62,4 %	41,0 %
42,1 %	42,6 %	56,0 %	74,8 %	65,0 %	59,8 %
<u>23,1 %</u>	59,5 %	41,1 %	66,1 %	64,7 %	67,2 %

- a) Quelle amplitude de classes et quelle limite inférieure pour la 1^{re} classe devrait-on choisir pour grouper ces données en classes ?
 b) Construire le tableau de distribution du pourcentage d'étudiants en formation technique.

(Pour dépouiller rapidement les données, il est fortement recommandé d'utiliser la technique de dénombrement présentée à la page 18.)

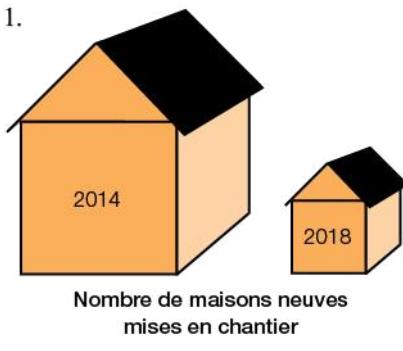
- c) Analyser la distribution.

12. a) Des étudiants contestant la hausse des frais de scolarité ont distribué le tract reproduit ci-dessous pour inviter leurs compagnons à venir manifester leur désaccord devant le Parlement. La représentation graphique de l'effet appréhendé par les étudiants au sujet de la hausse des frais de scolarité est-elle honnête ? Justifier la réponse.

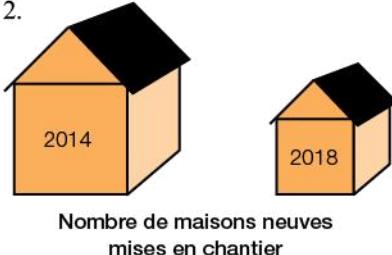


- b) On estime que le nombre de maisons neuves mises en chantier en 2018 sera égal à la moitié de ce qu'il était en 2014. Lequel des deux graphiques suivants donne la meilleure représentation de cette estimation ?

1.



2.



1.3 Les mesures de tendance centrale

Nous avons vu dans la section précédente comment présenter les données d'une série statistique sous forme de tableau ou de graphique afin d'en faire une première analyse. Les mesures de tendance centrale étudiées dans la présente section servent à raffiner cette analyse. Ces mesures permettent de représenter une série statistique par un seul nombre. Nous étudierons trois mesures de tendance centrale: la moyenne, le mode et la médiane.

1.3.1 La moyenne

Si l'on désire représenter une série de données par un et un seul nombre, comme les notes à un examen, la première mesure à laquelle on pense est la moyenne des données. La moyenne est la mesure de tendance centrale la plus connue et la plus utilisée pour représenter les données d'une série statistique. Avec la mise en situation suivante, nous apprendrons à représenter graphiquement une moyenne et à la calculer de trois façons différentes: avec les données brutes, avec les effectifs et avec les pourcentages du tableau de distribution.

MISE EN

SITUATION

Reprendons la série statistique donnant le nombre de programmes en techniques administratives offerts dans chacun des 48 collèges publics du Québec en 2010.

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

Source : Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

Calcul de la moyenne avec des données brutes

Comme nous le savons, le calcul de la moyenne avec des données brutes consiste à additionner l'ensemble des données, puis à diviser la somme obtenue par le nombre de données.

Moyenne avec des données brutes

$$\text{Moyenne} = \frac{\text{somme des données}}{\text{nombre total de données}}$$

Pour la moyenne des données de la mise en situation, on a :

$$\mu = \frac{2 + 1 + 3 + \dots + 4 + 4}{48} = 3,4 \text{ programmes}$$

Interprétation

En 2010, les collèges publics québécois offrent en moyenne 3,4 programmes en techniques administratives par collège.

NOTE

Le résultat du calcul d'une moyenne ne doit pas être arrondi à l'entier sous prétexte que les données brutes sont entières : la moyenne est un nombre théorique. Nous conviendrons de conserver une décimale après la virgule.

Écriture symbolique

Nous noterons la moyenne par le symbole μ , qui se lit «mu» (m dans l'alphabet grec), et le nombre de données par la lettre N .

En représentant chaque donnée de la série statistique par les symboles : x_1, x_2, x_3, x_4 et ainsi de suite, on obtient la formule suivante pour décrire le calcul d'une moyenne avec des données brutes :

$$\mu = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$$

On peut simplifier l'écriture de cette formule à l'aide de la notation sigma, symbolisée par \sum (S dans l'alphabet grec). Ce symbole indique que l'on doit faire la somme de tous les termes de forme x_i , l'indice i variant de 1 à N .

$$\mu = \frac{\sum x_i}{N}, \text{ pour } i \text{ variant de 1 à } N$$

NOTE

Dans le cadre d'un sondage, on emploie le symbole \bar{x} (x barre) pour désigner la moyenne des données de l'échantillon et μ pour la moyenne des données de la population. De même, on utilise le symbole n pour désigner le nombre de données de l'échantillon et N pour le nombre de données de la population. La formule pour obtenir la moyenne des données d'un échantillon s'écrit donc :

$$\bar{x} = \frac{\sum x_i}{n}, \text{ pour } i \text{ variant de 1 à } n$$

Calcul de la moyenne avec les effectifs du tableau de distribution

Il est plus rapide de calculer la moyenne d'une série statistique avec les effectifs du tableau de distribution qu'avec les données brutes. Rappelons le tableau de distribution du nombre de programmes en techniques administratives.

Répartition des 48 collèges publics selon le nombre de programmes en techniques administratives offerts, Québec, 2010

Nombre de programmes en techniques administratives	Nombre de collèges	Pourcentage de collèges
0	1	2,1 %
1	1	2,1 %
2	11	22,9 %
3	9	18,8 %
4	16	33,3 %
5	9	18,8 %
6	1	2,1 %
Total	48	100,1 %

Source: Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

Avec les effectifs du tableau de distribution, on calcule ainsi la moyenne du nombre de programmes en techniques administratives offerts par les collèges publics québécois :

$$\mu = \frac{0 \times 1 + 1 \times 1 + 2 \times 11 + 3 \times 9 + 4 \times 16 + 5 \times 9 + 6 \times 1}{48} = 3,4 \text{ programmes}$$

NOTE

Il est préférable de prendre l'habitude de respecter l'ordre d'écriture (valeur \times effectif) des produits des termes ci-dessus, car c'est dans cet ordre qu'il faut entrer les données lorsqu'on utilise le mode statistique d'une calculatrice.

Moyenne avec les effectifs

$$\text{Moyenne} = \frac{\text{somme des produits de chaque valeur de la variable par son effectif}}{\text{nombre total de données}}$$

Écriture symbolique

En notant par la lettre n_i l'effectif de la valeur x_i , on obtient la formule suivante :

$$\mu = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{N}, \text{ où } k \text{ représente le nombre de valeurs de la variable}$$

En utilisant la notation sigma, on obtient :

$$\mu = \frac{\sum x_i n_i}{N}, \text{ pour } i \text{ variant de 1 à } k$$

Calcul de la moyenne avec les pourcentages du tableau de distribution

Comment faut-il procéder pour calculer une moyenne si le tableau de distribution ne fournit que les pourcentages de données pour chaque valeur de la variable ?

Pour trouver une formule qui permettrait de calculer une moyenne en utilisant les pourcentages, reprenons le calcul de la moyenne avec les effectifs et apportons les modifications d'écriture suivantes :

$$\mu = \frac{0 \times 1 + 1 \times 1 + 2 \times 11 + \cdots + 6 \times 1}{48} = 3,4$$

$$\mu = \frac{0 \times 1}{48} + \frac{1 \times 1}{48} + \frac{2 \times 11}{48} + \cdots + \frac{6 \times 1}{48} = 3,4$$

$$\mu = 0 \times \frac{1}{48} + 1 \times \frac{1}{48} + 2 \times \frac{11}{48} + \cdots + 6 \times \frac{1}{48} = 3,4$$

$$\mu = 0 \times 0,021 + 1 \times 0,021 + 2 \times 0,229 + \cdots + 6 \times 0,021 = 3,4$$

En exprimant les décimales en pourcentage, on obtient la fréquence relative en pourcentage associée à chaque valeur :

$$\mu = 0 \times 2,1 \% + 1 \times 2,1 \% + 2 \times 22,9 \% + \cdots + 6 \times 2,1 \% = 3,4 \text{ programmes}$$

Moyenne avec les pourcentages

Moyenne = somme des produits de chaque valeur de la variable par son pourcentage

NOTE

Il est à remarquer que nous n'avons pas à diviser, dans ce cas-ci, par le total des données, car cette division a déjà été faite dans le calcul du pourcentage.

Écriture symbolique

En notant f_i la fréquence relative de la valeur x_i , on obtient la formule suivante :

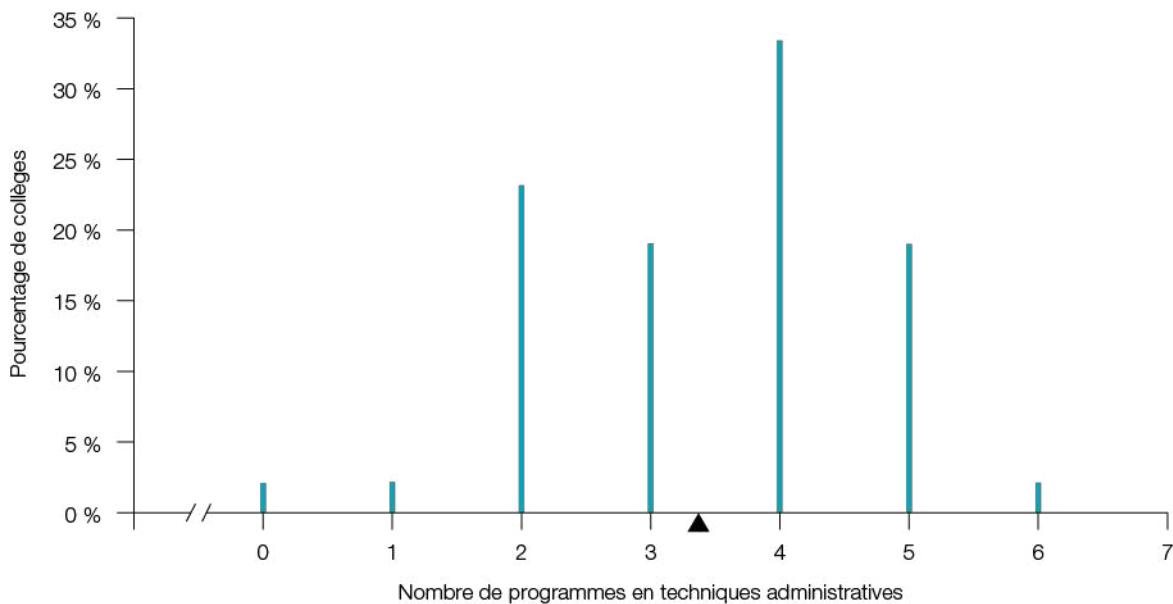
$$\mu = x_1 f_1 + x_2 f_2 + \dots + x_k f_k = \sum x_i f_i$$

pour i variant de 1 à k , où k est le nombre de valeurs de la variable.

Représentation graphique de la moyenne

Voici le diagramme en bâtons de la distribution du nombre de programmes en techniques administratives.

Répartition des 48 collèges publics selon le nombre de programmes en techniques administratives offerts, Québec, 2010



En plaçant un **pivot** sous l'axe horizontal du diagramme en bâtons à l'endroit où se situe la moyenne, on remarque que la moyenne correspond graphiquement au centre d'équilibre du diagramme. On applique ici le principe des balançoires à bascule : on doit s'imaginer que l'axe horizontal est une planche sous laquelle il faut placer un pivot pour que les bâtons du diagramme se trouvent en position d'équilibre sur cette planche.

EXEMPLE 1

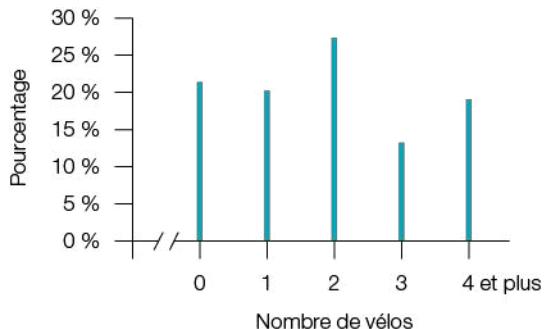
Combien de vélos les ménages québécois possèdent-ils ? Voici des statistiques à ce sujet.

Répartition des ménages selon le nombre de vélos, Québec, 2010

Nombre de vélos	Pourcentage de ménages
0	21 %
1	20 %
2	27 %
3	13 %
4 et plus	19 %
Total	100 %

Source: Vélo Québec. *État de la pratique du vélo au Québec en 2010*, mai 2011.

Répartition des ménages selon le nombre de vélos, Québec, 2010



- Estimer graphiquement la moyenne de cette distribution : $\mu \approx \underline{\hspace{2cm}}$.
- Calculer et interpréter la moyenne en remplaçant la catégorie «4 et plus» par le nombre «4».

Solution

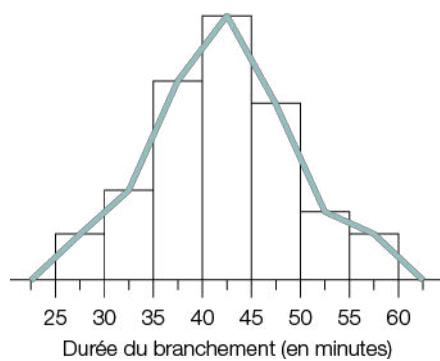
EXEMPLE 2

À la suite de l'étude menée auprès d'un échantillon de 40 clients d'un café (voir la page 21), on a construit le tableau de distribution et l'histogramme suivants.

Répartition des 40 clients de l'échantillon selon la durée du branchement au réseau Wi-Fi

Durée du branchement (en minutes)	Nombre de clients
$25 \leq X < 30$	2
$30 \leq X < 35$	4
$35 \leq X < 40$	9
$40 \leq X < 45$	12
$45 \leq X < 50$	8
$50 \leq X < 55$	3
$55 \leq X < 60$	2
Total	40

Répartition des 40 clients de l'échantillon selon la durée du branchement au réseau Wi-Fi



- a) Placer un pivot sous l'axe horizontal du graphique afin d'estimer la moyenne de la distribution.

$$\mu \approx \underline{\hspace{2cm}}$$

- b) Pour calculer la durée moyenne du branchement des clients au réseau, la formule utilisant les effectifs est tout indiquée, puisque le tableau nous les fournit. Rappelons que cette formule nécessite de multiplier chaque valeur par son effectif. Or, si nous savons qu'il y a 2 données dans la classe $25 \leq X < 30$, nous ne connaissons pas leur valeur. Il en est de même pour les autres classes. Que faire ?

Quelle valeur pourrions-nous utiliser pour représenter la valeur des données d'une classe ?

- c) Calculer et interpréter la moyenne de la distribution.

Solution

- d) La moyenne que l'on obtiendrait en utilisant les données brutes de la page 21 donnerait-elle la même réponse ?

À RETENIR

Pour effectuer des calculs sur des données groupées en classes, on utilise le centre de la classe pour représenter les valeurs des données d'une classe.

À moins d'avis contraire, une classe ouverte se verra attribuer une amplitude égale à celle des autres classes aux fins des calculs.

EXERCICE DE COMPRÉHENSION | 1.6

Quelle est l'intensité de l'utilisation d'Internet chez les jeunes de 16 à 24 ans ?

Voici des statistiques à ce sujet.

Répartition des internautes de 16 à 24 ans selon le temps d'utilisation d'Internet, Canada, 2012

Centre de classe	Nombre d'heures par semaine	Pourcentage d'internautes
	Moins de 5 h	23,4 %
	[5 h; 10 h[27,1 %
	[10 h; 20 h[25,4 %
	[20 h; 30 h[12,1 %
	30 h et plus	12,0 %
	Total	100,0 %

Source: Statistique Canada. Tableau 358-0220, CANSIM, novembre 2013.

- Quel pourcentage d'internautes de 16 à 24 ans sont connectés à Internet pendant 10 heures ou plus par semaine ?
- En moyenne, pendant combien d'heures par semaine un internaute de 16 à 24 ans est-il connecté à Internet ? (Attribuer une amplitude de 10 h à la dernière classe.)

Moyenne pondérée

Nous étudierons ce type de moyenne à partir de la mise en situation suivante :

MISE EN SITUATION

Supposons qu'un étudiant ait obtenu les notes suivantes, sur 100 points, en français :

Examen 1 : 65

Examen 2 : 70

Travail : 75

❓ Trouver sa moyenne pour la session si la pondération de chaque évaluation est :

Examen 1 : 25 %

Examen 2 : 35 %

Travail : 40 %

Moyenne =

Comme vous pouvez le constater, dans le calcul de cette moyenne, nous n'avons pas accordé la même importance à chaque évaluation, puisque la moyenne n'a pas été obtenue en additionnant les notes pour ensuite diviser cette somme par 3. Cette façon de procéder n'aurait pas eu de sens dans le contexte du problème. Nous avons plutôt donné un poids différent (une pondération) à chaque évaluation.

Lorsque le calcul d'une moyenne se fait en multipliant chaque valeur d'une série par sa pondération, on dit que l'on calcule la moyenne pondérée de la série. La pondération est déterminée selon l'importance qui est accordée à chaque valeur par rapport aux autres valeurs de la série. La somme des pondérations doit toujours être égale à 100 %.

EXEMPLE

Quel est le salaire moyen d'un titulaire d'un DEC en formation technique à son entrée sur le marché du travail ? Une étude auprès des diplômés de la promotion 2013 révèle que le salaire hebdomadaire moyen des techniciens qui ont décroché un emploi à temps plein à la fin de leurs études est de 686 \$ pour les femmes et de 764 \$⁴ pour les hommes.

Source: Ministère de l'Enseignement supérieur. *La relance au collégial en formation technique – 2013. La situation d'emploi des personnes diplômées. Enquêtes de 2011/2012/2013*, avril 2014.

- a) Le calcul suivant donne-t-il le salaire moyen d'un technicien, pour un emploi à temps plein, à son entrée sur le marché du travail ? Commenter.

$$\frac{686 \text{ \$} + 764 \text{ \$}}{2} = 725 \text{ \$}$$

Solution

- b) Sachant que 64 % des nouveaux diplômés qui occupent un emploi à temps plein sont des femmes, déterminer le salaire moyen d'un technicien, pour un emploi à temps plein, à son entrée sur le marché du travail.

Solution

EXERCICE DE COMPRÉHENSION | 1.7

Avoir le pied dansant, est-ce que c'est payant ?

En 2010, une étude effectuée auprès des 650 danseurs professionnels du Québec indique un revenu moyen de 18 514 \$ pour les 124 danseurs ayant moins de 6 ans d'expérience, de 25 323 \$ pour les 290 danseurs ayant de 6 à 15 ans d'expérience et de 35 809 \$ pour les 236 danseurs ayant plus de 15 ans d'expérience.

À partir de ces statistiques, calculer le revenu moyen d'un danseur professionnel québécois et le comparer au revenu personnel moyen des Québécois, qui était de 34 000 \$ en 2010.

Source: Institut de la statistique du Québec. *Enquête auprès des danseurs et chorégraphes du Québec, 2010*, janvier 2013.

4. Une partie de l'écart salarial entre les hommes et les femmes s'explique par le fait que les hommes travaillent en moyenne plus d'heures par semaine que les femmes, soit 39,2 heures comparativement à 36,6 heures.

Solution

1.3.2 Le mode et la classe modale

Le **mode** est la valeur ou la catégorie qui revient le plus souvent dans une série statistique. La **classe modale** est la classe qui regroupe le plus de données. On considère le centre de cette classe comme une approximation du mode de la distribution.

NOTE

Le mode et la classe modale sont significatifs seulement si leur fréquence est nettement différente de celle des autres valeurs ou classes.

EXEMPLE 1

Déterminer et interpréter le mode de la distribution suivante.

Répartition des 48 collèges publics selon le nombre de programmes en techniques administratives offerts, Québec, 2010

Nombre de programmes en techniques administratives	0	1	2	3	4	5	6	Total
Nombre de collèges	1	1	11	9	16	9	1	48
Pourcentage de collèges	2,1 %	2,1 %	22,9 %	18,8 %	33,3 %	18,8 %	2,1 %	100,0 %

Solution

Le mode est _____.

Interprétation

En 2010, une pluralité de collèges publics québécois (33,3 %) offrent quatre programmes en techniques administratives.

NOTE

Le mot **pluralité** signifie «le plus grand nombre». Il indique que le pourcentage entre parenthèses est le plus grand pourcentage de la distribution (quand il est supérieur à 50 %, on peut employer le mot **majorité**).

EXEMPLE 2

Déterminer et interpréter la classe modale de la distribution suivante.

Répartition des 40 clients de l'échantillon selon la durée du branchement au réseau Wi-Fi

Durée du branchement (en minutes)	Nombre de clients	Pourcentage de clients
$25 \leq X < 30$	2	5,0 %
$30 \leq X < 35$	4	10,0 %
$35 \leq X < 40$	9	22,5 %
$40 \leq X < 45$	12	30,0 %
$45 \leq X < 50$	8	20,0 %
$50 \leq X < 55$	3	7,5 %
$55 \leq X < 60$	2	5,0 %
Total	40	100,0 %

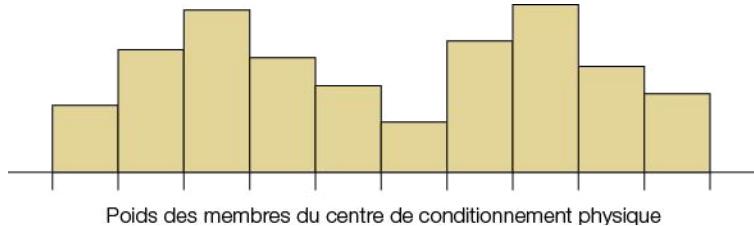
La classe modale est _____.

Interprétation

Une pluralité des clients de l'échantillon (30 %) ont été connectés au réseau pendant 40 à 45 minutes.

EXEMPLE 3

Un centre mixte de conditionnement physique a groupé les données portant sur le poids de ses membres. L'histogramme ci-contre présente la distribution.



- a) On dit d'une telle distribution qu'elle est **bimodale**, car deux classes se démarquent des autres avec approximativement la même fréquence. Pouvez-vous indiquer la cause de cette bimodalité ?

Solution

NOTE

Le caractère bimodal d'une distribution indique parfois la présence de deux sous-populations plus homogènes que la population globale.

- b) Serait-il judicieux d'utiliser la moyenne de cette série de données comme mesure de tendance centrale pour représenter le poids des membres ?

Solution

EXEMPLE 4

Pour chacune des situations suivantes, donner et interpréter la meilleure mesure de tendance centrale.

- a) Dans un sondage, on a posé la question suivante : « Quelle est votre principale source d'information pour consulter l'actualité ou les nouvelles ? » Voici la distribution des réponses obtenues chez les répondants de 18 à 24 ans.

Répartition des répondants de 18 à 24 ans selon la principale source d'information consultée pour l'actualité et les nouvelles, Québec, 2011

Principale source d'information	Internet	Radio	Télévision	Presse écrite	Total
Pourcentage	53 %	10 %	26 %	11 %	100 %

Source: CEFRIQ. NETendances 2011: Internet comme source d'information des Québécois, vol. 2, n° 4.

Solution

- b) Voici les salaires de cinq ingénieurs travaillant pour une même entreprise :

41 500 \$ 42 250 \$ 58 550 \$ 64 750 \$ 120 800 \$

Solution

La prochaine section permettra de définir une nouvelle mesure de tendance centrale qui pourra être appliquée à l'exemple précédent.

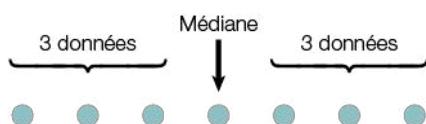
1.3.3 La médiane

La médiane, que l'on note « Me », est la valeur qui partage une série de données ordonnées en deux parties égales, chacune comprenant le même nombre de données. En d'autres termes, une valeur est la médiane d'une série de données ordonnées s'il y a autant de données à gauche qu'à droite de cette valeur.

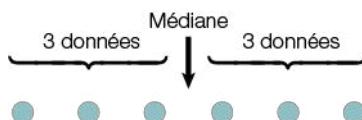
Données brutes

La définition de la médiane nous amène aux déductions suivantes :

- Pour une série statistique comptant un nombre impair de données ordonnées, pour qu'il y ait autant de données à gauche qu'à droite de la médiane, cette dernière doit être égale à la donnée centrale de la série. À titre d'exemple, pour sept données ordonnées, la médiane est la 4^e donnée.



- Pour une série statistique comptant un nombre pair de données ordonnées, pour qu'il y ait autant de données à gauche qu'à droite de la médiane, il faut choisir une valeur comprise entre les deux données centrales. Par convention, la médiane est égale à la valeur moyenne des deux données centrales de la série. À titre d'exemple, pour six données ordonnées, la médiane est la moyenne de la 3^e et de la 4^e donnée.



Pour déterminer la médiane avec des données brutes, on applique la procédure suivante :

1. On ordonne les données.
2. On détermine le nombre total de données de la série statistique.
3. Si ce nombre est impair, la médiane est la valeur de la donnée centrale de la série.
Si ce nombre est pair, la médiane est la moyenne des valeurs des deux données centrales.

EXEMPLE

Trouver la médiane des séries statistiques suivantes.

- a) Salaire de cinq ingénieurs travaillant pour une même entreprise :

41 500 \$ 42 250 \$ 58 550 \$ 64 750 \$ 120 800 \$

La médiane est _____. C'est la meilleure mesure de tendance centrale de la série, car elle n'est pas influencée par les valeurs extrêmes, contrairement à la moyenne.

Interprétation

Dans cette entreprise, au moins 50 % des ingénieurs gagnent 58 550 \$ ou moins.

NOTE

On dit «au moins» 50 %, car trois ingénieurs sur cinq gagnent 58 550 \$ ou moins.

b) Poids à la naissance de 10 nouveau-nés :

2 350 g	3 150 g	3 252 g	3 334 g	3 552 g
3 843 g	3 926 g	4 125 g	4 650 g	3 684 g

Solution

Interprétation

50 % des nouveau-nés pèsent moins de _____.

Données groupées par valeurs

On peut déterminer la médiane soit avec les effectifs, soit avec les pourcentages associés aux valeurs.

- Pour déterminer la médiane avec les effectifs :
On applique la même procédure que pour des données brutes.
- Pour déterminer la médiane avec les pourcentages :
 1. On repère la première valeur associée à un pourcentage cumulé égal ou supérieur à 50 %.
 2. Si le cumul des données est supérieur à 50 %, la médiane est égale à la valeur repérée.
Si le cumul des données est égal à 50 %, la médiane est la moyenne de la valeur repérée et de la valeur suivante.

EXEMPLE

a) Déterminer et interpréter la médiane de la distribution suivante.

Répartition des Centres de la petite enfance (CPE) d'une ville selon le nombre d'éducatrices

Nombre d'éducatrices	4	5	6	7	Total
Nombre de CPE	3	6	4	2	15
Pourcentage de CPE	20,0 %	40,0 %	26,7 %	13,3 %	100,0 %

Solution

- Calcul de la médiane avec les effectifs :

Avec un total de 15 données, un nombre impair, la médiane est la valeur de la 8^e donnée.

Médiane = 5 éducatrices

- Calcul de la médiane avec les pourcentages :

Pour obtenir un pourcentage cumulé d'au moins 50 %, il faut additionner les pourcentages associés aux valeurs 4 et 5, soit 20 % + 40 %, ce qui donne 60 %, une valeur supérieure à 50 %.

Médiane = 5 éducatrices

Interprétation

Au moins 50 % des CPE de cette ville ont 5 éducatrices ou moins.

b) Déterminer et interpréter la médiane de la distribution suivante.

Répartition de 160 entreprises selon le nombre de cadres

Nombre de cadres	2	3	4	5	6	Total
Nombre d'entreprises	16	24	40	48	32	160
Pourcentage d'entreprises	10 %	15 %	25 %	30 %	20 %	100 %

Solution

- Calcul de la médiane avec les effectifs :

Avec un total de 160 données, un nombre pair, la médiane est égale à la moyenne de la 80^e et de la 81^e donnée.

$$\text{Médiane} = \frac{4+5}{2} = 4,5 \text{ cadres}$$

- Calcul de la médiane avec les pourcentages :

Pour obtenir un pourcentage cumulé d'au moins 50 %, il faut additionner les pourcentages associés aux valeurs 2, 3 et 4, soit 10 % + 15 % + 25 %, ce qui donne exactement 50 %. La médiane est la moyenne des valeurs 4 et 5.

$$\text{Médiane} = \frac{4+5}{2} = 4,5 \text{ cadres}$$

En raison du contexte, on utilisera 4 (ou 5) au lieu de 4,5 dans l'interprétation.

Interprétation

50 % des entreprises ont 4 cadres ou moins (ou 5 cadres ou plus).

EXERCICES DE COMPRÉHENSION | 1.8

1. Vrai ou faux ? Si la médiane d'un examen sur 100 points est 68, on peut alors dire qu'au moins 50 % des étudiants ont une note de 68. _____
2. Calculer et interpréter la médiane des distributions suivantes.

a)

Répartition des employés qui ont pris un congé de maladie dans la dernière année selon le nombre de jours d'absence

Nombre de jours	1	2	3	4	5 et plus	Total
Nombre d'employés	9	6	7	3	5	30

$$\text{Médiane} =$$

Interprétation

- b) Un sondage effectué auprès des étudiants d'un cégep a permis de construire le tableau suivant.

Répartition des répondants selon le nombre de repas pris à la cafétéria durant la dernière semaine

Nombre de repas	0	1	2	3	4	5	Total
Pourcentage de répondants	30 %	15 %	15 %	18 %	12 %	10 %	100 %

Médiane = _____

Interprétation

3. Donner et interpréter la meilleure mesure de tendance centrale pour la distribution suivante.

Répartition des 50 albums¹ les plus vendus selon la langue d'enregistrement, Québec, 2012

Langue d'enregistrement	Anglais	Français	Autres	Total
Nombre d'albums	35	14	1	50

1. Inclut les albums sur support physique et les albums numériques.

Source: Observatoire de la culture et des communications du Québec. Mai 2013.

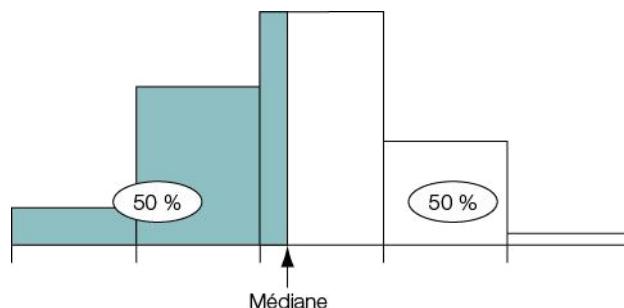
À titre d'information, les 5 premiers albums de la liste des 50 albums les plus vendus au Québec en 2012 sont : *Sans attendre* (Céline Dion), *Star Académie 2012* (Artistes variés), *21* (Adele), *Star Académie Noël* (Artistes variés), *Mes amours mes amis* (Paul Daraîche).

Meilleure mesure de tendance centrale :

Interprétation

Données groupées en classes

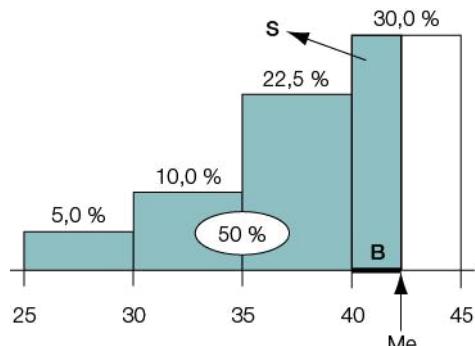
Lorsque les données d'une série statistique sont groupées en classes, la médiane est égale à la valeur sur l'axe horizontal qui divise la surface de l'histogramme (donc les données, en vertu du principe de la proportionnalité) en deux parties égales.



MISE EN SITUATION

Cherchons la médiane de la distribution suivante.

Esquisse de l'histogramme



Répartition des 40 clients de l'échantillon selon la durée du branchement au réseau Wi-Fi

Durée du branchement (en minutes)	Pourcentage de clients
$25 \leq X < 30$	5,0 %
$30 \leq X < 35$	10,0 %
$35 \leq X < 40$	22,5 %
$40 \leq X < 45$	30,0 %
$45 \leq X < 50$	20,0 %
$50 \leq X < 55$	7,5 %
$55 \leq X < 60$	5,0 %
Total	100,0 %

Notre objectif est de trouver une durée de branchement sur l'axe horizontal tel que 50 % de la surface de l'histogramme (donc 50 % des données) se situe à gauche de ce nombre. Pour obtenir 50 % de la surface, il faut additionner la surface du 1^{er} rectangle, du 2^e rectangle, du 3^e rectangle et une partie **S**, de la surface du 4^e rectangle.

- Surface : $50 \% = (5 \% + 10 \% + 22,5 \%) + S$

$$50 \% = 37,5 \% + S$$

$$S = 50 \% - 37,5 \%$$

$$S = 12,5 \%$$

- Médiane = $40 \text{ min} + B \text{ min}$, où **B** est la longueur de la base du rectangle de surface **S**. On trouve la valeur **B** par une règle de trois établissant un rapport entre la surface et la base du 4^e rectangle et la surface **S** et la base **B** du rectangle construit à l'intérieur de ce rectangle.

<u>Surface</u>	<u>Base</u>
30,0 %	\rightarrow 5 min
12,5 %	\rightarrow B min

D'où $B = \frac{12,5 \times 5}{30} = 2,1 \text{ min}$

$$\text{Médiane} = 40 \text{ min} + 2,1 \text{ min} = 42,1 \text{ min}$$

Interprétation

On peut estimer que, pour 50 % des clients de l'échantillon, le temps de branchement au réseau Wi-Fi a duré moins de 42,1 minutes.

EXEMPLE

Le tableau suivant donne la distribution de l'âge des arbres recensés sur un terrain boisé. Trouver et interpréter la médiane de cette distribution.

Répartition des arbres selon l'âge

Âge (en ans)	Pourcentage
[0; 10[8 %
[10; 20[28 %
[20; 30[32 %
[30; 40[20 %
[40; 50[12 %
Total	100,0 %

Solution

Esquisse de l'histogramme

Quelle mesure de tendance centrale faut-il utiliser ?

Chaque mesure de tendance centrale présente des avantages et des inconvénients. La moyenne est certainement la mesure la plus couramment utilisée pour représenter une série de données, mais on ne peut pas l'employer si la variable est qualitative. Son principal inconvénient est qu'elle peut être influencée par quelques valeurs extrêmes de la série statistique. Dans ce cas, on choisit la médiane comme mesure de tendance centrale.

La médiane est une mesure intéressante, car elle donne le centre de la distribution. Une différence importante entre la moyenne et la médiane indique que certaines données de la série statistique sont beaucoup plus grandes ou beaucoup plus petites que les autres.

Le mode est la mesure de tendance centrale qu'il faut utiliser si la variable est qualitative. Dans le cas d'une variable quantitative, le mode est significatif uniquement si sa fréquence est élevée.

Voici un tableau dans lequel on associe mesures de tendance centrale et types de variable.

Type de variable	Mesure possible
Variable qualitative	Mode
Variable quantitative	Moyenne Mode Médiane

1.4 Les mesures de position (quantiles)

Les mesures de position que sont les quantiles et la cote z servent à situer une donnée par rapport aux autres données d'une série statistique. Dans la présente section, nous étudions les quantiles. Par contre, l'étude de la cote z sera présentée à la section 1.6, après l'élaboration des outils mathématiques nécessaires à sa compréhension. La mise en situation suivante sert à illustrer le concept de quantile pour des données groupées en classes.

MISE EN SITUATION

Reprenons la distribution de l'âge des arbres recensés dans un boisé.

Répartition des arbres selon l'âge

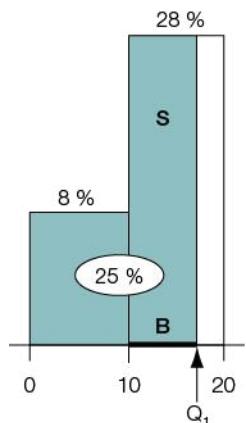
Âge (en ans)	Pourcentage
[0; 10[8 %
[10; 20[28 %
[20; 30[32 %
[30; 40[20 %
[40; 50[12 %
Total	100 %

Nous avons déjà estimé que 50 % des arbres avaient moins de 24,4 ans. Supposons qu'un voisin curieux pose la question suivante au propriétaire du boisé : « Le quart de vos arbres ont moins de quel âge ? »

Pour répondre à cette question, il faut trouver un âge, disons Q_1 , tel que 25 % des arbres aient un âge inférieur à Q_1 . La marche à suivre pour déterminer l'âge Q_1 est analogue à celle qui a servi à calculer la médiane.

Solution

Esquisse de l'histogramme



- Surface : $25 \% = 8 \% + S$
 $S = 25 \% - 8 \%$
 $S = 17 \%$
- $Q_1 : 10 + B$
On a : Surface Base
28 % \rightarrow 10 ans
17 % \rightarrow B ans
D'où $B = \frac{17 \times 10}{28} = 6,1$ ans
 $Q_1 = 10 + 6,1 = 16,1$ ans

Interprétation

On peut estimer que 25 % des arbres ont moins de 16,1 ans.

La valeur Q_1 est le 1^{er} quartile de la distribution. Les quartiles partagent une distribution en quatre parties égales comprenant 25 % des données. La médiane correspond au 2^e quartile : $Me = Q_2$.

Les **quantiles** sont des valeurs qui partagent une distribution en un certain nombre de parties égales. Les plus utilisés sont :

- les **quartiles** (Q_1, Q_2, Q_3), qui partagent une distribution en quatre parties comprenant 25 % des données ;
- les **quintiles** (V_1, V_2, V_3, V_4), qui partagent une distribution en cinq parties comprenant 20 % des données ;
- les **déciles** (D_1, D_2, \dots, D_9), qui partagent une distribution en dix parties comprenant 10 % des données ;
- les **centiles** (C_1, C_2, \dots, C_{99}), qui partagent une distribution en cent parties comprenant 1 % des données.

EXERCICES DE COMPRÉHENSION | 1.9

1. Quel âge ont les propriétaires de PME (petites et moyennes entreprises) ? Le tableau suivant donne des statistiques à ce sujet.

Répartition des propriétaires de PME selon l'âge, Canada, 2011

Âge	Pourcentage
Moins de 40 ans	12 %
[40 ans; 50 ans[28 %
[50 ans; 60 ans[40 %
60 ans et plus	20 %
Total	100 %

Source: Industrie Canada. *Principales statistiques relatives aux petites entreprises – Août 2013.*

- a) Calculer le 2^e décile de la distribution. (Fermer la 1^{re} classe en lui attribuant une amplitude de 10 ans.)

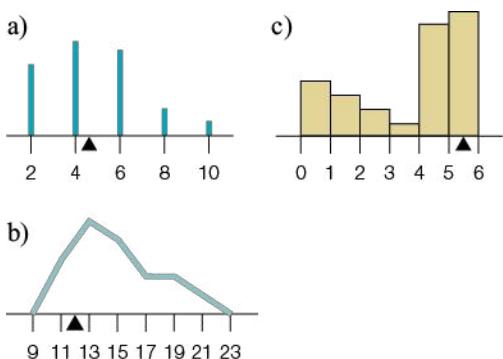
Solution

Esquisse de l'histogramme

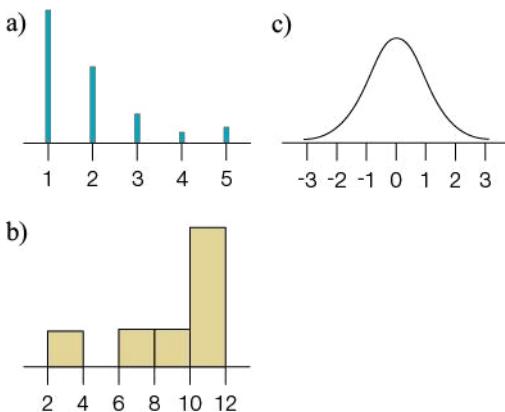
- b) Vrai ou faux. On interprète le décile trouvé en a) ainsi :
On peut estimer que 20 % des propriétaires de PME canadiennes ont 42,9 ans en 2011. _____
2. 50 ans correspond à quel quintile ? _____
3. Compléter. 60 ans correspond au _____ centile de la distribution.

EXERCICES 1.3

1. Le pivot placé sous l'axe horizontal de chacun des graphiques suivants permet d'estimer la moyenne de la distribution représentée. Y a-t-il des cas où cette estimation est de toute évidence erronée ?



2. Placer un pivot permettant d'estimer la moyenne des distributions suivantes.



3. Pour les séries statistiques suivantes, donner et interpréter chacune des mesures de tendance centrale, puis indiquer laquelle serait la plus représentative de la série statistique.

- a) Le nombre de calendriers vendus en une journée par sept personnes :

7 8 6 9 6 36 10

- b) Le nombre de spectateurs à chacune des six représentations d'une pièce de théâtre :

724 802 715 825 650 790

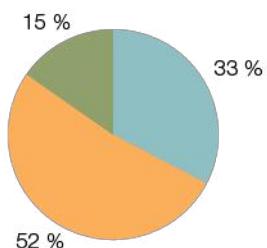
4. Quel est le rythme de production des écrivains professionnels⁵? Se cantonnent-ils dans un seul genre littéraire (roman, poésie, théâtre, etc.)? En 2010, une enquête effectuée auprès des 1 510 écrivains professionnels du Québec donne une réponse à ces questions.

Source: Institut de la statistique du Québec. *Les écrivains québécois. Portrait des conditions de pratique de la profession littéraire au Québec, 2010*, septembre 2011.

- a) Donner et interpréter la meilleure mesure de tendance centrale de la distribution ci-dessous.

Répartition des écrivains professionnels selon le rythme de publication, Québec, 2010

- Un livre ou moins tous les trois ans
- Un livre tous les deux ans
- Au moins un livre par année



- b) On compte 680 femmes parmi les 1 510 écrivains professionnels québécois. La moyenne de livres édités en carrière par celles-ci est de 10 livres alors qu'elle est de 11,8 livres pour les hommes. Utiliser ces informations pour calculer le nombre moyen de livres édités en carrière pour l'ensemble des auteurs québécois.

5. Écrivains qui ont publié au moins deux livres au cours de leur carrière.

- c) Déterminer et interpréter la moyenne, le mode et la médiane de la distribution suivante. Remplacer la catégorie «4 et plus» par le nombre «4» dans les calculs.

Répartition des écrivains professionnels selon le nombre de genres littéraires auxquels appartiennent leurs publications, Québec, 2010

Nombre de genres littéraires	Nombre d'écrivains	Pourcentage
1	512	33,9 %
2	450	29,8 %
3	285	18,9 %
4 et plus	263	17,4 %
Total	1 510	100,0 %

5. Peut-on vivre de sa plume au Québec? Voici des statistiques sur le revenu des écrivains.

Répartition des écrivains professionnels selon le revenu tiré des activités de création littéraire, Québec, 2010

Revenu (en \$)	Pourcentage
Moins de 5 000	64,9 %
[5 000; 20 000[22,1 %
[20 000; 40 000[8,8 %
[40 000; 60 000[2,2 %
60 000 et plus	2,0 %
Total	100,0 %

Source: Institut de la statistique du Québec. *Les écrivains québécois. Portrait des conditions de pratique de la profession littéraire au Québec, 2010*, septembre 2011.

- a) Parmi les 1 510 écrivains professionnels, combien ont tiré un revenu d'au moins 40 000 \$ de leurs activités de création littéraire?
- b) Donner et interpréter la classe modale de la distribution.
- c) Donner et interpréter la médiane de la distribution.
- d) Donner et interpréter la moyenne de la distribution, puis expliquer l'écart que l'on observe entre la moyenne et la médiane. Attribuez une amplitude de 20 000 \$ à la dernière classe.
- e) Quelle mesure de tendance centrale représente le mieux le revenu que les auteurs professionnels tirent de leur création?
6. En 2011-2012, 26 316 cégepiens de la formation technique ont reçu une aide financière pour leurs études. À combien s'élève cette aide? Voici de l'information sur les montants attribués.

Répartition des bénéficiaires du Programme de prêts et bourses de la formation technique selon le montant¹ d'aide attribué, 2011-2012

Montant attribué	Pourcentage
Moins de 2 000 \$	26,3 %
[2 000 \$; 4 000 \$[20,1 %
[4 000 \$; 6 000 \$[17,7 %
[6 000 \$; 8 000 \$[16,3 %
[8 000 \$; 10 000 \$[9,7 %
10 000 \$ et plus	9,9 %
Total	100,0 %

1. Prêt + bourse.

Source: Ministère de l'Enseignement supérieur. *Statistiques. Rapport 2011-2012*, 2014.

- a) Quel pourcentage de bénéficiaires de la formation technique ont reçu 6 000 \$ ou plus du Programme de prêts et bourses?
- b) Calculer et interpréter la moyenne de la distribution.
- c) Donner et interpréter la classe modale.
- d) Donner et interpréter la médiane.
- e) Donner et interpréter le 1^{er} quartile.

7. On mise sur l'immigration pour contrer le vieillissement de la population québécoise. Les nouveaux immigrants sont-ils plus jeunes que les Québécois? Voici des statistiques à ce sujet.

Répartition des immigrants accueillis au Québec en 2012 selon le groupe d'âge

Âge	Pourcentage
Moins de 15 ans	21,4 %
[15 ans; 30 ans[29,4 %
[30 ans; 45 ans[39,2 %
45 ans et plus	10,0 %
Total	100,0 %

Source: Institut de la statistique du Québec. Août 2013.

- a) En 2012, l'âge médian des Québécois est de 41,5 ans. Qu'en est-il de l'âge médian des immigrants que le Québec a accueillis cette année-là? Interpréter cette mesure.
- b) Donner et interpréter la classe modale de la distribution.
- c) Calculer et interpréter la moyenne de la distribution.
- d) Déterminer et interpréter le 2^e quintile (V_2) de la distribution.
- e) À quel décile correspond un âge de 45 ans? Interpréter cette mesure.

8. Calculer la note moyenne des cours suivants en tenant compte du nombre de crédits pour chaque cours. Quel nom donne-t-on à ce type de moyenne ?

Cours	Nombre de crédits	Note (sur 100 points)
Mathématiques	3	60
Histoire	2	70
Géographie	2	65
Français	3	80

9. Un ménage est constitué d'une personne ou d'un groupe de personnes qui occupent le même logement. De combien de personnes sont composés les ménages québécois en 2011 ?

Répartition des ménages selon le nombre de personnes par ménage, Québec, 2011

Nombre de personnes par ménage	Pourcentage
1	32,2 %
2	34,8 %
3	14,6 %
4	12,4 %
5	4,2 %
6 et plus	1,8 %
Total	100,0 %

Source: Statistique Canada. Recensement 2011.

- a) Donner le type de la variable étudiée.
 - b) Analyser la distribution du nombre de personnes par ménage.
 - c) Calculer et interpréter la moyenne en remplaçant la catégorie «6 et plus» par «6».
 - d) Calculer et interpréter la médiane.
10. Le poids moyen des personnes du groupe A est de 66,3 kg et celui du groupe B, de 60,2 kg. Supposons que l'on réunisse les deux groupes. Peut-on, dans les conditions suivantes, calculer le poids moyen du nouveau groupe ? Si oui, donner la moyenne.
- a) Les deux groupes contiennent le même nombre de personnes.
 - b) On ne connaît pas le nombre de personnes de chaque groupe.
 - c) Le groupe A comprend 10 personnes et le groupe B, 40 personnes.
11. Si l'on compare les naissances de 2011 à celles de 1991, observe-t-on une différence en ce qui concerne la distribution de l'âge de la mère ? Le tableau suivant permet d'effectuer cette analyse.

Répartition des naissances selon l'âge de la mère, Québec, 1991 et 2011

Âge de la mère (en ans)	Pourcentage des naissances	
	1991	2011
Moins de 20	4,1 %	2,5 %
[20; 25[20,4 %	13,7 %
[25; 30[41,0 %	32,7 %
[30; 35[26,4 %	33,8 %
[35; 40[7,2 %	14,4 %
40 et plus	0,9 %	2,9 %
Total	100,0 %	100,0 %

Source: Institut de la statistique du Québec. Naissances et fécondité, 2012.

- a) Compléter l'énoncé.

On peut estimer que la moyenne d'âge des mères qui ont donné naissance à un enfant était de 28,2 ans en 1991 et de _____ ans en 2011, une augmentation de _____ an(s) sur une période de 20 ans.

- b) Compléter l'énoncé.

L'âge médian des mères qui ont donné naissance à un enfant peut être estimé à 28,1 ans en 1991 et à _____ ans en 2011, une augmentation de _____ an(s) sur une période de 20 ans.

- c) Calculer et interpréter le 1^{er} décile de l'âge des mères qui ont donné naissance à un enfant en 2011.
d) Le 4^e quintile de la distribution de l'âge de la mère pour l'année 2011 est 34,6 ans. Interpréter cette mesure.

12. Selon une étude de Statistique Canada, dans les villes canadiennes de moins de 250 000 habitants, les travailleurs mettent 19 minutes en moyenne pour se rendre au travail. Qu'en est-il des régions métropolitaines comme Montréal où la congestion routière est particulièrement fréquente ?

Répartition des travailleurs selon la durée des déplacements pour se rendre au travail, région de Montréal, 2010

Durée (en min)	[0; 15[[15; 30[[30; 45[45 et plus	Total
Pourcentage	20 %	27 %	26 %	27 %	100 %

Source: Statistique Canada. Se rendre au travail: résultats de l'enquête sociale générale de 2010, mai 2013.

- a) Calculer la moyenne de la distribution ci-dessus, l'interpréter et la comparer avec le temps moyen que prennent les travailleurs des villes de moins de 250 000 habitants pour se rendre au travail.

- b) Quel pourcentage des travailleurs de la région de Montréal consacrent au moins une heure trente par jour à faire l'aller-retour entre la maison et le travail, si l'on suppose qu'ils prennent le même temps pour aller au travail que pour en revenir ?

- c) Trouver la médiane et l'interpréter.
d) Donner et interpréter le 1^{er} décile (D_1).
e) Donner et interpréter le 1^{er} quartile (Q_1).

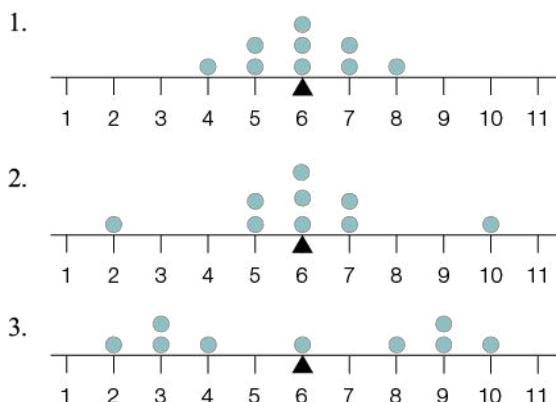
1.5 Les mesures de dispersion

Les mesures de dispersion permettent de mesurer la dispersion des données d'une série statistique autour de la moyenne. Nous étudierons plus particulièrement l'étendue et l'écart type.

1.5.1 L'étendue

MISE EN SITUATION

Analysons les trois séries statistiques suivantes, représentées par un pictogramme où chaque point est une donnée de la série dont la valeur correspond à la position du point sur l'axe. Pour chaque série, la moyenne des données est indiquée par le pivot.



Ces trois séries statistiques présentent des moyennes identiques, mais la dispersion des données autour de ces moyennes diffère d'une série à l'autre. Les données de la série 1 sont assez concentrées autour de la moyenne, alors que la série 2 comporte deux données qui sont passablement éloignées de la moyenne. Les données de la série 3 sont encore plus dispersées par rapport à la moyenne que celles de la série 2. Comment peut-on mesurer mathématiquement cette dispersion ?

L'étendue de la série pourrait peut-être nous permettre de la mesurer. Calculons l'étendue de chacune des séries :

	Série 1	Série 2	Série 3
Étendue	4	8	8

Comme l'étendue ne tient compte que de la plus grande et de la plus petite valeur de la série, cette mesure n'est pas assez subtile pour mesurer la différence de dispersion entre les données de la série 2 et celles de la série 3.

En fait, l'étendue sert très peu comme mesure de dispersion. On l'utilise surtout en contrôle de la qualité, où les échantillons sont souvent de petite taille (quatre ou cinq éléments) ; dans ce cas, l'étendue donne une bonne idée de la dispersion des résultats de l'échantillon.

1.5.2 La variance et l'écart type

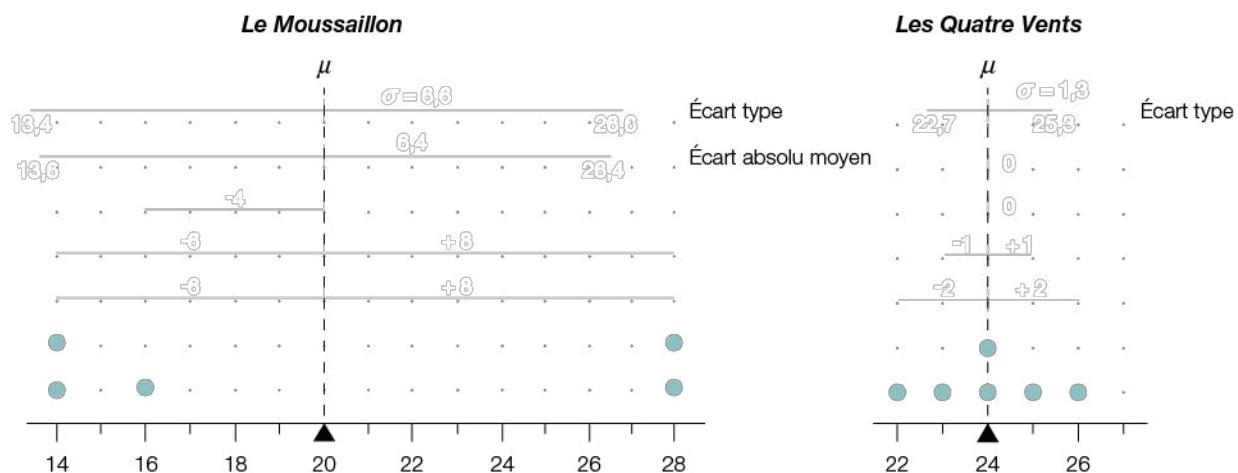
L'écart type que nous présenterons ici est une mesure de dispersion qui, contrairement à l'étendue, tient compte de toutes les valeurs de la série de données.

MISE EN SITUATION

Supposons qu'un étudiant de cégep offre ses services à titre de matelot pendant les vacances. Il a reçu deux offres d'emploi avec des conditions de travail semblables. Il décide de faire son choix en fonction de l'information qu'on lui a donnée sur l'âge moyen de ses compagnons de voyage. Il a donc préféré l'offre d'emploi du voilier *Le Moussaillon*, où l'âge moyen est de 20 ans, à celui du voilier *Les Quatre Vents*, où il est de 24 ans. Les pictogrammes ci-dessous montrent la distribution de l'âge pour chaque voilier. Notre cégepien a-t-il fait le bon choix ?

Bien sûr que non : il devra passer ses vacances en compagnie de M. et M^{me} Tremblay et de leurs trois neveux, alors que l'autre voilier semble accueillir l'équipage dynamique qu'il recherchait. La différence principale entre ces deux distributions est la dispersion des données par rapport à la moyenne. Comment s'y prendre pour mesurer cette dispersion ?

- Dans un premier temps, comme nous voulons que la mesure tienne compte de toutes les données, nous visualiserons l'écart entre chaque donnée et la moyenne de la série par un segment de droite.
- ?** Pour chaque pictogramme, superposer, sur les trois premières lignes en pointillé au-dessus des points, les segments de droite tracés entre chaque donnée et la moyenne, puis inscrire sur chaque segment la valeur de l'écart représenté.



- Cherchons maintenant un nombre qui sera un bon représentant des écarts observés. Pour le voilier *Le Moussaillon*, nous avons les écarts suivants : (-6), (-6), (-4), (8), (8). Lorsque l'on veut représenter des données par un et un seul nombre, notre premier réflexe est d'utiliser la moyenne. C'est ce que nous ferons ici.

- ?** Calculer la moyenne des cinq écarts trouvés à la page précédente :

Moyenne des écarts =

Ne soyez pas surpris du résultat obtenu : la moyenne des écarts donnera toujours 0. En effet, la moyenne étant le centre d'équilibre du pictogramme, la somme des écarts situés à gauche de la moyenne (-16, dans ce cas-ci) sera toujours égale, mais de signe négatif, à la somme des écarts situés à droite de la moyenne (16).

- Pour contourner le problème causé par les écarts négatifs, tout en gardant l'idée d'utiliser une moyenne pour mesurer la dispersion, regardons les choses sous un autre angle : au lieu de dire qu'il y a un écart de -6 ans entre 14 et 20 ans, nous dirons qu'il y a une distance de 6 ans entre ces deux valeurs. Rappelons que l'on utilise la valeur absolue pour indiquer que l'on ne doit pas tenir compte du signe devant un nombre, par exemple, $|-6| = 6$.

- ?** Calculer la moyenne des distances entre chaque donnée et la moyenne. On donne le nom d'**écart absolu moyen** au résultat obtenu.

Moyenne des distances =

Sur l'avant dernière ligne en pointillé du pictogramme «*Le Moussaillon*», tracer un segment de droite de longueur 6,4 de part et d'autre de la moyenne pour représenter l'écart absolu moyen.

- Comme la présence de valeurs absolues dans une formule complique souvent son utilisation, nous allons transformer cette formule pour qu'elle soit plus facile à exploiter. Un moyen approprié pour éliminer les valeurs absolues consiste à utiliser l'égalité voulant que la valeur absolue d'un nombre élevée au carré donne le même résultat que donnerait ce nombre élevé au carré. Par exemple, $|-6|^2 = (-6)^2 = 36$. Appliquons cette équivalence à l'écart absolu moyen calculé ci-dessus.

- ?** Élever chacun des écarts au carré dans le calcul de l'écart absolu moyen. On donne le nom de **variance** à ce résultat.

Variance =

La variance est la moyenne des carrés des écarts et elle s'exprime en unités carrées, en ans carrés dans ce cas-ci.

Il est évident que 43 ans^2 ne peut pas représenter les écarts observés : dire que les données se dispersent de 43 ans^2 autour de la moyenne serait quelque peu saugrenu. Que faire pour obtenir une mesure de dispersion un peu plus sensée ?

Il faut tout simplement extraire la racine carrée de la variance pour obtenir un écart «typique» de tous les écarts entre les données et la moyenne. On donne le nom d'**écart type** à cet écart noté σ , qui se lit «sigma» (s dans l'alphabet grec); la variance, quant à elle, est notée σ^2 .

- ?** Calculer l'écart type de l'âge de l'équipage du *Moussaillon*, puis le représenter sur le pictogramme par un segment de droite tracé de part et d'autre de la moyenne et y inscrire sa valeur.

Écart type : $\sigma =$

Utilité de l'écart type pour l'analyse de données

Généralement, on trouve la plupart des données d'une distribution entre la moyenne moins un écart type et la moyenne plus un écart type, soit entre $\mu - \sigma$ et $\mu + \sigma$. Lorsque la distribution prend la forme d'une cloche (modèle normal), environ les deux tiers des données seront comprises dans cet intervalle. En donnant ce sens à l'écart type calculé à la page précédente, on obtient l'interprétation qui suit.

Interprétation de l'écart type

La plupart des matelots (3 sur 5) du voilier *Le Mousaillon* ont un âge se situant à $\pm 6,6$ ans de la moyenne d'âge de l'équipage de ce navire, soit entre 13,4 ans et 26,6 ans.

- Pour le voilier *Les Quatre Vents*, nous avons les écarts suivants entre les données et la moyenne : (-2), (-1), (0), (0), (1), (2).

 Calculer et interpréter l'écart type de l'âge de l'équipage de ce voilier, puis le représenter sur le pictogramme par un segment de droite tracé de part et d'autre de la moyenne et y inscrire sa valeur.

Variance : $\sigma^2 =$

Écart type : $\sigma =$

Interprétation

La plupart des matelots (4 sur 6) du voilier *Les Quatre Vents* ont un âge se situant à _____ an(s) de la moyenne d'âge de l'équipage, soit entre _____ ans et _____ ans.

La formule suivante permet de généraliser les résultats :

Écart type et variance

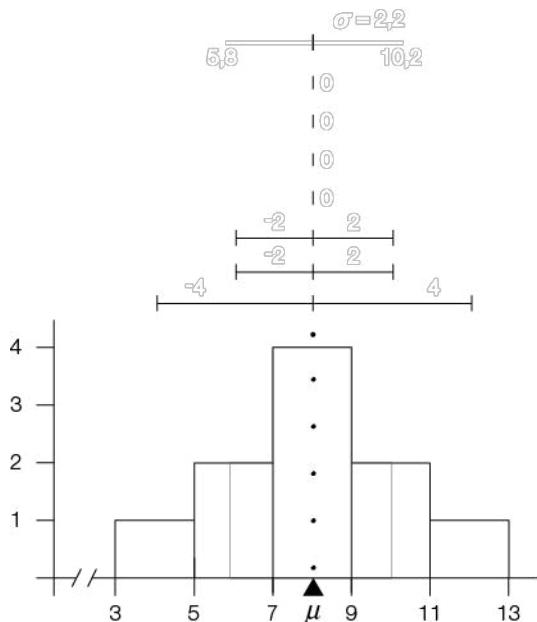
$$\text{Écart type } (\sigma) = \sqrt{\text{variance } (\sigma^2)} = \sqrt{\frac{\text{somme des carrés des écarts à la moyenne}}{\text{nombre total de données}}}$$

$$\text{Écart type } (\sigma) = \sqrt{\frac{\sum (x_i - \mu)^2 n_i}{N}} \quad (\text{ou } \sigma = \sqrt{\sum (x_i - \mu)^2 f_i} \text{ avec les fréquences relatives})$$

EXEMPLE

- Compléter le tableau de distribution en utilisant l'information fournie par l'histogramme.
- Indiquer sur les segments de droite au-dessus de l'histogramme la valeur des écarts à la moyenne.
- Calculer, interpréter et représenter graphiquement l'écart type de la distribution.

Solution



Moyenne : $\mu =$

Variance : $\sigma^2 =$

Écart type : $\sigma =$

Interprétation

La plupart des données se situent entre _____ et _____.

Tableau de distribution

Variable	Effectif
[3 ; 5[
[5 ; 7[
[7 ; 9[
[9 ; 11[
[11 ; 13[
Total	

Écart type corrigé dans le cas d'un échantillon

Dans le cadre d'une étude par sondage, l'écart type de l'échantillon est retenu pour estimer celui de la population. Toutefois, les statisticiens ont démontré que l'estimation est meilleure si l'on divise le numérateur de la formule de l'écart type par le nombre de données de l'échantillon, moins 1. On obtient ainsi l'écart type corrigé, qui prend la notation s . En se rappelant que la moyenne de l'échantillon se note \bar{x} et le nombre de données n , on obtient la formule suivante pour calculer l'écart type corrigé :

Écart type corrigé et variance corrigée

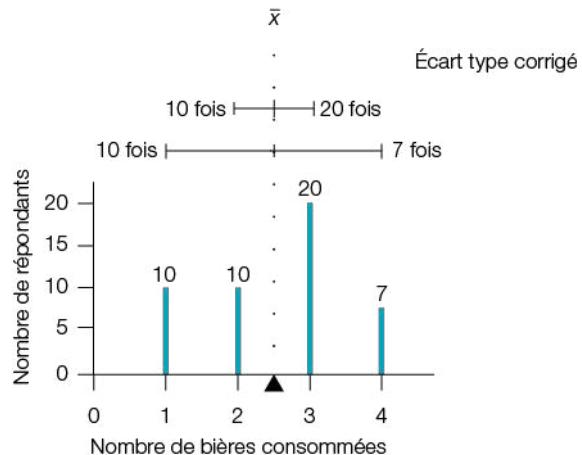
$$\text{Écart type corrigé } (s) = \sqrt{\text{variance corrigée } (s^2)} = \sqrt{\frac{\text{somme des carrés des écarts à la moyenne}}{\text{nombre total de données} - 1}}$$

$$\text{Écart type corrigé } (s) = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{n - 1}}$$

EXEMPLE

À la suite d'un 5 à 7 organisé par une entreprise pour souligner le départ d'un employé, on a demandé à un échantillon de 47 personnes combien de bières elles avaient consommées. Le diagramme suivant donne la répartition des répondants selon le nombre de bières consommées. Calculer et représenter graphiquement l'écart type corrigé de cette distribution.

Solution



Moyenne : $\bar{x} =$

Variance corrigée : $s^2 =$

Écart type corrigé : $s =$

Interprétation

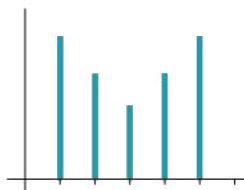
_____ des répondants ont consommé _____ bières pendant le 5 à 7.
(Les nombres 2 et 3 sont les entiers compris dans l'intervalle [1,5 ; 3,5]).

EXERCICES DE COMPRÉHENSION | 1.10

1. Sans faire de calculs, indiquer le numéro du diagramme en bâtons ci-dessous :

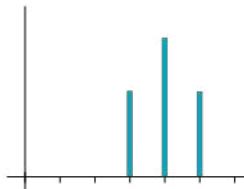
a) qui a la plus petite moyenne : _____

1.



b) qui a le plus petit écart type : _____

2.



c) dont la valeur de l'écart type est égale à la longueur du petit segment de droite qui figure sous le graphique 2 : _____

Écart type —————

- 2. a) Une série A représente l'âge de cinq membres d'une famille et une série B, l'âge des étudiants d'une classe de cégep. Laquelle des deux séries aura le plus grand écart type? _____
- b) Un professeur fait passer un test de classement en mathématiques dans ses deux groupes. Les deux groupes ont obtenu la même moyenne, mais l'écart type du groupe A est plus grand que celui du groupe B. Dans quel groupe y a-t-il moins de disparité entre les étudiants en mathématiques? _____
- c) Vrai ou faux? Toutes les données d'une distribution dont la moyenne est 70 et l'écart type, 10 sont comprises entre 60 et 80. _____
- d) Dans une région aride du globe, on enregistre la précipitation quotidienne (en millimètres) durant 60 jours consécutifs. La moyenne des 60 données est égale à 0. Que vaut l'écart type?
- _____

3. Nous avons vu, dans la mise en situation de la section 1.5.2 «La variance et l'écart type» (*voir la page 62*), que la moyenne d'âge de l'équipage du voilier *Le Moussaillon* était de 20 ans avec un écart type de 6,6 ans. Si l'équipage est toujours le même dans six ans, quels seront alors la moyenne et l'écart type de l'âge de l'équipage?

$$\mu = \text{_____} \text{ et } \sigma = \text{_____}$$

4. La moyenne de la distribution suivante est 10.

Valeur	6	8	14	20	Total
Effectif	3	3	2	1	9

- a) Calculer la variance et l'écart type de la distribution.

$$\text{Variance: } \sigma^2 =$$

$$\text{Écart type: } \sigma =$$

- b) Calculer la variance corrigée et l'écart type corrigé de la distribution.

$$\text{Variance corrigée: } s^2 =$$

$$\text{Écart type corrigé: } s =$$

1.5.3 Le coefficient de variation

Le coefficient de variation permet de mesurer la **dispersion relative** d'une série de données. Nous allons découvrir cette mesure à partir de la mise en situation suivante.

MISE EN

SITUATION

Supposons que, en 1930, le salaire moyen des six couturières d'une petite entreprise était de 46 \$ par semaine, avec un écart type de 10,30 \$. En ce temps-là, les ouvrières étaient payées à la pièce. Voici la distribution des salaires des couturières :

Salaires, 1930 : 30 \$ 37 \$ 44 \$ 50 \$ 55 \$ 60 \$

Moyenne : 46 \$

Écart type : 10,30 \$

Supposons qu'en 2012, le salaire moyen de six couturières (bien entendu, ce ne sont plus celles de 1930 !) était de 356 \$ par semaine, avec un écart type de 10,30 \$. Voici la distribution des salaires des couturières :

Salaires, 2012 : 340 \$ 347 \$ 354 \$ 360 \$ 365 \$ 370 \$

Moyenne : 356 \$

Écart type : 10,30 \$

Il est facile de comprendre que, même si les écarts types sont identiques dans les deux cas, leur importance n'est pas la même par rapport à la moyenne ; une variation de 10 \$ sur un salaire moyen de 46 \$ est beaucoup plus importante qu'une variation de 10 \$ sur un salaire moyen de 356 \$.

On peut facilement observer qu'en 1930 il y avait beaucoup plus de disparité entre les salaires : ceux-ci variaient du simple au double, de 30 \$ à 60 \$, ce qui n'était pas du tout le cas en 2012.

On a voulu mesurer mathématiquement l'importance relative de l'écart type par rapport à la moyenne. Cette mesure se traduit par le **coefficient de variation**, que l'on note CV, obtenu en divisant l'écart type par la moyenne de la distribution que l'on exprime par la suite en pourcentage.

Voici le coefficient de variation des salaires des couturières :

$$\text{En 1930: } \text{CV} = \frac{10,30 \$}{46 \$} \times 100 \% = 22,4 \%$$

$$\text{En 2012: } \text{CV} = \frac{10,30 \$}{356 \$} \times 100 \% = 2,9 \%$$

Interprétation

Les coefficients de variation indiquent qu'en 1930, il y avait beaucoup plus de disparité entre les salaires qu'en 2012.

Le coefficient de variation est une mesure de dispersion relative des données ; on le calcule ainsi :

Coefficient de variation

$$CV = \frac{\text{écart type}}{\text{moyenne}} \times 100 \% = \frac{\sigma}{\mu} \times 100 \%$$

NOTE

Si l'on travaille avec un échantillon, on remplace σ par s et μ par \bar{x} dans la formule.

Homogénéité des données

Le coefficient de variation permet de mesurer l'homogénéité des données d'une série statistique. Plus le coefficient de variation est faible, plus les données sont homogènes et plus la moyenne est représentative des données. On considère qu'un coefficient de variation inférieur à 15 % indique une bonne homogénéité des données. Il est à remarquer que le coefficient de variation est une mesure pure, sa valeur n'a pas d'unités de mesure ; cela permet de comparer l'homogénéité de plusieurs séries statistiques, même si les données de ces dernières ne sont pas exprimées dans les mêmes unités de mesure.

NOTE

On ne peut pas calculer le coefficient de variation pour des données mesurées selon une échelle d'intervalle, car son calcul nécessite l'utilisation de la division, une opération qui n'est pas permise avec cette échelle de mesure. À titre d'exemple, supposons que la température moyenne en janvier soit de -5 °C avec un écart type de 2 °C. En divisant ces deux nombres, on obtient un coefficient de variation de -40 %, un résultat plutôt surprenant !

EXEMPLE

Supposons que l'on désire comparer les revenus des médecins de la Russie avec ceux du Québec. Voici la moyenne et l'écart type respectifs des revenus :

	RUSSIE	QUÉBEC
Revenu moyen:	180 000 roubles	264 000 dollars
Écart type du revenu:	7 200 roubles	33 520 dollars
Coefficient de variation:	_____	_____

À quel endroit observons-nous :

- la plus grande disparité des revenus ? _____
- la distribution des revenus la plus homogène ? _____

1.5.4 L'utilisation du mode statistique de la calculatrice

Lorsqu'une série compte un grand nombre de données, le calcul de la moyenne et de l'écart type peut s'avérer laborieux, surtout si ces données ne sont pas groupées par valeurs ou par classes. C'est pourquoi nous n'hésiterons pas à utiliser la calculatrice pour faciliter le travail.

Nous présentons dans cette section la façon d'utiliser le mode statistique de la calculatrice scientifique de base (modèle Sharp EL-531W) et de la calculatrice graphique (modèle TI-84 Plus). Même si le modèle de votre calculatrice est différent des modèles mentionnés, les procédures présentées devraient convenir. Si ce n'est pas le cas, essayez les autres procédures suggérées (entre parenthèses dans le texte) ou consultez le guide d'utilisation de votre calculatrice.

EXERCICES DE COMPRÉHENSION | 1.11

Des deux guides d'utilisation présentés ci-dessous et à la page suivante, choisir celui qui s'applique au modèle de votre calculatrice, puis l'utiliser pour calculer la moyenne, l'écart type et l'écart type corrigé des trois distributions suivantes. S'assurer d'obtenir les réponses données.

1. Données brutes : 21,8 26,8 32,5 28,4

Réponses attendues : $\bar{x} = 27,4$ $\sigma = 3,8$ $s = 4,4$

2. Distribution avec des effectifs :

Valeur	2	8	12	Total
Effectif	4	3	7	14

Réponses attendues : $\bar{x} = 8,3$ $\sigma = 4,3$ $s = 4,4$

3. Distribution avec des classes et des pourcentages :

Classe	[2; 4[[4; 6[[6; 8[Total
Pourcentage	25,7 %	44,3 %	30,0 %	100,0 %

Réponses attendues : $\bar{x} = 5,1$ $\sigma = 1,5$ $s = \text{aucune valeur ou erreur}^6$

Guide d'utilisation de la calculatrice scientifique de base

Modèle Sharp EL-531W (ou modèle équivalent)

POUR ACCÉDER AU MODE STATISTIQUE

Appuyer sur **MODE**, sélectionner l'option **STAT** en appuyant sur **1**, puis l'option **SD** (statistique descriptive) en appuyant sur **0**.

(Pour certaines calculatrices, on appuie sur **MODE**, puis on sélectionne l'option **SD** ou **STAT_x**. Si un sous-menu s'affiche, on choisit l'option permettant d'effectuer des calculs à une variable.)

6. Le nombre total de données étant inconnu, il est impossible de calculer l'écart type corrigé. Si vous obtenez une valeur, c'est sans doute parce que vous avez saisi les pourcentages tels quels, sans les exprimer sous forme décimale : la calculatrice considère alors que vous avez un total de 100 données, ce qui n'est pas le cas.

POUR ENTRER LES DONNÉES

Données brutes

- Saisir la 1^{re} valeur, puis appuyer sur **M+** (**data**) pour l'entrer en mémoire.
- Faire de même pour les autres valeurs.

On peut vérifier à tout moment le nombre de données entrées en appuyant sur **RCL**, puis sur **0** (**n**).

(Certaines calculatrices affichent le nombre de données entrées, au fur et à mesure.)

Données groupées par valeurs ou par classes

- Saisir la 1^{re} valeur (ou le centre de classe), puis appuyer sur **STO** (**x, y**).
- Saisir l'effectif (ou la fréquence relative) associé à la valeur puis appuyer sur **M+** (**data**) pour entrer le tout en mémoire.
- Faire de même pour les autres valeurs.

On peut vérifier à tout moment le nombre de données entrées en appuyant sur **RCL**, puis sur **0** (**n**).

POUR AFFICHER LES MESURES

Moyenne : appuyer sur **RCL**, puis sur **4** (**\bar{x}**).

(Autre façon : **2nd F**, puis **\bar{x}** .)

Écart type : appuyer sur **RCL**, puis sur **6** (**σx**).

(Autre façon : **2nd F**, puis **σ** ou **σn** ou **$x \sigma n$** .)

Écart type corrigé : appuyer sur **RCL**, puis sur **5** (**sx**).

(Autre façon : **2nd F**, puis **s** ou **$sn - 1$** ou **$xsn - 1$** .)

POUR VIDER LA MÉMOIRE DE LA CALCULATRICE

Avant chaque nouveau problème, on vide la mémoire de la calculatrice en appuyant sur **2nd F**, puis sur **MODE** (**CA**). (Pour certaines calculatrices, on appuie sur **2nd F** puis sur **DEL**.)

Guide d'utilisation de la calculatrice graphique

Modèle TI-84 Plus (ou modèle équivalent)

POUR ACCÉDER AU MODE STATISTIQUE ET EFFACER LES DONNÉES EN MÉMOIRE

- Appuyer sur **STAT**.
- Placer le curseur sur le menu **EDIT**, puis sur **1: EDIT** et appuyer sur **ENTER**.
- Placer le curseur sur le titre de colonne **L1**, appuyer sur **CLEAR**, puis sur **ENTER** pour effacer le contenu de la colonne.
- Reprendre la procédure précédente pour effacer le contenu de la 2^e colonne.

POUR ENTRER LES DONNÉES

Données brutes

- Saisir la 1^{re} valeur dans la colonne **L1**, puis appuyer sur **ENTER**.
- Faire de même pour les autres valeurs.

Données groupées par valeurs ou par classes

- Saisir la 1^{re} valeur, dans la colonne **L1**, puis appuyer sur **ENTER**. Faire de même pour les autres valeurs.
- Saisir le 1^{er} effectif (ou la 1^{re} fréquence relative) dans la colonne **L2**, puis appuyer sur **ENTER**. Faire de même pour les autres effectifs.

POUR AFFICHER LES MESURES

- Appuyer sur **STAT**, sélectionner le menu **CALC**, puis **1: 1-VAR STATS** et appuyer sur **ENTER**.
- Appuyer sur **2nd**, sur **1** (**L1**), puis sur **ENTER**.

La moyenne **\bar{x}** , l'écart type **σx** , l'écart type corrigé **sx** et le nombre de données **n** s'affichent.

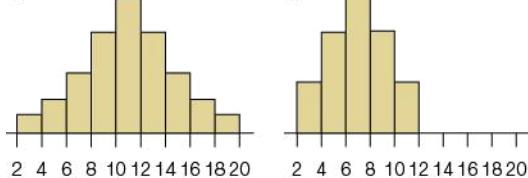
- Appuyer sur **STAT**, sélectionner le menu **CALC**, puis **1: 1-VAR STATS** et appuyer sur **ENTER**.
- Appuyer sur **2nd**, sur **1** (**L1**), sur **,**, sur **2nd**, sur **2** (**L2**), et enfin sur **ENTER**.

La moyenne **\bar{x}** , l'écart type **σx** , l'écart type corrigé **sx** et le nombre de données **n** s'affichent.

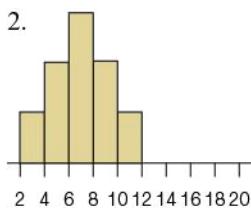
EXERCICES 1.4

1. a) Estimer graphiquement la moyenne de chacun des histogrammes suivants.

1.



2.



- b) Lequel de ces histogrammes a le plus grand écart type ?
- c) Pour l'histogramme 2, utiliser un rapport de surfaces pour déterminer le pourcentage de données dans la classe [6 ; 8].
2. Voici des statistiques sur les matchs de hockey joués par les Canadiens de Montréal en 2013-2014.

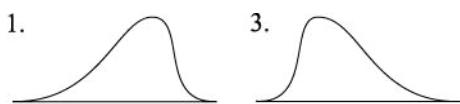
Répartition des matchs joués par les Canadiens de Montréal selon le nombre de buts comptés, saison régulière, 2013-2014

Nombre de buts	Nombre de matchs
0	6
1	22
2	15
3	15
4	13
5	7
6	3
7	1
Total	82

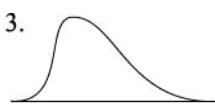
Source: Le site officiel des Canadiens de Montréal. 2014.

- a) Donner le nom et le type de la variable.
- b) Donner le mode et la médiane de la distribution et interpréter ces deux mesures.
- c) Calculer la moyenne et l'écart type de la distribution, et interpréter ces deux mesures.
3. a) Indiquer par un pivot la position de la moyenne sur chacun des graphiques suivants.

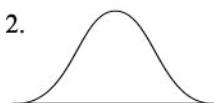
1.



3.



2.



- b) Laquelle des affirmations suivantes est vraie ? L'écart type du graphique 1 est :
- i) plus petit que celui du graphique 2.

ii) égal à celui du graphique 3.

iii) plus grand que celui du graphique 3.

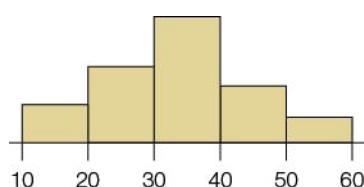
4. a) Soit A, la série des 365 températures quotidiennes à Montréal en 2013, et B, celle des 365 températures quotidiennes à Miami la même année. D'après vous, laquelle des deux séries présente le plus grand écart type ?

- b) Selon Statistique Canada, le revenu personnel médian des Québécois en 2011 est de 27 400 \$ alors que le revenu personnel moyen est de 35 900 \$. Qu'est-ce qui peut expliquer un tel écart entre ces deux mesures ? Quelle mesure est-il préférable d'utiliser pour cette variable ?

Source: Statistique Canada. Tableau 202-0402, CANSIM, juin 2013.

- c) Parmi les trois nombres suivants, lesquels ne peuvent être l'écart type de l'histogramme ?

11,2 28,1 1,2



5. Dans chacun des cas suivants, dire si la mesure de tendance centrale donnée est possible compte tenu des restrictions imposées. Si non, dire pourquoi. Si oui, donner un exemple de résultats satisfaisant à toutes ces conditions. La représentation de la situation à l'aide d'un pictogramme facilite la recherche de la solution.

- a) Il y a cinq données ; la plus petite donnée de la série est 4 ; l'étendue est 10 ; $\mu = 14$.
- b) Il y a cinq données ; la plus petite donnée de la série est 4 ; l'étendue est 10 ; la médiane égale 14.
- c) Il y a cinq données ; la plus petite donnée de la série est 50 ; la plus grande est 100 ; $\mu = 55$.

6. Pour un échantillon de six années, de 2006 à 2011, on a compilé le nombre d'accidents avec dommages corporels impliquant une motoneige au Québec.

293 333 381 338 293 306

Source: Société de l'assurance automobile du Québec. Dossier statistique – Bilan 2011, accidents, parc automobile, permis de conduire, juin 2012.

- a) Donner l'étendue de la série statistique.
- b) Calculer et interpréter la moyenne et l'écart type corrigé de l'échantillon.
- c) Donner et interpréter la médiane.

7. Soit la distribution suivante.

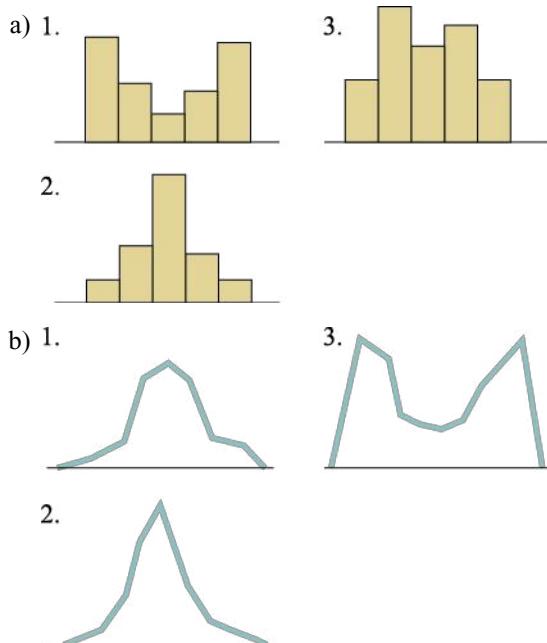
Répartition des professeurs de cégep selon l'âge, Québec, 2009-2010

Âge (en ans)	Pourcentage de professeurs
Moins de 30	10 %
$30 \leq X < 40$	27 %
$40 \leq X < 50$	27 %
$50 \leq X < 60$	28 %
60 et plus	8 %
Total	100 %

Source: Ministère de l'Éducation, du Loisir et du Sport, et Ministère de l'Enseignement supérieur. *Statistiques de l'éducation, édition 2011, 2013.*

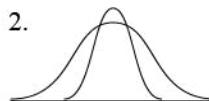
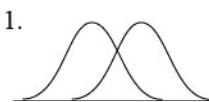
- a) Calculer et interpréter la moyenne et l'écart type de la distribution.
- b) La distribution est-elle homogène? Justifier la réponse.
- c) Déterminer et interpréter la médiane de la distribution.

8. Ordonner les graphiques suivants selon l'écart type, du plus petit au plus grand.



9. Parmi les deux graphiques suivants :

- a) lequel présente deux distributions ayant la même moyenne ?
- b) lequel présente deux distributions ayant le même écart type ?



10. a) Intuitivement, laquelle des deux séries statistiques suivantes vous paraît la plus homogène ? Pourquoi ?

Série A : 2 4 6 8 10

Série B : 102 104 106 108 110

- b) Mesurer mathématiquement l'homogénéité de ces deux séries.

11. Il y a une trentaine d'années, c'était surtout des jeunes qui achetaient des motocyclettes. Aujourd'hui, en raison des coûts d'achat et d'assurances de ces véhicules, les choses ont changé comme en témoignent les statistiques suivantes.

Répartition des propriétaires d'une motocyclette selon l'âge, Québec, 2011

Âge du propriétaire	Pourcentage
Moins de 25 ans	2,8 %
[25 ans; 35 ans[10,7 %
[35 ans; 45 ans[22,5 %
[45 ans; 55 ans[37,2 %
55 ans et plus	26,8 %
Total	100,0 %

Source: Société de l'assurance automobile du Québec. *Dossier statistique – Bilan 2011, accidents, parc automobile, permis de conduire, juin 2012.*

- a) Indiquer et interpréter la classe modale.
- b) Calculer la moyenne et l'écart type de la distribution et interpréter ces deux mesures.
- c) Les mesures calculées en b) sont-elles exactes ou approximatives ?
- d) Compléter l'énoncé. À partir du tableau, on peut estimer qu'au Québec, en 2011, 50 % des propriétaires d'une motocyclette ont plus de _____ ans, alors que les moins de _____ ans comptent pour seulement 10 % des propriétaires.

1.6 Les mesures de position (cote z)

La cote z permet de situer une donnée par rapport aux autres données d'une série statistique. La mise en situation suivante va nous aider à comprendre cette notion.

MISE EN

SITUATION

Un employeur désire engager un étudiant pour l'été afin qu'il l'aide à terminer une étude de marché. Comme le travail requiert des connaissances en statistique, il décide de choisir la personne la plus performante parmi les quatre candidats qui ont suivi un cours de statistique. Voici le dossier scolaire des candidats.

Candidat	Note	Moyenne du groupe	Écart type du groupe
Loïc	85	75	10
Zoé	76	70	3
Alexia	70	60	4
William	75	80	5

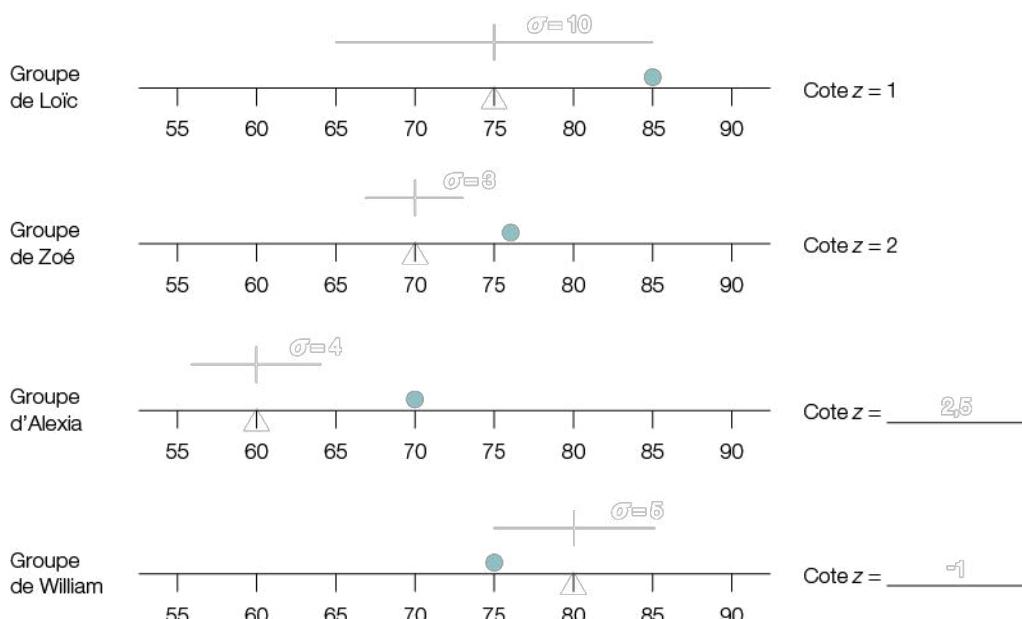
Si l'employeur se fie à la note seulement, quel devrait être son choix ? _____

S'il considère en plus la moyenne du groupe ? _____

Et s'il tient compte de la note, de la moyenne et de l'écart type, qui devrait-il embaucher ? _____

Est-ce le bon choix ? Pourquoi ?

Pour mieux comprendre, représentons sur les diagrammes suivants la moyenne du groupe par un pivot et l'écart type par un segment de droite tracé de part et d'autre de la moyenne. La note de l'étudiant ou de l'étudiante est indiquée par un point. Pour ne pas alourdir la lecture du graphique, nous ne représentons pas les notes des autres étudiants du groupe.



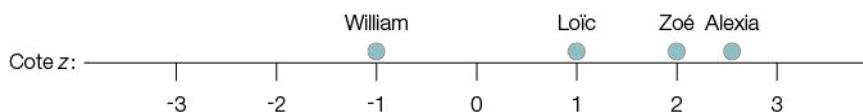
Position de chaque note par rapport aux autres notes du groupe :

- Comme la note de Loïc se situe à 1 écart type au-dessus de la moyenne de son groupe, on dit que sa cote z est 1.
- Comme la note de Zoé se situe à 2 écarts types au-dessus de la moyenne de son groupe, on dit que sa cote z est 2.

La **cote z** donne la mesure, en nombre d'écarts types, de l'écart entre une valeur et la moyenne.

 Trouver la cote z pour les notes d'Alexia et de William et l'inscrire à droite du pictogramme.

La cote z permet de comparer les notes des quatre étudiants en les ramenant sur une même échelle, celle des cotes z .



De ce qui précède, nous pouvons déduire la formule permettant de calculer la cote z d'une donnée.

Cote z

$$\text{Cote } z = \frac{\text{valeur} - \text{moyenne}}{\text{écart type}}$$

Valeurs possibles pour une cote z

La cote z est particulièrement utile pour comparer des résultats de nature différente. Il est important de savoir qu'une cote z plus grande que 2 ou plus petite que -2 est assez rare, et qu'une cote z plus grande que 3 ou plus petite que -3 est rare. C'est pour cette raison que les valeurs indiquées sur une échelle de cote z sont généralement comprises entre -3 et +3. En fait, on a établi que, dans une série de données, au maximum :

12,5 % des données ont une cote $z \geq 2$	12,5 % des données ont une cote $z \leq -2$
8,0 % des données ont une cote $z \geq 2,5$	8,0 % des données ont une cote $z \leq -2,5$
5,5 % des données ont une cote $z \geq 3$	5,5 % des données ont une cote $z \leq -3$
4,1 % des données ont une cote $z \geq 3,5$	4,1 % des données ont une cote $z \leq -3,5$

NOTE

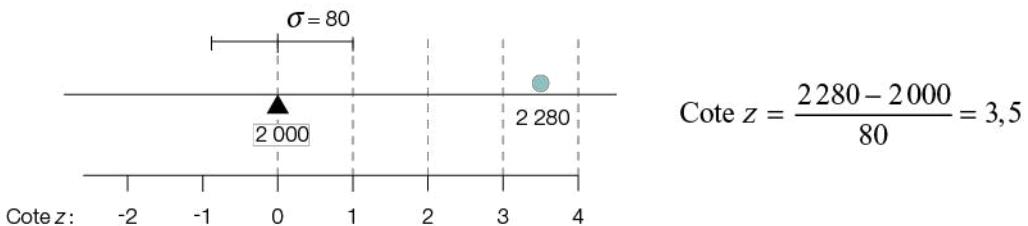
Nous verrons au chapitre 3 que, si le polygone de fréquences d'une distribution a la forme d'une cloche, alors le pourcentage de données ayant une cote z particulière est bien inférieur à la valeur indiquée dans le tableau. Par exemple, seulement 2,3 % des données ont une cote z plus grande que 2, et seulement 0,1 % ont une cote z plus grande que 3.

EXEMPLE

Chaque semaine, un géant de l'alimentation publie une circulaire annonçant les soldes du jeudi pour toutes ses épiceries. Le gérant de l'une de ces épiceries décide un jour d'en faire un peu plus en plaçant une annonce dans le journal local. Le jeudi suivant la parution de l'annonce, il reçoit 2 280 clients alors qu'habituellement le jeudi la moyenne est de 2 000 clients avec un écart type de 80 clients. Peut-il en conclure que son annonce dans le journal local a eu de l'effet ? Un écart de 280 clients par rapport à la moyenne est-il significatif ?

Solution

Effectuons une analyse graphique de la situation en plaçant la moyenne, l'écart type et le résultat obtenu sur un pictogramme.

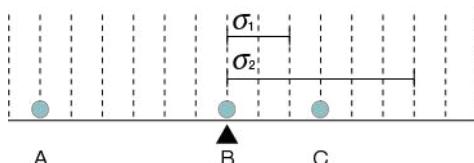


Avec 2 280 clients, on obtient une cote z de 3,5 ; cet écart de 280 clients par rapport à la moyenne est exceptionnel. On peut donc affirmer sans grand risque d'erreur que la publicité supplémentaire a entraîné cette augmentation remarquable de clientèle.

On peut même affirmer que, normalement, en se fiant aux valeurs possibles pour une cote z , un tel résultat ne se produirait au maximum que 4 fois sur 100 : sur 100 jeudis on aurait 2 280 clients ou plus au maximum 4 fois.

Si la distribution du nombre de clients a la forme d'une cloche, un tel résultat ne se produit que 2 fois sur 10 000.

EXERCICES DE COMPRÉHENSION | 1.12



Solution

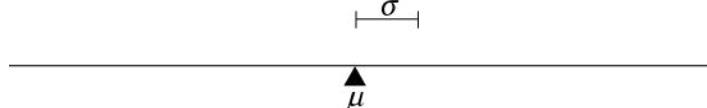
- a) Avec l'écart type σ_1
Cote z de A = _____
Cote z de B = _____
Cote z de C = _____

- b) Avec l'écart type σ_2

Cote z de A = _____

Cote z de B = _____

Cote z de C = _____

- 2. Vrai ou faux ? Si votre cote z à un examen est 2, alors :
- vous avez deux points au-dessus de la moyenne. _____
 - votre note est égale à deux fois l'écart type du groupe. _____
 - votre note se situe à deux écarts types au-dessus de la moyenne. _____
 - votre note est égale à la moyenne de l'examen plus deux fois l'écart type. _____
3. Une cote z de 2 à un examen peut-elle être considérée comme faible, moyenne, très bonne ou exceptionnelle ? _____
4. a) À l'aide des cotes z des points A, B et C données ci-dessous, situer ces points sur le pictogramme.
- Cote z de A = 2 Cote z de B = -1 Cote z de C = -1,5
- 
- b) Calculer l'écart entre chaque point du pictogramme et la moyenne si l'écart type est de 10.
- Écart entre A et μ = _____
- Écart entre B et μ = _____
- Écart entre C et μ = _____
- c) En considérant les écarts calculés en b), déterminer la valeur de A, de B et de C si la moyenne est 50.
- A = _____ B = _____ C = _____

5. Nous avons souligné que la cote z permet de comparer des données même si celles-ci proviennent de domaines bien différents. En voici un exemple.

On veut trouver le meilleur vendeur du mois. Cet honneur sera décerné à la personne s'étant le plus distinguée dans son domaine. Trois concurrents sont en lice ; voici la description de la performance de chacun des candidats, en un mois :

- Mia a vendu 85 tablettes de chocolat pour les activités sportives de son école, alors que la moyenne est de 52 tablettes par étudiant avec un écart type de 13 tablettes.
- Thomas a vendu 25 automobiles, alors que la moyenne de ventes est de 12 automobiles avec un écart type de 6.
- Lucie a vendu 75 abonnements au *Journal de Québec*, alors que la moyenne de ventes est de 47 abonnements avec un écart type de 10.

Qui devrait être nommé « meilleur vendeur du mois » ? Justifier ce choix.

Solution

Cote de rendement au collégial

Jusqu'en 1995, on utilisait la cote z pour classer les étudiants de cégep qui désiraient s'inscrire à un programme universitaire contingenté. La sélection était alors équitable seulement pour les étudiants qui appartenaient à des groupes comparables au cégep : ceux qui provenaient de groupes homogènes constitués d'étudiants très forts ne pouvaient obtenir une cote z aussi élevée que s'ils avaient fait partie d'un groupe plus hétérogène. On a donc décidé d'ajuster la cote z au moyen d'un indicateur de la force du groupe de l'étudiant. Cela a donné naissance à la cote de rendement au collégial (CRC) ou cote r , que l'on calcule de la façon suivante :

$$\text{CRC} = (\text{cote } z \text{ de l'étudiant} + \text{indicateur de la force du groupe} + 5) \times 5$$

Cote z de l'étudiant

On calcule cette cote z ainsi : Cote $z = \frac{\text{note} - \mu_G}{\sigma_G}$

- où :
- Note désigne la note de l'étudiant pour un cours donné ;
 - μ_G est la moyenne, pour le groupe, des notes supérieures ou égales à 50 ;
 - σ_G est l'écart type, pour le groupe, des notes supérieures ou égales à 50.

Indicateur de la force du groupe (IFG)

On calcule l'IFG ainsi : $\text{IFG} = \frac{\text{MS}_G - 75}{14}$

- où :
- MS désigne la moyenne pondérée des notes de 4^e et 5^e secondaire d'un étudiant ;
 - MS_G est la moyenne des MS des n étudiants du groupe :
- $$\text{MS}_G = \frac{\text{MS}_1 + \text{MS}_2 + \text{MS}_3 + \dots + \text{MS}_n}{n}$$
- 75 est considéré comme la moyenne provinciale des MS des étudiants acceptés au cégep ;
 - 14 est considéré comme l'écart type provincial des MS des étudiants acceptés au cégep.

Rôle des constantes

L'addition de la constante 5 vise à éliminer les valeurs négatives de la CRC.

La multiplication par 5 permet de déterminer l'amplitude de l'échelle de la CRC.

EXEMPLE

Cote de rendement au collégial selon la cote z de l'étudiant et la force du groupe

Force du groupe selon MS_G	Cote z	Indicateur de la force du groupe (IFG)	Cote de rendement au collégial (CRC)
Fort: $\text{MS}_G = 92$	1,5	$\text{IFG} = \frac{92 - 75}{14} = 1,2$	$\text{CRC} = (1,5 + 1,2 + 5) \times 5 = 38,5$
	0		$\text{CRC} = (0 + 1,2 + 5) \times 5 = 31$
	-1,5		$\text{CRC} = (-1,5 + 1,2 + 5) \times 5 = 23,5$
Moyen: $\text{MS}_G = 75$	1,5	$\text{IFG} = \frac{75 - 75}{14} = 0$	$\text{CRC} = (1,5 + 0 + 5) \times 5 = 32,5$
	0		$\text{CRC} = (0 + 0 + 5) \times 5 = 25$
	-1,5		$\text{CRC} = (-1,5 + 0 + 5) \times 5 = 17,5$
Faible: $\text{MS}_G = 65$	1,5	$\text{IFG} = \frac{65 - 75}{14} = -0,7$	$\text{CRC} = (1,5 - 0,7 + 5) \times 5 = 29$
	0		$\text{CRC} = (0 - 0,7 + 5) \times 5 = 21,5$
	-1,5		$\text{CRC} = (-1,5 - 0,7 + 5) \times 5 = 14$

EXERCICES 1.5

1. On a demandé à un laboratoire spécialisé en contrôle de la qualité d'évaluer le mélange bitumineux fabriqué par deux usines ayant répondu à un appel d'offres. Le responsable de l'étude décide de prélever un échantillon de 50 cylindres de bitume dans la production de chaque usine et de mesurer la résistance à la compression de celui-ci. Les résultats apparaissent dans les tableaux suivants.

Usine A

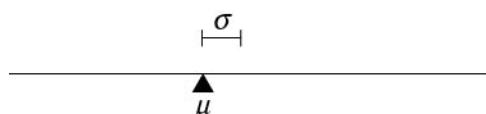
Résistance à la compression (en kg/cm ²)	Nombre de cylindres
[70; 75[2
[75; 80[4
[80; 85[7
[85; 90[12
[90; 95[11
[95; 100[11
[100; 105[3
Total	50

Usine B

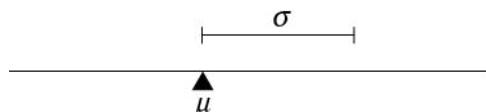
Résistance à la compression (en kg/cm ²)	Nombre de cylindres
[70; 80[4
[80; 90[7
[90; 100[19
[100; 110[12
[110; 120[6
[120; 130[2
Total	50

- a) Calculer la moyenne et l'écart type corrigé de l'échantillon de chaque usine.
 b) Le contrat sera attribué à l'usine qui produit le béton le plus homogène. À quelle usine sera-t-il accordé ?
 2. a) Sur chacun des graphiques suivants, placer un point ayant la cote z indiquée.

i) Cote $z = -2,5$

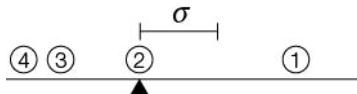


ii) Cote $z = 0,5$



- b) Dans chaque cas, donner l'écart entre le point et la moyenne si l'écart type est de 8 dans le graphique i) et de 20 dans le graphique ii).

3. On a représenté ci-dessous la note à un examen (sur 100 points) de quatre étudiants d'un groupe.



- a) Quelle cote z chacun de ces étudiants a-t-il à son examen ?

- b) Quel est l'écart entre chaque note et la moyenne du groupe si l'écart type est de 10 points ?

- c) Donner la note obtenue dans chaque cas si la moyenne de l'examen est de 65 points.

4. Deux étudiantes appartenant à deux groupes différents ont eu la même note à un examen, mais la cote z d'Élodie est plus grande que celle de Jade. Elles ont toutes deux une note supérieure à la moyenne.

- a) Si la moyenne est la même pour les deux groupes, pour lequel l'écart type est-il le plus petit ?

- b) Si l'écart type est le même pour les deux groupes, pour lequel la moyenne est-elle la plus faible ?

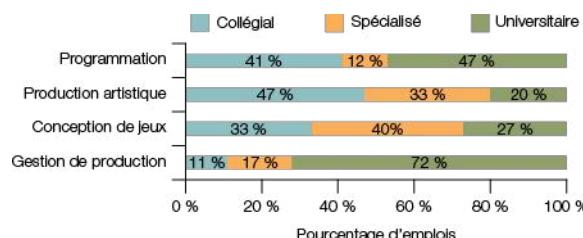
5. a) Supposons que les enseignants québécois gagnent en moyenne 50 377 \$ par année avec un écart type de 4 789 \$ et que les enseignants français gagnent en moyenne 35 244 € (euros) avec un écart type de 4 977 €. Calculer le coefficient de variation pour chaque groupe d'enseignants et interpréter les résultats.

- b) Un professeur enseigne à trois groupes qui ont obtenu les moyennes suivantes sur 100 points : le groupe A, qui compte 33 étudiants, a une moyenne de 69 ; le groupe B, qui compte 25 étudiants, a une moyenne de 74 ; le groupe C, qui compte 22 étudiants, a une moyenne de 80. Calculer la moyenne pour l'ensemble de tous les étudiants.

6. Un optométriste vous a informé que votre pression intra-oculaire est de 23 mm de mercure. Pour une population de 100 000 personnes de votre âge, la pression moyenne est de 17 mm de mercure avec un écart type de 2,4 mm. Combien, au maximum, y a-t-il de personnes dans la population dont la pression est au moins aussi éloignée de la moyenne que la vôtre ? (Utiliser le tableau sur les valeurs possibles pour une cote z à la page 75.)

7. Quel type de diplôme faut-il pour travailler dans l'industrie du jeu vidéo ? Voici des statistiques à ce sujet.

Répartition des emplois dans l'industrie du jeu vidéo, par secteur, selon le type de diplôme, Québec, 2012



Source: TECHNOCompétences. *L'emploi dans l'industrie du jeu électronique au Québec en 2012*, avril 2013.

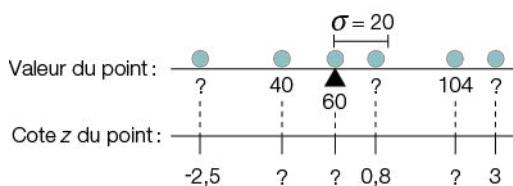
- a) Quel secteur de l'industrie du jeu vidéo compte la plus grande proportion :
- de diplômés universitaires ?
 - de détenteurs d'un DEC ?
 - de détenteurs d'un diplôme spécialisé ?
- b) Compléter l'énoncé. Les diplômés du collégial détiennent _____ % des emplois du secteur de la conception de jeux et _____ % du secteur de la programmation.
8. Un commerçant pense que ce sont des travaux effectués par la municipalité qui ont entraîné une diminution de l'achalandage et, du même

coup, une baisse de ses recettes. Dans la plainte qu'il adresse aux autorités municipales, il précise que ses recettes sont en moyenne de 20 000 \$ par jour et que, la journée où les travaux ont été effectués, elles n'ont été que de 19 500 \$. La Ville réplique que, dans le cas de recettes moyennes de 20 000 \$, un écart de 500 \$ est trop petit pour être significatif. Selon le commerçant, l'écart type de ses recettes quotidiennes est de 100 \$. Qui a raison ?

9. La moyenne des notes à un examen sur 100 points est de 60 et l'écart type, de 10. La distribution des notes a la forme d'une cloche.

- Comment doit-on interpréter l'écart type ?
- La cote z de la note de Lucie est de 1,5. Comment faut-il interpréter cette mesure ?
- Combien de points Lucie a-t-elle de plus que la moyenne ?
- Quelle note Lucie a-t-elle obtenue à l'examen ?

10. À l'aide de l'information donnée pour chacun des points du pictogramme ci-dessous, déterminer, selon le cas, la valeur ou la cote z de chaque point du graphique.



Traitement de données (une variable)

Pour présenter et analyser les données d'une variable, on applique la procédure suivante :

1. On détermine le type de la variable.
2. On construit le tableau et le graphique appropriés au type de variable étudié, puis on effectue une première analyse des données.
3. On complète l'analyse des données en calculant et en interprétant les mesures de tendance centrale et de dispersion pertinentes.

Types de variables, de tableaux, de graphiques et de mesures

Types de variables et échelles de mesure	Tableaux et graphiques ⁷	Mesures possibles
Qualitative (non numérique) <ul style="list-style-type: none"> • nominale (absence de relation d'ordre); échelle nominale • ordinale (existence d'une relation d'ordre); échelle ordinale 	Tableau de distribution Diagramme à rectangles (verticaux ou horizontaux) Diagramme circulaire Diagramme linéaire	Mode
Quantitative (numérique) <ul style="list-style-type: none"> • discrète (on ne peut pas augmenter la précision); échelle ordinale, d'intervalle ou de rapport • continue (on peut augmenter la précision); échelle ordinale, d'intervalle ou de rapport 	Tableau de distribution Diagramme en bâtons Tableau de distribution Histogramme Polygone de fréquences Ogive (courbe de fréquences cumulées)	Moyenne Mode Médiane Quantiles Cote z Écart type Coefficient de variation

Définition, calcul et interprétation des mesures

Mesures de tendance centrale

Mesure	Définition et calcul	Interprétation
Moyenne	Centre d'équilibre du graphique de la distribution. Calcul: mode statistique de la calculatrice.	En moyenne...
Moyenne pondérée	Donne la moyenne de valeurs qui n'ont pas le même poids. Calcul: somme des produits de chaque valeur par sa pondération.	
Mode	Valeur, catégorie ou classe ayant la plus grande fréquence.	Une pluralité... (%)... (mode)...

7. Pour les caractéristiques des graphiques, voir «Quel graphique faut-il construire?» à la page 34.

Mesure	Définition et calcul	Interprétation
Médiane	<p>Divise une série de données en deux parties égales.</p> <p><u>Données groupées en classes</u> Valeur sur l'axe horizontal qui partage la surface de l'histogramme en deux parties égales.</p> <p><u>Données non groupées en classes</u></p> <ul style="list-style-type: none"> • Avec les effectifs Si le nombre total de données est : <ul style="list-style-type: none"> – pair : moyenne des deux données centrales ; – impair : valeur de la donnée centrale. • Avec les pourcentages On repère la valeur qui donne un cumul d'au moins 50 % des données. La médiane est : <ul style="list-style-type: none"> – la valeur repérée, si le cumul est supérieur à 50 % ; – la moyenne de la valeur repérée et de la suivante, si le cumul est égal à 50 %. 	<p><u>Données groupées en classes</u> On peut estimer que 50 %... ont moins de (<i>médiane</i>).</p> <p><u>Données non groupées en classes</u> Selon le cas : <ul style="list-style-type: none"> • 50 %... ont (<i>médiane</i>) ou moins. • Au moins 50 %... ont (<i>médiane</i>) ou moins. </p>

Mesures de position

Mesure	Définition et calcul	Interprétation
Quantiles	<p>Valeurs qui partagent une série de données en un certain nombre de parties égales.</p> <p>Les centiles (C_i) divisent la série de données en 100 parties égales, les déciles (D_i) la divisent en 10 parties, les quintiles (V_i) la divisent en 5 parties et les quartiles (Q_i) la divisent en 4 parties.</p> <p><u>Données groupées en classes</u> Pour trouver le centile C_i, on détermine, sur l'axe horizontal de l'histogramme, la valeur qui laisse à sa gauche i % de la surface totale.</p>	<p><u>Données groupées en classes</u> On peut estimer que i %... ont moins de (<i>centile</i>).</p>
Cote z	<p>Mesure, en nombre d'écart types, l'écart entre une valeur et la moyenne de la distribution :</p> $\text{Cote } z = \frac{\text{valeur} - \text{moyenne}}{\text{écart type}}$	<p>La valeur se situe à (<i>cote z</i>) écart(s) type(s) de la moyenne.</p>

Mesures de dispersion

Mesure	Définition et calcul	Interprétation
Écart type	<p>Peut être considéré comme une approximation de la moyenne des distances entre chaque donnée et la moyenne des données.</p> <p>Calcul : mode statistique de la calculatrice.</p>	<p>La plupart des données se situent entre (<i>moyenne – écart type</i>) et (<i>moyenne + écart type</i>).</p>
Coefficient de variation (CV)	<p>Mesure la dispersion relative des données :</p> $CV = \frac{\text{écart type}}{\text{moyenne}} \times 100\%$	<p>Si le CV < 15 %, la distribution est homogène.</p>

EXERCICES RÉCAPITULATIFS

1. Vous rêvez de vous lancer en affaires? Les statistiques suivantes devraient vous intéresser, elles sont tirées d'une étude réalisée en 2013 auprès des propriétaires d'une PME⁸ québécoise.

- a) À la question «Quel âge aviez-vous lorsque vous avez démarré votre entreprise?», on a obtenu les réponses ci-dessous.

Répartition des propriétaires d'une PME selon leur âge au démarrage de l'entreprise, Québec, 2013

Âge au démarrage de l'entreprise	Pourcentage de propriétaires
Moins de 25 ans	16,4 %
[25 ans; 35 ans[35,1 %
[35 ans; 45 ans[28,8 %
[45 ans; 55 ans[16,2 %
55 ans et plus	3,5 %
Total	100,0 %

Source: Fondation de l'entrepreneurship. *Indice entrepreneurial québécois 2013. Les entrepreneurs du Québec font-ils preuve d'audace?*

- i) Calculer et interpréter la moyenne et l'écart type de la distribution.
ii) La distribution est-elle homogène?
iii) Déterminer et interpréter la cote z de l'âge d'une personne qui avait 58 ans au démarrage de son entreprise.
- b) L'étude révèle que 40 % des propriétaires d'une PME dirigent leur entreprise depuis moins de cinq ans. On a demandé à ces derniers combien d'argent ils avaient dû investir pour démarrer leur entreprise et combien d'emplois ils avaient créé, en excluant le leur.

Tableau 1

Répartition des nouveaux propriétaires d'une PME selon le montant investi au démarrage de l'entreprise, Québec, 2013

Montant investi	Pourcentage
Moins de 100 000 \$	84,1 %
[100 000 \$; 500 000 \$[14,9 %
500 000 \$ et plus	1,0 %
Total	100,0 %

Source: Fondation de l'entrepreneurship. *Indice entrepreneurial québécois 2013. Les entrepreneurs du Québec font-ils preuve d'audace?*

8. Petite et moyenne entreprise.

Tableau 2

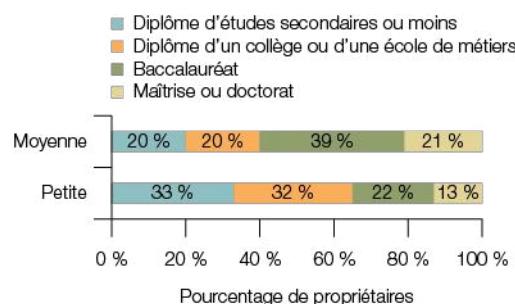
Répartition des nouveaux propriétaires d'une PME selon le nombre d'emplois créés au démarrage de l'entreprise en excluant le leur, Québec, 2013

Nombre d'emplois	Pourcentage
Aucun	59,0 %
De 1 à 3	26,7 %
De 4 à 5	7,4 %
6 et plus	6,9 %
Total	100,0 %

Source: Fondation de l'entrepreneurship. *Indice entrepreneurial québécois 2013. Les entrepreneurs du Québec font-ils preuve d'audace?*

- i) Déterminer et interpréter la médiane de la distribution du tableau 1.
ii) Déterminer et interpréter le mode de la distribution du tableau 2.
c) Quel est le niveau d'instruction des propriétaires d'une PME ?

Répartition des propriétaires d'une PME, par taille d'entreprise, selon le niveau d'instruction le plus élevé atteint, Canada, 2013



Pourcentage de propriétaires

Source: Industrie Canada. *Principales statistiques relatives aux petites entreprises*, août 2013.

Compléter chaque énoncé.

- i) Le pourcentage de propriétaires diplômés d'un collège ou d'une école de métiers est plus élevé dans les petites entreprises que dans les moyennes entreprises, soit _____ % contre _____ %.
ii) Les statistiques révèlent que _____ % des propriétaires d'une moyenne entreprise ont un diplôme universitaire alors que ce pourcentage est de _____ % chez les propriétaires d'une petite entreprise.

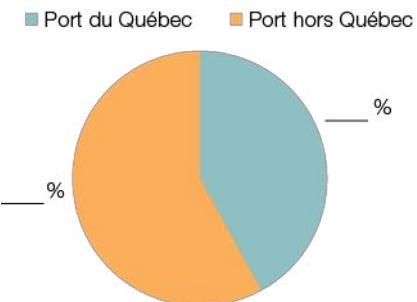
2. **Répartition des divorces selon la durée du mariage, Québec, 2008**

Durée du mariage	Pourcentage
Moins de 5 ans	16,4 %
[5 ans; 10 ans[21,3 %
[10 ans; 15 ans[15,5 %
[15 ans; 25 ans[25,2 %
[25 ans; 45 ans[21,6 %
Total	100,0 %

Source: Institut de la statistique du Québec. 2011.

- a) Vrai ou faux ?
- i) Au Québec, 37,7 % des mariages sont rompus après moins de 10 ans de vie commune.
 - ii) Pour le quart des divorces survenus en 2008, le couple était marié depuis 15 à 25 ans.
- b) On note que les classes du tableau de distribution n'ont pas toutes la même amplitude. Construire l'histogramme qui conviendrait à cette distribution.
- c) Calculer et interpréter le 1^{er} quartile de la distribution.
3. Afin de mesurer l'apport économique du tourisme de croisière sur le Saint-Laurent, un sondage est effectué auprès d'un échantillon de 2 330 croisiéristes. Voici quelques-unes des questions posées.
- Source: Tourisme Québec. *Étude auprès des croisiéristes et des membres d'équipage des navires de croisières dans les ports du Saint-Laurent*, juin 2013.
- Q1. À quel port avez-vous embarqué pour cette croisière?
- 1. Port du Québec 2. Port hors Québec
- Aux personnes qui ont amorcé leur croisière au Québec, on a demandé :
- Q2. Si vous avez séjourné au Québec avant de monter à bord du navire, combien de nuits a duré le séjour? _____
- Q3. En ce qui concerne les procédures d'embarquement sur le navire, diriez-vous que vous êtes...
- 1. Très satisfait 3. Moyennement satisfait
 - 2. Satisfait 4. Insatisfait
- a) Pour chacune des questions posées, indiquer le type de variable et l'échelle de mesure.
- b) Les réponses à la question Q1 indiquent que 979 des 2 330 croisiéristes de l'échantillon ont embarqué à bord du navire dans un port du Québec. Le diagramme suivant présente ces statistiques. Le compléter.

Répartition des croisiéristes de l'échantillon selon le port d'embarquement, Québec, 2013



- c) Voici la distribution des réponses à la question Q2.

Répartition des répondants ayant séjourné au Québec avant de monter à bord du navire selon le nombre de nuuitées du séjour, Québec, 2013

Nombre de nuuitées	Nombre de répondants
1	333
2	186
3	78
4	39
Total	636

- i) Quel type de graphique convient pour représenter la distribution?
- ii) Calculer et interpréter la médiane de la distribution.
- iii) Calculer et interpréter la moyenne de la distribution.

4. a) Si l'on désire grouper en classes 400 données dont la plus petite est 54 et la plus grande 984, quelle devrait être l'amplitude des classes? Justifier le choix de l'amplitude et décrire la 1^{re} classe.

- b) En 2010-2011, les étudiants bénéficiaires d'un prêt et d'une bourse ont reçu les montants suivants : une moyenne de 8 547 \$ pour les étudiants universitaires, de 6 613 \$ pour les étudiants du collégial et de 7 436 \$ pour les étudiants en formation professionnelle au secondaire. Calculer le montant moyen remis à l'ensemble des 95 235 étudiants bénéficiaires d'un prêt et d'une bourse sachant qu'ils se répartissent ainsi : 50 638 sont à l'université, 27 083 sont au collégial et 17 514 sont en formation professionnelle au secondaire.

Source: Ministère de l'Enseignement supérieur. *Statistiques. Aide financière aux études. Rapport 2010-2011, 2013*.

5. INCROYABLE, MAIS VRAI!

Pour augmenter le placement d'annonces publicitaires, le quotidien *Le Grand Journal* a distribué un dépliant promotionnel auprès des entreprises de sa

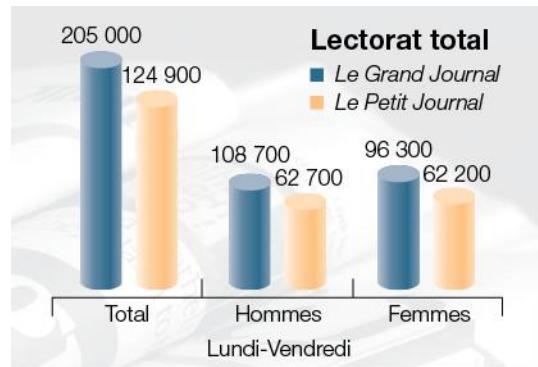
région. Ce dépliant compare *Le Grand Journal* avec son concurrent régional *Le Petit Journal* à l'aide de graphiques. Voici un de ces graphiques :



- La situation décrite et les chiffres donnés sont réels, seuls les noms des journaux ont été changés.

- Pourquoi peut-on dire que ce graphique est trompeur?
- Le graphique précédent amène à penser qu'on a volontairement tenté d'avantagez *Le Grand*

Journal. Une analyse du graphique ci-dessous, extrait du même dépliant, montre qu'il n'en est rien : on fait plutôt face à un manque de connaissances mathématiques de base des concepteurs du dépliant. En effet, si l'on respecte le principe de proportionnalité entre volume et effectif, indiquer le pourcentage de la hauteur des cylindres du *Grand Journal* auquel devrait correspondre la hauteur des cylindres du *Petit Journal*. Faire le calcul pour les catégories «Total», «Hommes» et «Femmes».



PRÉPARATION À L'EXAMEN

Pour préparer votre examen, assurez-vous d'avoir les compétences suivantes :

		Si vous avez la compétence, cochez.
Terminologie et variable		
• À partir de l'énoncé d'une étude, déterminer si c'est un recensement ou un sondage, identifier l'unité statistique et décrire l'échantillon ainsi que la population étudiée.	<input type="radio"/>	
• Déterminer le type d'une variable (qualitative nominale ou ordinaire, quantitative continue ou discrète).	<input type="radio"/>	
• Déterminer l'échelle de mesure (nominale, ordinaire, d'intervalle ou de rapport).	<input type="radio"/>	
Tableau de distribution d'une variable		
• Construire le tableau de distribution d'une variable selon les règles.	<input type="radio"/>	
• Analyser un tableau de distribution.	<input type="radio"/>	
Représentation graphique de la distribution d'une variable		
• Choisir la représentation graphique appropriée à chaque type de variable.	<input type="radio"/>	
• Construire selon les règles et interpréter les données des graphiques suivants:		
– Diagramme à rectangles verticaux ou horizontaux	<input type="radio"/>	
– Diagramme circulaire	<input type="radio"/>	
– Diagramme linéaire	<input type="radio"/>	
– Diagramme en bâtons	<input type="radio"/>	
– Histogramme avec classes égales ou inégales	<input type="radio"/>	
– Polygone de fréquences	<input type="radio"/>	
– Ogive ou courbe de fréquences cumulées	<input type="radio"/>	
• Déterminer le pourcentage de données dans une classe à l'aide d'un rapport de surfaces.	<input type="radio"/>	
Mesures de tendance centrale		
• Calculer une moyenne avec les données brutes, les effectifs ou les pourcentages.	<input type="radio"/>	
• Estimer une moyenne graphiquement.	<input type="radio"/>	
• Reconnaître et calculer une moyenne pondérée.	<input type="radio"/>	
• Trouver et interpréter le mode ou la classe modale d'une distribution.	<input type="radio"/>	
• Calculer et interpréter la médiane d'une distribution.	<input type="radio"/>	
• Choisir la meilleure mesure de tendance centrale dans une situation donnée.	<input type="radio"/>	
Mesures de dispersion		
• Calculer et interpréter l'écart type d'une distribution.	<input type="radio"/>	
• Comparer la dispersion de distributions à partir de leur représentation graphique.	<input type="radio"/>	
• Calculer et interpréter un coefficient de variation.	<input type="radio"/>	
Mesures de position		
• Calculer et interpréter des quantiles (déciles, quintiles, quartiles et centiles).	<input type="radio"/>	
• Cote z:		
– Calculer et interpréter une cote z;	<input type="radio"/>	
– Utiliser la cote z pour comparer des données;	<input type="radio"/>	
– Trouver la cote z ou la valeur d'une donnée à partir d'un pictogramme où la moyenne, l'écart type et la donnée sont représentés.	<input type="radio"/>	

Chapitre 2

Les probabilités



OBJECTIF DU CHAPITRE

Appliquer la théorie des probabilités à la résolution de problèmes concrets.

OBJECTIFS DU LABORATOIRE

Le laboratoire 2 vise à utiliser Excel pour construire un tableau de distribution conditionnelle, calculer des probabilités et étudier la relation entre deux variables à l'aide de la théorie des probabilités.

On prend beaucoup de décisions dans la vie en considérant la probabilité qu'un événement se produise. Parce qu'on estime qu'il y a un risque que notre voiture soit accidentée, que nos biens soient volés ou que notre voyage soit annulé, on contracte des assurances pour se faire indemniser. On participe à une loterie parce qu'on estime avoir des chances de gagner. De même, un entrepreneur se lance en affaires parce qu'il croit que sa probabilité de réussir est élevée.

Le présent chapitre est consacré à l'étude des probabilités. Les concepts abordés sont à la base des lois de probabilité étudiées au chapitre 3.

2.1 Le lien entre les probabilités et l'inférence statistique

Quel lien y a-t-il entre la probabilité et l'inférence statistique, cette branche de la statistique qui permet de généraliser à une population les mesures prises sur un échantillon ?

La mise en situation suivante illustre ce lien en mettant en parallèle les types de questionnements propres à chacune des théories.

MISE EN

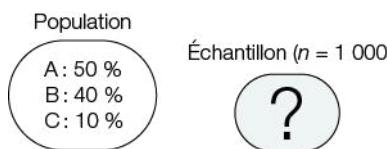
SITUATION

Problématique en probabilité

Situation de départ

On connaît les caractéristiques de la population. Par exemple, lors d'une élection, les votes sont répartis ainsi :

Parti A : 50 % Parti B : 40 % Parti C : 10 %



Type de questionnement

Si l'on prélève au hasard 1 000 personnes parmi celles qui ont voté :

- Les chances que seulement 20 personnes de l'échantillon aient voté pour le parti A sont-elles faibles, moyennes ou fortes ?
- Les chances que 510 personnes de l'échantillon aient voté pour le parti A sont-elles faibles, moyennes ou fortes ?

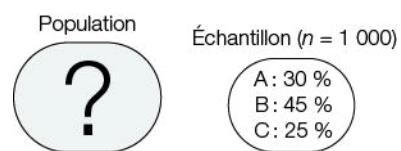
L'étude des probabilités vise à répondre à des questions de ce type et à découvrir les lois de probabilité qui s'appliquent à des expériences aléatoires. Quand on connaît bien ces lois, il est facile de répondre aux questions posées en inférence statistique.

Problématique en inférence statistique

Situation de départ

On connaît les caractéristiques d'un échantillon. Par exemple, un sondage mené auprès d'un échantillon aléatoire de 1 000 électeurs indique les intentions de vote suivantes :

Parti A : 30 % Parti B : 45 % Parti C : 25 %



Type de questionnement

- Sachant que 50 % des électeurs ont voté pour le parti A aux dernières élections, doit-on s'étonner de ce que seulement 30 % des répondants au sondage appuient ce parti ?
Peut-on attribuer l'écart de 20 % au hasard de l'échantillonnage ou doit-on plutôt y voir une baisse des appuis dont jouit le parti A ?
- Peut-on utiliser le résultat de 30 % d'appuis au parti A dans l'échantillon pour prédire, avec une certaine certitude, le pourcentage d'électeurs de la population qui appuient le parti A actuellement ?

2.2 La terminologie

Voici la définition de termes qu'il faut connaître avant d'aborder l'étude des probabilités.

Expérience aléatoire

Toute expérience dont l'issue dépend du hasard est une **expérience aléatoire**. Elle se caractérise par le fait que l'on peut, *a priori*, décrire ou énumérer les résultats possibles de l'expérience. Mais on ne peut prévoir quel résultat sera obtenu une fois l'expérience réalisée.

Espace échantillonnaux

L'**espace échantillonnaux**, représenté symboliquement par la lettre S, est l'ensemble de tous les résultats possibles d'une expérience aléatoire.

Événement

Un **événement** est un sous-ensemble de l'espace échantillonnaux. On le note généralement par une lettre majuscule qui évoque, de préférence, le nom de l'événement.

On peut décrire un événement avec des mots, ce qu'on appelle la **description en compréhension**, ou en énumérant ses éléments, ce qu'on nomme la **description en extension**.

EXEMPLE

Expérience aléatoire	Espace échantillonnaux	Événement
a) Lancer un dé.	$S = \{1, 2, 3, 4, 5, 6\}$ Chaque fois qu'on réalise l'expérience, on obtient un seul des six nombres de S.	Description en compréhension : I: « obtenir un nombre impair » Description en extension : $I = \{1, 3, 5\}$
b) Piger un échantillon de deux personnes parmi Léa, Emma, Mathis et Olivier.	$S = \{\{Léa, Emma\}, \{Léa, Mathis\}, \{Léa, Olivier\}, \{Emma, Mathis\}, \{Emma, Olivier\}, \{Mathis, Olivier\}\}$	G: « piger deux garçons » $G = \{Mathis, Olivier\}$
c) Demander à un électeur, choisi au hasard, pour lequel des partis A, B ou C il a l'intention de voter.	$S = \{\text{parti A, parti B, parti C, indécis, refuse de répondre}\}$	V: « l'électeur exprime son choix de parti » $V = \{\text{parti A, parti B, parti C}\}$
d) Vérifier le taux de compactage d'un échantillon aléatoire de sol.	$S = [0 \% ; 100 \%]$	T: « un taux de compactage d'au moins 90 % » $T = [90 \% ; 100 \%]$

2.3 Le calcul d'une probabilité

Quand on réalise une expérience aléatoire, on veut généralement connaître les chances qu'un événement donné se produise : c'est ce que l'on appelle la **probabilité** d'un événement. Voici deux façons d'aborder le calcul d'une probabilité.

Probabilité classique

Si l'on considère *a priori* que les éléments de l'espace échantillonnaux associé à un événement A sont **équiprobables** (c'est-à-dire que les éléments de S ont tous la même chance de se produire), alors la probabilité de A, notée P(A), se calcule comme suit.

Probabilité classique

$$P(A) = \frac{\text{nombre de résultats favorables à } A}{\text{nombre de résultats possibles}} = \frac{n(A)}{n(S)}$$

EXEMPLE 1

Une expérience aléatoire consiste à piger une carte dans un jeu de 52 cartes à jouer. On s'intéresse aux deux événements suivants :

A : «piger un as» N : «piger une carte noire»

Calculer la probabilité de l'événement A et de l'événement N.

Solution

S =

P(A) =

P(N) =

EXEMPLE 2

Une expérience aléatoire consiste à piger un échantillon de deux personnes parmi les quatre personnes suivantes : Léa, Emma, Mathis et Olivier. Quelle est la probabilité de piger deux garçons ?

Solution

S = {{Léa, Emma}, {Léa, Mathis}, {Léa, Olivier}, {Emma, Mathis}, {Emma, Olivier}, {Mathis, Olivier}}

G =

P(G) =

Probabilité empirique (ou fréquentiste)

MISE EN SITUATION

Quelle est la probabilité qu'un détenteur d'un DEC en formation technique décroche un emploi à la fin de ses études ? Peut-on déterminer cette probabilité avec la formule de probabilité classique ? Effectuer le calcul pour vérifier.

Solution

Intuitivement, on perçoit que cette probabilité n'a pas de sens. En fait, on ne peut pas appliquer la définition classique, car les éléments de S ne sont pas équiprobables.

Lorsque les éléments d'un espace échantillonnal ne sont pas équiprobables, on ne peut pas calculer la probabilité d'un événement en utilisant la définition classique ; il faut recourir à la définition empirique d'une probabilité.

La **probabilité empirique** d'un événement, contrairement à la probabilité classique, est fondée sur l'observation des résultats obtenus après plusieurs répétitions de l'expérience aléatoire. On considère que la fréquence relative de réalisation d'un événement A après n répétitions de l'expérience aléatoire est une bonne approximation de la probabilité de A. La probabilité empirique d'un événement A se calcule ainsi :

Probabilité empirique
$P(A) = \frac{\text{nombre de fois que l'événement A se produit}}{\text{nombre de répétitions de l'expérience aléatoire}} = \frac{n_A}{n} = \text{fréquence relative de A}$

EXEMPLE 1

Répartition des détenteurs d'un DEC en formation technique entrés sur le marché du travail en 2011-2012 selon la situation de l'emploi en juin 2013

Situation de l'emploi	Programme d'études de la formation technique	
	Ensemble des programmes	Administration, commerce et informatique
Ont un emploi	11 332	1 382
N'ont pas d'emploi	441	71
Total	11 773	1 453

Source: Ministère de l'Enseignement supérieur. *La relance au collégial en formation technique – 2013. La situation d'emploi de personnes diplômées. Enquêtes de 2011/2012/2013*, 2014.

Utiliser les statistiques du tableau pour estimer la probabilité qu'un détenteur d'un DEC en formation technique décroche un emploi dans l'année suivant la fin de ses études. Cette probabilité est-elle la même pour les diplômés du secteur administration, commerce et informatique ?

Solution

Ensemble des programmes : $P(E) =$

Administration, commerce et informatique : $P(E) =$

Il est à souligner que, contrairement à la probabilité classique, la probabilité empirique peut varier au fil du temps. Par exemple, en 2009, la probabilité qu'un diplômé en formation technique décroche un emploi à la fin de ses études était de 95,6 %.

EXEMPLE 2

En 2011, le Québec comptait 1 762 hôtels classifiés. Voici la distribution de la classification.

Répartition des hôtels selon la classification, Québec, 2011

Classification	Nombre	Pourcentage
1 étoile	253	14,4 %
2 étoiles	722	41,0 %
3 étoiles	536	30,4 %
4 étoiles	231	13,1 %
5 étoiles	20	1,1 %
Total	1 762	100,0 %

Source: Institut de la statistique du Québec. Données fournies par le ministère du Tourisme du Québec. *Établissements d'hébergement selon la classification officielle, Québec, 2011*, février 2012.

On choisit un hôtel au hasard. Quelle est la probabilité que ce soit un hôtel 4 étoiles ou plus ?

Solution

Soit l'événement C : «avoir une cote de 4 étoiles ou plus».

2.4 Les événements particuliers

Un événement étant un sous-ensemble de l'espace échantillonnal, la définition, la notation et la représentation graphique d'un événement particulier reposent sur la théorie des ensembles. Le lancer du dé va servir à illustrer chaque type d'événements. On sait que, dans ce cas, l'espace échantillonnal est $S = \{1, 2, 3, 4, 5, 6\}$.

Événement certain

Un événement égal à l'espace échantillonnal S est un **événement certain**.

Par exemple, pour le lancer d'un dé, l'événement A : «obtenir un nombre inférieur à 7» est un événement certain, puisque $A = S$.

Événement impossible

Un événement qui ne peut se produire est un **événement impossible**; il est égal à l'ensemble vide (\emptyset).

Par exemple, l'événement A: «obtenir un nombre supérieur à 6» est un événement impossible, puisque $A = \emptyset$.

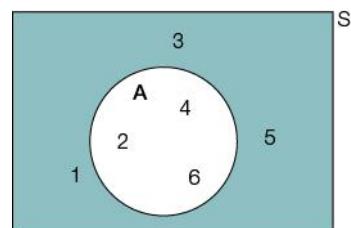
Événement contraire

L'**événement contraire** d'un événement A, noté A' , est l'événement composé des résultats de S qui n'appartiennent pas à A.

Par exemple, si A est l'événement «obtenir un nombre pair», alors A' est l'événement «ne pas obtenir un nombre pair». La partie ombrée du diagramme de Venn ci-contre correspond à l'événement A' .

$$A = \{2, 4, 6\} \quad A' = \{1, 3, 5\}$$

$$\text{On a: } n(A') = n(S) - n(A) = 6 - 3 = 3$$



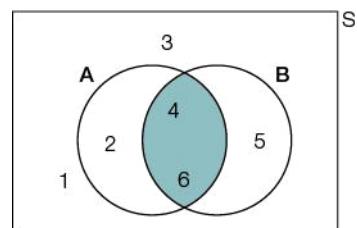
Intersection de deux événements ($A \cap B$)

L'**intersection** de deux événements A et B, notée $A \cap B$ (lire «A et B»), est un événement composé des résultats de S qui appartiennent à la fois à A et à B.

Par exemple, si A est l'événement «obtenir un nombre pair» et B est l'événement «obtenir un nombre supérieur à 3», alors $A \cap B$ est l'événement «obtenir un nombre pair supérieur à 3» :

$$A = \{2, 4, 6\} \quad B = \{4, 5, 6\} \quad A \cap B = \{4, 6\}$$

$$\text{On a: } n(A \cap B) = 2$$

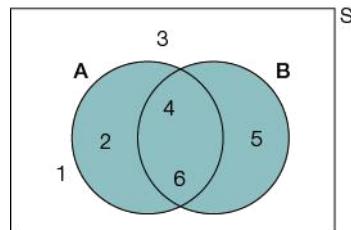


Union de deux événements ($A \cup B$)

L'**union** de deux événements A et B, notée $A \cup B$ (lire «A ou B»), est un événement composé des résultats de S qui appartiennent à A ou à B, ou aux deux à la fois.

Attention !

Le connecteur «ou» peut être inclusif ou exclusif. En théorie des probabilités, il a toujours un sens inclusif (A ou B ou les deux).



Par exemple, si l'événement A est «obtenir un nombre pair» et l'événement B, «obtenir un nombre supérieur à 3», alors $A \cup B$ est l'événement «obtenir un nombre pair ou un nombre supérieur à 3» :

$$A = \{2, 4, 6\} \quad B = \{4, 5, 6\} \quad A \cup B = \{2, 4, 5, 6\}$$

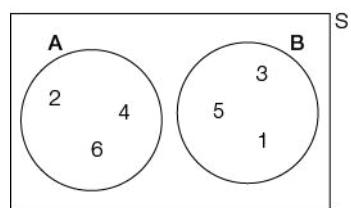
Le nombre de résultats favorables à l'événement $A \cup B$ se calcule ainsi :

$$n(A \cup B) = n(A) + n(B) - n(A \cap B) = 3 + 3 - 2 = 4$$

Il est important de comprendre que l'on retranche $n(A \cap B)$ de la somme $n(A) + n(B)$ pour ne pas compter les éléments de $A \cap B$ deux fois. En effet, les nombres 4 et 6 sont comptés parmi les 3 éléments de A et sont de nouveau comptés parmi les 3 éléments de B.

Événements incompatibles

Deux événements sont dits **incompatibles** s'ils ne peuvent se produire simultanément. Si deux événements A et B sont incompatibles, alors $A \cap B = \emptyset$.



Par exemple, A : «obtenir un nombre pair» et B : «obtenir un nombre impair» sont deux événements incompatibles, car l'événement $A \cap B$: «obtenir un nombre qui est à la fois pair et impair» est impossible.

On a $A \cap B = \emptyset$ et $n(A \cap B) = 0$.

Si deux événements A et B sont incompatibles, alors $n(A \cup B) = n(A) + n(B)$, car dans ce cas $n(A \cap B) = 0$.

Par exemple, pour l'événement $A \cup B$: «obtenir un nombre pair ou un nombre impair», on a $n(A \cup B) = n(A) + n(B) = 3 + 3 = 6$. Ici, l'événement $A \cup B$ est un événement certain.

EXEMPLE 1

Une expérience aléatoire consiste à piger une carte dans un jeu de 52 cartes. Soit les événements :

N : «piger une carte noire» F : «piger une figure (roi, dame, valet)» C : «piger un cœur»

Décrire en compréhension et donner le nombre de résultats des événements suivants :

- a) C' b) $F \cap C$ c) $N \cup F$ d) $C \cap N$

Solution

- a) C' :

$$n(C') =$$

- b) $F \cap C$:

$$n(F \cap C) =$$

c) $N \cup F$:

$$n(N \cup F) =$$

d) $C \cap N$:

$$n(C \cap N) =$$

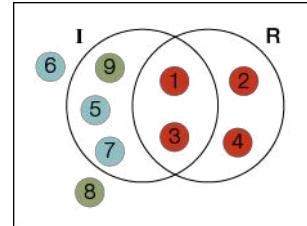
EXEMPLE 2

On pige une boule dans une urne qui contient neuf boules : quatre boules rouges numérotées 1, 2, 3, 4 ; trois boules bleues numérotées 5, 6, 7 ; deux boules vertes numérotées 8, 9.

a) Calculer la probabilité que la boule pigée ne soit pas rouge.

b) Calculer la probabilité que la boule pigée soit identifiée par un numéro impair ou qu'elle soit rouge.

Solution



2.5 Les propriétés des probabilités

Voici les principales propriétés des probabilités.

Propriété 1 La probabilité d'un événement A est toujours comprise entre 0 et 1 (ou 0 % et 100 %) :

$$0 \leq P(A) \leq 1 \quad \text{ou} \quad 0 \% \leq P(A) \leq 100 \%$$

Propriété 2 La probabilité d'un événement certain est 1 (ou 100 %) et celle d'un événement impossible est 0 (ou 0 %) :

$$P(S) = 1 \text{ (ou } 100\%) \quad \text{et} \quad P(\emptyset) = 0 \text{ (ou } 0\%)$$

Propriété 3 La probabilité de l'événement contraire, A' , d'un événement A est égale à 1 (ou 100 %) moins la probabilité de A :

Probabilité de l'événement contraire

$$P(A') = 1 - P(A) \quad \text{ou} \quad P(A') = 100 \% - P(A)$$

Par exemple, si l'on pige une carte au hasard dans un jeu de 52 cartes, il y a 25 % de chances (13/52) de piger un cœur. On peut en déduire qu'il y a 75 % de chances de ne pas piger un cœur.

Propriété 4 La probabilité que l'événement (A ou B) se produise, notée $P(A \cup B)$, est égale à la somme des probabilités des événements A et B moins la probabilité que A et B se produisent simultanément, celle-ci étant notée $P(A \cap B)$. Cette propriété est appelée règle d'addition :

Règle d'addition

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

On prouve cette règle ainsi :

$$\begin{aligned} P(A \cup B) &= \frac{n(A \cup B)}{n(S)} \\ &= \frac{n(A) + n(B) - n(A \cap B)}{n(S)} \\ &= \frac{n(A)}{n(S)} + \frac{n(B)}{n(S)} - \frac{n(A \cap B)}{n(S)} \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

NOTE

Si deux événements A et B sont incompatibles (c'est-à-dire qu'ils ne peuvent pas se produire simultanément), alors $P(A \cup B) = P(A) + P(B)$, car $P(A \cap B) = P(\emptyset) = 0$.

EXEMPLE 1

Un professeur a demandé à ses élèves s'ils avaient accès à un ordinateur portable ou à une tablette numérique à la maison. La compilation des réponses révèle que 30 % d'entre eux ont accès à un ordinateur portable, 15 % à une tablette numérique et 10 % aux deux appareils.

- Présenter les statistiques recueillies dans un tableau.
- Calculer la probabilité qu'un élève ait accès à un ordinateur portable ou à une tablette numérique.
- Calculer la probabilité qu'un élève ait accès à un ordinateur portable, mais pas à une tablette numérique.
- Calculer la probabilité qu'un élève n'ait ni accès à un ordinateur portable ni accès à une tablette numérique.

Solution

- Soit les événements suivants :

$S = \{\text{L'ensemble des élèves de la classe}\}$

O : «avoir accès à un ordinateur portable»

T : «avoir accès à une tablette numérique»

Le tableau ci-contre présente les statistiques recueillies.

	T	T'	Total
O	10 %	20 %	30 %
O'	5 %	65 %	70 %
Total	15 %	85 %	100 %

- Comme on ne peut pas utiliser la définition de probabilité pour calculer $P(O \cup T)$, car $n(O \cup T)$ est inconnu, nous appliquerons la règle d'addition :

$$\begin{aligned} P(O \cup T) &= P(O) + P(T) - P(O \cap T) \\ &= 30 \% + 15 \% - 10 \% = 35 \% \end{aligned}$$

Le tableau montre qu'en additionnant 30 % et 15 % on compte deux fois 10 % : on doit donc enlever un des deux 10 %.

c) $P(O \cap T') = 20\%$

d) $P(O' \cap T') = 65\%$

L'encadré suivant suggère une démarche qui facilite grandement la résolution de problèmes de probabilité.

Démarche de résolution de problèmes de probabilité

- On assigne une lettre majuscule à chacun des événements en cause.
- Si la donnée du problème donne de l'information sur l'intersection de deux événements, $n(A \cap B)$ ou $P(A \cap B)$, on construit l'un ou l'autre des tableaux suivants.

Si l'on a $n(A)$, $n(B)$, $n(A \cap B)$ et $n(S)$

	B	B'	Total
A	$n(A \cap B)$	$n(A \cap B')$	$n(A)$
A'	$n(A' \cap B)$	$n(A' \cap B')$	$n(A')$
Total	$n(B)$	$n(B')$	$n(S)$

Si l'on a $P(A)$, $P(B)$ et $P(A \cap B)$

	B	B'	Total
A	$P(A \cap B)$	$P(A \cap B')$	$P(A)$
A'	$P(A' \cap B)$	$P(A' \cap B')$	$P(A')$
Total	$P(B)$	$P(B')$	$P(S)$

- On exprime la question en employant le symbolisme suivant:

P : «La probabilité de...» ou «les chances d'avoir...»

\cap : et

\cup : ou (inclusif)

- On calcule la probabilité demandée en utilisant:

- la définition de probabilité, si l'on peut dénombrer les résultats favorables à l'événement;
- les propriétés, si l'on ne peut pas dénombrer les résultats favorables à l'événement.

EXEMPLE 2

En 2011-2012, 166 112 étudiants ont fait une demande d'aide financière au gouvernement: 154 475 demandeurs ont reçu un prêt et 101 479 ont reçu une bourse. Parmi les récipiendaires d'une bourse, 2 252 n'ont pas reçu de prêt.

Source: Ministère de l'Enseignement supérieur. *Statistiques. Rapport 2011-2012*, 2014.

On choisit au hasard un demandeur d'aide financière.

- Quelle est la probabilité qu'il ait reçu un prêt et une bourse ?
- Quelle est la probabilité qu'il ait reçu un prêt, mais pas une bourse ?
- Quelle est la probabilité qu'il n'ait reçu ni prêt, ni bourse ?
- Quelle est la probabilité qu'il ait reçu un prêt ou une bourse ?

Solution

$S = \{\text{L'ensemble des 166 112 demandeurs}\}$

P : «recevoir un prêt»

B : «recevoir une bourse»

	P	P'	Total
B			
B'			
Total			

EXEMPLE 3

Le tableau suivant dresse le profil des médecins québécois en fonction du sexe et de l'âge.

Répartition des médecins selon le groupe d'âge et le sexe, Québec, 2013

Âge (en ans)	Femmes (F)	Hommes (H)	Total
Moins de 40 (A)	3 238	1 748	4 986
De 40 à 49 (B)	2 534	2 029	4 563
De 50 à 59 (C)	2 311	3 161	5 472
60 et plus (D)	923	3 874	4 797
Total	9 006	10 812	19 818

Source: Collège des médecins du Québec. *Répartition des médecins selon le groupe d'âge et selon le sexe*, 31 décembre 2013.

D'après les statistiques de 2013 :

- quelle est la probabilité qu'un médecin soit une femme ?
- quelle est la probabilité qu'un médecin ait moins de 40 ans ?
- quelle est la probabilité qu'un médecin soit une femme de moins de 40 ans ?
- quelle est la probabilité qu'un médecin soit un homme ou qu'il fasse partie du groupe des 50 à 59 ans ?

Solution

EXERCICE DE COMPRÉHENSION | 2.1

Un sondage effectué auprès d'un échantillon d'adultes québécois révèle les statistiques suivantes : 55 % ont fait du vélo durant l'année, 48 % ont accès à une auto et ont fait du vélo durant l'année et 12 % n'ont pas accès à une auto.

Source: Vélo Québec. *État de la pratique du vélo au Québec en 2010*, mai 2011.

En se basant sur ces statistiques, calculer la probabilité qu'un adulte québécois choisi au hasard :

- n'ait pas accès à une auto et n'ait pas fait de vélo durant l'année.
- n'ait pas accès à une auto ou n'ait pas fait de vélo durant l'année.

Solution

$S = \{\text{Ensemble des répondants}\}$

V : « faire du vélo »

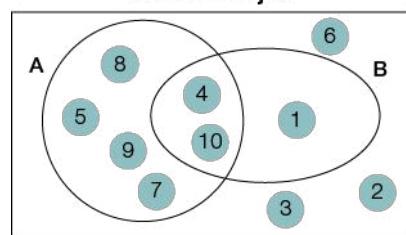
A : « avoir accès à une auto »

	V	V'	Total
A			
A'			
Total			

EXERCICES 2.1

- Dans chaque cas, décrire en extension l'espace échantillonnal.
 - Le nombre de truites prises par un pêcheur en une journée, le quota quotidien étant de 15.
 - Un psychologue scolaire note le temps que met un enfant à lire une page d'un texte, le temps maximal alloué étant de 5 min.
 - Lors d'un test psychologique, on compte le nombre de pièces d'un casse-tête de 20 pièces qu'un jeune enfant réussit à placer en 10 min.
 - On note le mois de naissance d'une personne choisie au hasard.
- On lance une pièce de monnaie trois fois et on note le résultat de chaque lancer.
 - Décrire en extension l'espace échantillonnal.
 - Décrire en extension chacun des événements suivants et en calculer la probabilité.
 - A : « obtenir exactement deux piles ».
 - B : « obtenir seulement des faces ».
 - C : « obtenir plus de piles que de faces ».
 - D : « obtenir autant de piles que de faces ».
 - E : « obtenir moins de quatre piles ».
- Parmi les 52 cartes à jouer, on pige successivement deux cartes au hasard. Soit les événements :
 - R_1 : « la première carte pigée est rouge »;
 - R_2 : « la deuxième carte pigée est rouge »;
 - D_1 : « la première carte pigée est une dame »;
 - D_2 : « la deuxième carte pigée est une dame ».
 Décrire en compréhension les événements suivants :
 - $D_1 \cap R_1$
 - $R_1 \cap R_2$
 - $D_1 \cup D_2$
 - $R_1 \cap R'_2$
- Dans un pot contenant des jetons numérotés de 1 à 10, on pige au hasard un jeton, puis on le place sur le cercle de la planche de jeu qui contient le numéro.

Planche de jeu



Exprimer l'événement décrit en langage mathématique, puis en calculer la probabilité :

- a) «le jeton se situe dans la zone A»;
 - b) «le jeton se situe dans la zone A et dans la zone B»;
 - c) «le jeton se situe dans la zone A ou dans la zone B»;
 - d) «le jeton se situe ni dans la zone A, ni dans la zone B»;
 - e) «le jeton se situe dans la zone A, mais pas dans la zone B».
5. En 2011, il y a eu 88 618 naissances au Québec, dont 45 313 garçons. On sait qu'un nouveau-né de faible poids (moins de 2 500 g) risque davantage d'avoir des problèmes de santé. En 2011, on a dénombré 5 012 nouveau-nés, dont 2 383 garçons, qui pesaient moins de 2 500 g.
- Source:** Institut de la statistique du Québec. *Naissances selon le poids à la naissance, le groupe d'âge de la mère et le sexe, Québec, 2011*, 6 août 2013.
- Selon ces statistiques :
- a) Quels sont les risques qu'un nouveau-né pèse moins de 2 500 g à la naissance ?
 - b) Quels sont les risques qu'un nouveau-né soit une fille pesant moins de 2 500 g ?
 - c) Quelle est la probabilité qu'un nouveau-né ne soit ni un garçon, ni un bébé de moins de 2 500 g ?
 - d) Quelle est la probabilité qu'un nouveau-né soit un garçon ou un bébé pesant moins de 2 500 g ?
 - e) La probabilité empirique que le nouveau-né soit un garçon est-elle de 50 % ?
6. Une étude effectuée en 2012 auprès des ménages québécois révèle les statistiques suivantes : 81,5 % sont branchés à Internet; 23,4 % ont des enfants et sont branchés à Internet; 17,1 % n'ont pas d'enfants et ne sont pas branchés à Internet.
- Source:** Institut de la statistique du Québec. *Enquête québécoise sur l'accès des ménages à Internet 2012*, 2013.
- Selon ces statistiques, quelle est la probabilité qu'un ménage québécois :
- a) ne soit pas branché à Internet ?
 - b) n'ait pas d'enfants ?
 - c) soit branché à Internet et n'ait pas d'enfants ?
 - d) ne soit pas branché à Internet ou ait des enfants ?
7. a) En utilisant le fait qu'une année compte 365 jours, calculer la probabilité classique qu'une personne soit née en janvier.

b) *A priori*, quelle hypothèse a été formulée pour répondre à la question a) ?

c) En utilisant les statistiques du tableau suivant, calculer la probabilité empirique qu'une personne soit née en janvier. Comparer le résultat avec la probabilité classique calculée en a) et commenter.

Répartition du nombre¹ de naissances, de décès et de mariages par mois, Québec, 2012

Mois	Naissances	Décès	Mariages
Janvier	7 100	5 600	650
Février	6 650	5 000	800
Mars	7 200	5 400	800
Avril	7 050	5 000	850
Mai	7 600	4 950	1 850
Juin	7 150	4 600	3 000
Juillet	7 700	4 850	3 700
Août	8 000	4 800	4 350
Septembre	7 850	4 600	3 750
Octobre	7 850	5 000	1 850
Novembre	7 450	4 900	850
Décembre	7 100	6 100	1 150
Total	88 700	60 800	23 600

1. Les nombres sont arrondis à 50 près.

Source: Institut de la statistique du Québec. *Naissances, décès, mariages par mois*, septembre 2013.

d) D'après les données du tableau, une personne a-t-elle statistiquement la même probabilité de décéder en mars qu'en septembre ?

e) Estimer les chances qu'un couple se marie entre le 1^{er} juin et le 31 août.

8. En 2010, on dénombre 650 danseurs professionnels au Québec : 472 femmes et 178 hommes. Des 490 danseurs gagnant 15 000 \$ ou plus par année, 151 sont des hommes.

Source: Institut de la statistique du Québec. *Enquête auprès des danseurs et chorégraphes du Québec, 2010*, juillet 2012.

En se basant sur ces statistiques, donner la probabilité qu'un danseur professionnel :

- a) soit un homme.
- b) ait un revenu inférieur à 15 000 \$.
- c) soit une femme ayant un revenu inférieur à 15 000 \$.
- d) soit un homme ayant un revenu inférieur à 15 000 \$.
- e) soit un homme ou une personne gagnant 15 000 \$ ou plus par année.

2.6 La probabilité conditionnelle

On appelle **probabilité conditionnelle** d'un événement A par rapport à un événement B la probabilité que l'événement A se réalise sachant que l'événement B s'est déjà réalisé.

La probabilité conditionnelle d'un événement A par rapport à un événement B se note $P(A | B)$ (lire «probabilité d'obtenir A si B s'est produit» ou «probabilité d'obtenir A sachant que B s'est produit»). Les deux mises en situation suivantes illustrent cette notion.

MISE EN

SITUATION 1

Reprendons le tableau de distribution de l'âge et du sexe des médecins québécois.

Répartition des médecins selon le groupe d'âge et le sexe, Québec, 2013

Âge (en ans)	Femmes (F)	Hommes (H)	Total
Moins de 40 (A)	3 238	1 748	4 986
De 40 à 49 (B)	2 534	2 029	4 563
De 50 à 59 (C)	2 311	3 161	5 472
60 et plus (D)	923	3 874	4 797
Total	9 006	10 812	19 818

Source: Collège des médecins du Québec. *Répartition des médecins selon le groupe d'âge et selon le sexe*, 31 décembre 2013.

Selon les données du tableau :

a) Quelle est la probabilité qu'un médecin soit une femme ?

$$P(F) =$$

b) Quelle est la probabilité qu'un médecin soit une femme s'il a moins de 40 ans ?

$$P(F | A) =$$

Symboliquement, on peut traduire ce raisonnement ainsi :

$$P(F | A) = \frac{n(F \cap A)}{n(A)}$$

Pour déterminer la probabilité conditionnelle, il faut tenir compte de l'information donnée : le médecin a moins de 40 ans (l'événement A est réalisé). On doit donc retirer de l'espace échantillonnaux tous les médecins âgés de 40 ans et plus : il ne reste alors que 4 986 médecins, dont 3 238 sont des femmes. Le rapport de ces deux nombres est la probabilité cherchée.

NOTE

On a affaire à une probabilité conditionnelle chaque fois qu'une information indique qu'il y a une restriction de l'espace échantillonnaux S : certains résultats ne peuvent pas être obtenus.

On lance un dé et on s'intéresse aux deux événements suivants :

Q : « obtenir un nombre supérieur à 4 »; on a $Q = \{5, 6\}$

I : « obtenir un nombre impair »; on a $I = \{1, 3, 5\}$

a) Quelles sont les chances d'obtenir un nombre supérieur à 4 ?

$$P(Q) =$$

b) Quelles sont les chances d'obtenir un nombre supérieur à 4, sachant qu'on a obtenu un nombre impair ?

$$P(Q | I) =$$

$$\text{Symboliquement on a : } P(Q | I) = \frac{n(Q \cap I)}{n(I)}$$

Des deux mises en situation précédentes, on déduit la formule suivante :

Probabilité conditionnelle

$$P(A | B) = \frac{n(A \cap B)}{n(B)} \quad \text{ou} \quad P(A | B) = \frac{P(A \cap B)}{P(B)} \quad \text{où } B \neq \emptyset$$

NOTE

La formule de droite, que l'on utilise quand il est impossible de dénombrer les éléments des événements A et B , s'obtient comme suit :

$$P(A | B) = \frac{n(A \cap B)}{n(B)} = \frac{\cancel{n(A \cap B)}}{\cancel{n(B)} / n(S)} = \frac{P(A \cap B)}{P(B)}$$

EXEMPLE 1

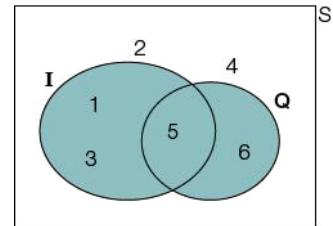
On pige une carte au hasard dans un jeu de 52 cartes.

a) Si l'on pige un cœur (C), quelle est la probabilité que ce soit une dame (D) ?

Solution

b) Quelle est la probabilité de piger un cœur (C) si la carte pigée est une dame (D) ?

Solution



EXEMPLE 2

D'où proviennent les passagers des navires de croisières qui sillonnent le Saint-Laurent? Voici des statistiques à ce sujet.

Répartition des passagers des navires de croisières internationales dans les ports du Saint-Laurent selon le type et la provenance, Québec, 2012

Provenance	Type de passager		Total
	Croisiériste (C)	Membre de l'équipage (E)	
États-Unis (US)	55,2 %	1,3 %	56,5 %
Europe (Eu)	16,7 %	1,9 %	18,6 %
Canada (Ca)	15,3 %	0,4 %	15,7 %
Autre (A)	4,9 %	4,2 %	9,1 %
Total	92,1 %	7,8 %	99,9 %

Source: Tourisme Québec. *Étude auprès des croisiéristes et des membres d'équipage des navires de croisières dans les ports du Saint-Laurent*, juin 2013.

- a) Quelle est la probabilité qu'un passager provienne des États-Unis? Cette probabilité est-elle la même si le passager est membre de l'équipage?
- b) Si un passager est d'origine européenne, quelle est la probabilité que ce soit un croisiériste?
- c) Quelle est la probabilité qu'un passager provienne d'un autre endroit que les États-Unis, l'Europe ou le Canada si c'est un membre de l'équipage?

Solution

EXERCICE DE COMPRÉHENSION | 2.2

En 2010, le Québec compte 1 500 écrivains professionnels¹ dont 710 ont moins de 45 ans. On dénombre 974 écrivains qui ont fait moins de 5 000 \$ avec leurs créations littéraires durant l'année. Parmi ceux-ci, 422 ont moins de 45 ans.

Source: Institut de la statistique du Québec. *Enquête auprès des écrivains du Québec, 2010*, 2013.

Utiliser ces statistiques pour répondre aux questions suivantes.

- a) Calculer la probabilité qu'un écrivain touche 5 000 \$ et plus pour ses créations littéraires s'il a 45 ans et plus.
- b) Quelle est la probabilité qu'un écrivain soit âgé de moins de 45 ans et touche moins de 5 000 \$ pour ses créations littéraires?

1. Écrivains qui ont publié au moins deux livres au cours de leur carrière.

- c) i) Quelle est la probabilité qu'un écrivain touche moins de 5 000 \$ pour ses créations littéraires ?
ii) Cette probabilité est-elle plus élevée chez les écrivains de moins de 45 ans ?

Solution

	Moins de 5 000 \$ (A)	5 000 \$ et plus (A')	Total
Moins de 45 ans (B)			
45 ans et plus (B')			
Total			

2.7 Les événements indépendants

Il est souvent intéressant de savoir si deux événements A et B ont une influence l'un sur l'autre, c'est-à-dire si le fait que l'événement B se soit produit modifie la probabilité de l'événement A. Si c'est le cas, les événements A et B sont dits **dépendants**; sinon A et B sont dits **indépendants**.

Événements indépendants

A et B sont indépendants si:

$$P(A) = P(A | B)$$

Attention !

Il ne faut pas confondre l'indépendance et l'incompatibilité de deux événements : A et B sont incompatibles si $A \cap B = \emptyset$.

Analogie

Comme la notion d'indépendance est fondamentale en statistique, il est important d'en bien saisir le sens. L'analogie suivante devrait aider à bien comprendre la définition d'indépendance énoncée ci-dessus.

On voudrait savoir si la consommation d'alcool d'André est influencée par la présence de l'un ou l'autre de ses amis (Bruno, Carl et Diane). Si c'est le cas, augmente-t-elle ou diminue-t-elle ?

Pour répondre à ces questions, il faut d'abord déterminer les habitudes de consommation d'André lorsqu'il n'est pas en présence d'un de ses amis, puis vérifier si la présence d'un de ses amis modifie ses habitudes de consommation.

Soit :

$P(A)$, la probabilité qu'André consomme plus de 3 bières lorsqu'il va dans un bar sans être accompagné d'un ami ;

$P(A | B)$, la probabilité qu'André consomme plus de 3 bières si Bruno l'accompagne ;

$P(A | C)$, la probabilité qu'André consomme plus de 3 bières si Carl l'accompagne ;
 $P(A | D)$, la probabilité qu'André consomme plus de 3 bières si Diane l'accompagne.

On observe que :

$$P(A) = 30 \%$$

$$P(A | B) = 40 \%$$

$$P(A | C) = 30 \%$$

$$P(A | D) = 20 \%$$

On tire les conclusions suivantes de ces statistiques :

- André est influencé par la présence de Bruno : la probabilité qu'il consomme plus de 3 bières augmente de 10 points de pourcentage si Bruno l'accompagne. Statistiquement, les événements A et B sont dépendants.
- André n'est pas influencé par la présence de Carl : la probabilité qu'il consomme plus de 3 bières ne change pas si Carl l'accompagne, car $P(A) = P(A | C) = 30 \%$. Statistiquement, les événements A et C sont indépendants.
- André est influencé par la présence de Diane : la probabilité qu'il consomme plus de 3 bières diminue de 10 points de pourcentage si Diane l'accompagne. Statistiquement, les événements A et D sont dépendants.

EXEMPLE 1

Les tableaux suivants sont tirés de l'étude portant sur les 1 500 écrivains professionnels du Québec présentée dans l'exercice de compréhension 2.2, à la page 102. Ils vont permettre d'analyser le revenu des écrivains en fonction de l'âge et du sexe.

a)

**Répartition des écrivains professionnels
selon l'âge et le revenu tiré de la création littéraire, Québec, 2010**

Âge	Revenu tiré de la création littéraire		Total
	Moins de 5 000 \$ (A)	5 000 \$ et plus (A')	
Moins de 45 ans (B)	28,1 %	19,2 %	47,3 %
45 ans et plus (B')	36,8 %	15,9 %	52,7 %
Total	64,9 %	35,1 %	100,0 %

Source: Institut de la statistique du Québec. *Enquête auprès des écrivains du Québec, 2010, 2013*.

Les statistiques du tableau permettent-elles de conclure que les événements A : «toucher moins de 5 000 \$ de ses créations littéraires» et B : «avoir moins de 45 ans» sont indépendants ? Justifier et interpréter la réponse.

Solution

Les événements sont indépendants si $P(A) = P(A | B)$.

On a : $P(A) = 64,9 \% \approx 65 \%$

$$P(A | B) = \frac{28,1 \%}{47,3 \%} = 59,4 \% \approx 59 \%$$

Puisque $P(A) \neq P(A | B)$, les événements A et B sont dépendants.

Interprétation

Globalement, si l'on ne tient pas compte de l'âge, la probabilité qu'un écrivain touche moins de 5 000 \$ de ses créations littéraires est de 65 %. Or, si l'écrivain a moins de 45 ans, cette probabilité baisse à 59 %. La variable «âge» fait diminuer de près de 6 points la probabilité qu'un écrivain touche moins de 5 000 \$: l'âge de l'écrivain a donc une influence sur le revenu tiré de la création littéraire.

b)

Répartition des écrivains professionnels selon le sexe et le revenu tiré de la création littéraire, Québec, 2010

Sexe	Revenu tiré de la création littéraire		Total
	Moins de 5 000 \$ (A)	5 000 \$ et plus (A')	
Femmes (F)	29,4 %	15,8 %	45,2 %
Hommes (H)	35,5 %	19,3 %	54,8 %
Total	64,9 %	35,1 %	100,0 %

Source: Institut de la statistique du Québec. *Enquête auprès des écrivains du Québec, 2010*, 2013.

Les statistiques du tableau permettent-elles de conclure que les événements A : «toucher moins de 5 000 \$ de ses créations littéraires» et F : «l'écrivain est une femme» sont indépendants ? Justifier et interpréter la réponse.

Solution

Interprétation

Le fait que l'écrivain soit une femme _____ la probabilité de toucher moins de 5 000 \$ de la création littéraire. Le revenu touché est _____ du sexe de l'écrivain.

EXEMPLE 2

L'étude de la distribution de l'âge et du sexe des médecins a permis de construire le tableau suivant.

Répartition des médecins selon le groupe d'âge et le sexe, Québec, 2013

Âge (en ans)	Femmes (F)	Hommes (H)	Total
Moins de 40 (A)	3 238	1 748	4 986
De 40 à 49 (B)	2 534	2 029	4 563
De 50 à 59 (C)	2 311	3 161	5 472
60 et plus (D)	923	3 874	4 797
Total	9 006	10 812	19 818

Source: Collège des médecins du Québec. *Répartition des médecins selon le groupe d'âge et selon le sexe, 31 décembre 2013*.

- a) Peut-on conclure de ces statistiques qu'il y a un lien entre l'âge d'un médecin et le sexe de celui-ci ? Pour répondre à cette question, vérifier, pour chaque tranche d'âge, s'il y a une dépendance entre l'âge d'un médecin et le fait que le médecin soit une femme, et interpréter le résultat.

Solution

Est-ce que $P(F) = P(F | A) = P(F | B) = P(F | C) = P(F | D)$?

OUI: indépendance entre le sexe et l'âge
NON: dépendance entre le sexe et l'âge

$$P(F) =$$

$$P(F | A) = \neq P(F) \Rightarrow F \text{ et } A \text{ sont dépendants.}$$

$$P(F | B) = \neq P(F) \Rightarrow F \text{ et } B \text{ sont dépendants.}$$

$$P(F | C) = \frac{2311}{5472} = 42,2 \% \neq P(F) \Rightarrow F \text{ et } C \text{ sont dépendants.}$$

$$P(F | D) = \frac{923}{4797} = 19,2 \% \neq P(F) \Rightarrow F \text{ et } D \text{ sont dépendants.}$$

Conclusion

Il y a _____ entre l'âge d'un médecin et le fait que le médecin soit une femme.

Interprétation

Globalement, si l'on ne tient pas compte de l'âge, la probabilité qu'un médecin soit une femme est de ____ %. Or, si l'on tient compte de l'âge, cette probabilité est plus élevée dans les deux premiers groupes d'âge, soit ____ % et ____ % respectivement, et plus basse dans les deux derniers groupes d'âge, soit ____ % et ____ % respectivement.

Une tendance se dégage de ces statistiques : plus un médecin est jeune, plus la probabilité que celui-ci soit une femme est élevée.

- b) Supposons qu'une analyse de la répartition des médecins selon la région de résidence et le sexe indique qu'il n'y a pas de lien entre le sexe du médecin et sa région de résidence. Dans ce cas :
- quelle est la probabilité qu'un médecin soit une femme si celui-ci réside dans les Laurentides (L) ?
 - quelle est la probabilité qu'un médecin soit une femme si celui-ci réside en Estrie (E) ?

Solution

Tableau de distribution conditionnelle

On présente souvent les données d'une étude dans un tableau de distribution conditionnelle afin de faciliter l'analyse de la dépendance entre deux variables. Ce tableau affiche la probabilité conditionnelle d'une variable par rapport aux catégories de l'autre variable. Il se caractérise par le fait que le total de chaque ligne (ou colonne) est de 100 %.

EXEMPLE

Voici le tableau de distribution conditionnelle représentant les données de l'exemple 2 qui précède.

Répartition des médecins selon le groupe d'âge et le sexe, Québec, 2013

Âge (en ans)	Femmes (F)	Hommes (H)	Total
Moins de 40 (A)	64,9 %	35,1 %	100,0 %
De 40 à 49 (B)	55,5 %	44,5 %	100,0 %
De 50 à 59 (C)	42,2 %	57,8 %	100,0 %
60 et plus (D)	19,2 %	80,8 %	100,0 %
Total	45,4 %	54,6 %	100,0 %

Source: Collège des médecins du Québec. *Répartition des médecins selon le groupe d'âge et selon le sexe*, 31 décembre 2013.

En comparant le pourcentage de la ligne «Total» à ceux des autres lignes, on mesure rapidement l'influence de l'âge sur la probabilité qu'un médecin soit une femme. De plus, il est facile de voir la tendance qui se dégage des données : plus le médecin est jeune, plus la probabilité qu'il soit une femme est élevée.

EXERCICE DE COMPRÉHENSION | 2.3

La consommation accrue d'eau embouteillée soulève des questions d'ordre environnemental et éthique. Le tableau suivant est tiré d'une étude visant à déterminer les caractéristiques des ménages qui consomment de l'eau embouteillée.

**Répartition des ménages, par tranche de revenu,
selon la consommation d'eau embouteillée à la maison, Canada, 2006**

Revenu du ménage	Consomme de l'eau embouteillée		Total
	Oui (O)	Non (N)	
Moins de 40 000 \$ (A)	25 %	75 %	100 %
[40 000 \$; 64 000 \$[(B)	29 %	71 %	100 %
[64 000 \$; 91 000 \$[(C)	32 %	68 %	100 %
91 000 \$ et plus (D)	33 %	67 %	100 %
Total	29 %	71 %	100 %

Source: Statistique Canada. *EnviroStats*, juin 2008.

a) Vrai ou faux ?

- Les événements O et A sont indépendants. _____
- Les événements O et B sont indépendants. _____
- Il y a un lien entre la consommation d'eau embouteillée à la maison et le revenu du ménage.

b) Compléter l'énoncé.

Globalement, la probabilité qu'un ménage consomme de l'eau embouteillée à la maison est de ____ %. Cette probabilité diminue de ____ points de pourcentage pour les ménages dont le revenu est inférieur à 40 000 \$ et augmente de ____ points de pourcentage pour les ménages dont le revenu est compris entre 64 000 \$ et 91 000 \$.

c) Les données de la colonne «Oui» permettent de dégager une tendance. Énoncer cette tendance.

2.8 La règle de multiplication

La définition de la probabilité conditionnelle permet de découvrir une nouvelle façon de calculer la probabilité que deux événements A et B se produisent simultanément, soit $P(A \cap B)$. On a :

- $P(A | B) = \frac{P(A \cap B)}{P(B)}$ [Définition de la probabilité conditionnelle]

- $P(A \cap B) = P(B) P(A | B)$ [On isole $P(A \cap B)$]

- Comme $A \cap B = B \cap A$, la dernière égalité s'écrit également
 $P(B \cap A) = P(B) P(A | B)$

En remplaçant la lettre B par la lettre A, et vice versa, on obtient une formule plus facile à mémoriser, car les lettres suivent l'ordre alphabétique :

$$P(A \cap B) = P(A) P(B | A)$$

On donne le nom de règle de multiplication à cette formule.

Règle de multiplication

$$P(A \cap B) = P(A) P(B | A)$$

NOTE

Si A et B sont indépendants alors $P(A \cap B) = P(A) P(B)$, car $P(B | A)$ est égal à $P(B)$.

Certains auteurs utilisent cette égalité pour vérifier si A et B sont indépendants : on vérifie si $P(A \cap B) = P(A) P(B)$. Si oui, alors A et B sont indépendants.

Comparaison des deux méthodes pour calculer $P(A \cap B)$

Nous connaissons maintenant deux méthodes pour calculer la probabilité de l'intersection de deux événements : la définition de probabilité et la règle de multiplication. Quelle méthode faut-il choisir pour calculer $P(A \cap B)$? Tout dépend de l'expérience aléatoire menée. La mise en situation suivante va permettre d'y voir un peu plus clair quant à la méthode à privilégier pour calculer $P(A \cap B)$.

MISE EN

SITUATION

Pour les deux expériences aléatoires suivantes, nous calculerons $P(A \cap B)$ en utilisant les deux méthodes de calcul, puis nous comparerons ces méthodes afin de déterminer la méthode la plus efficace.

1. On lance un dé deux fois. Soit les événements :

A : « obtenir un nombre pair au premier lancer »

B : « obtenir un nombre impair au second lancer »

Calculer la probabilité d'obtenir un nombre pair au premier lancer et un nombre impair au second lancer, soit $P(A \cap B)$.

Solution

- Calculons $P(A \cap B)$ à l'aide de la définition de probabilité :

Énumérons d'abord les éléments de S et de $A \cap B$ afin de les dénombrer :

$$\begin{aligned} S = & \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ & (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\} \end{aligned}$$

$$A \cap B = \{(2, 1), (2, 3), (2, 5), (4, 1), (4, 3), (4, 5), (6, 1), (6, 3), (6, 5)\}$$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{9}{36}$$

- Calculons $P(A \cap B)$ en appliquant la règle de multiplication :

$$P(A \cap B) = P(A) P(B | A) = \frac{3}{6} \times \frac{3}{6} = \frac{9}{36}$$

Ici, les événements A et B sont indépendants, car la probabilité d'obtenir un nombre impair au second lancer n'est pas influencée par le fait que l'on obtienne ou non un nombre pair au premier lancer, donc $P(B | A) = P(B)$, d'où l'on peut écrire : $P(A \cap B) = P(A) P(B)$.

- Comparons les deux méthodes :

Pour cette expérience aléatoire, la règle de multiplication est la méthode à privilégier, car elle est la plus rapide : il est en effet fastidieux de dénombrer les éléments de S et de $A \cap B$ afin de pouvoir appliquer la définition de probabilité.

- Le tableau suivant donne des statistiques sur les étudiants inscrits à un voyage culturel pendant la relâche.

Répartition des étudiants inscrits à un voyage culturel selon le sexe et la destination

Destination	Filles (F)	Garçons (G)	Total
New York (N)	44	51	95
Boston (B)	14	11	25
Total	58	62	120

On choisit un étudiant au hasard parmi les 120 étudiants inscrits à un voyage culturel. Calculer la probabilité que ce soit un garçon inscrit pour le voyage à New York.

Solution

- Calculons $P(G \cap N)$ à l'aide de la définition de probabilité :

$$P(G \cap N) = \frac{n(G \cap N)}{n(S)} = \frac{51}{120}$$

- Calculons $P(G \cap N)$ en appliquant la règle de multiplication :

$$P(G \cap N) = P(G) P(N | G) = \frac{62}{120} \times \frac{51}{62} = \frac{51}{120}$$

- Comparons les deux méthodes :

Pour ce type d'expérience aléatoire, le calcul à l'aide de la définition de probabilité est la méthode à privilégier, car elle est la plus simple : il y a moins d'opérations à effectuer.

Quelle méthode faut-il choisir pour calculer $P(A \cap B)$?

Si l'expérience aléatoire est décomposable en épreuves successives, comme piger deux cartes ou lancer un dé deux fois, il est préférable d'utiliser la règle de multiplication pour calculer $P(A \cap B)$, car celle-ci permet de traiter isolément chaque épreuve. Par contre, si l'expérience aléatoire est simple (ou non décomposable), comme choisir une personne dans un groupe ou piger une carte, il est préférable d'appliquer la définition de probabilité pour calculer $P(A \cap B)$.

EXEMPLE 1

On pique une carte dans un jeu de 52 cartes. Calculer la probabilité de piger un roi (R) noir (N).

Solution

EXEMPLE 2

Parmi les 12 figures d'un jeu de cartes, on pique successivement deux cartes sans remise (c'est-à-dire qu'on ne remet pas la première carte pigée dans le paquet). Calculer la probabilité de piger deux dames.

Solution

Il s'agit d'une expérience aléatoire décomposable en deux tirages : l'événement A : «piger deux dames» se produit uniquement si l'on obtient une dame à chaque tirage.

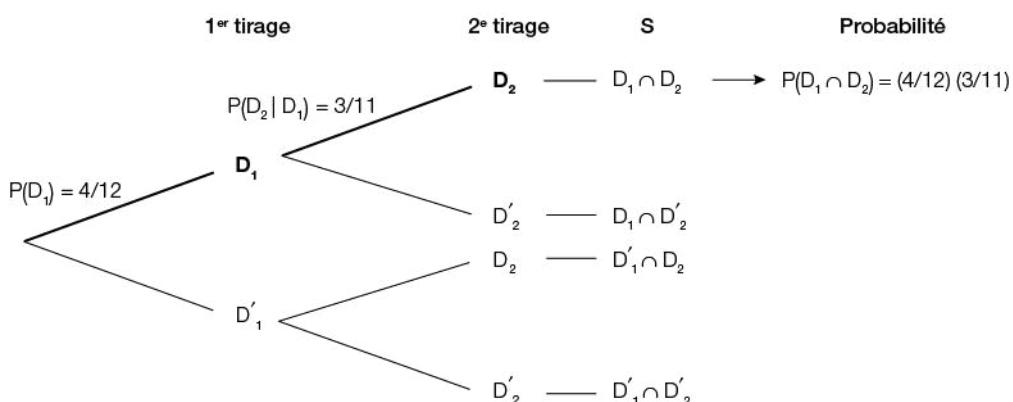
Soit les événements : D_1 : «piger une dame au premier tirage» et D_2 : «piger une dame au second tirage».

$$\begin{aligned} P(A) &= P(D_1 \cap D_2) \\ &= P(D_1) P(D_2 | D_1) \\ &= \end{aligned}$$

Diagramme en arbre

Un diagramme en arbre permet de représenter par une suite de branches tous les résultats possibles d'une expérience aléatoire décomposable en épreuves successives. La première branche indique la probabilité que l'événement A se réalise à la 1^{re} épreuve, soit $P(A)$; la branche suivante indique la probabilité que l'événement B se réalise à la 2^e épreuve sachant que l'événement A s'est réalisé à la 1^{re} épreuve, soit $P(A | B)$. En multipliant les probabilités de deux branches successives, ce qui revient à appliquer la règle de multiplication, on obtient la probabilité que l'événement A se réalise à la 1^{re} épreuve et que l'événement B se réalise à la 2^e épreuve, soit $P(A \cap B)$.

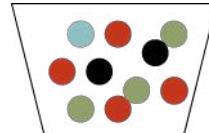
Le diagramme en arbre ci-dessous représente tous les résultats possibles de l'expérience aléatoire décrite à l'exemple 2 précédent. Le chemin tracé en gras mène à la probabilité demandée, soit $P(D_1 \cap D_2)$.



EXEMPLE 1

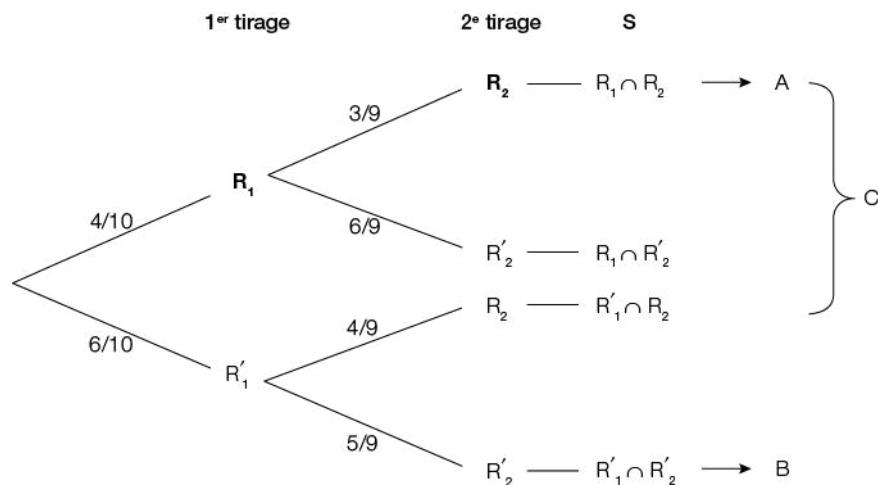
On pique successivement, sans remise, deux boules dans une urne qui en contient dix : quatre rouges, trois vertes, deux noires et une bleue. Calculer la probabilité de réalisation des événements suivants et indiquer le ou les chemins du diagramme en arbre menant à l'événement.

- a) A : «piger deux boules rouges»
- b) B : «ne piger aucune boule rouge»
- c) C : «piger au moins une boule rouge»



Solution

Comme la description des événements A, B et C indique que nous nous intéressons au fait que la boule pigée soit rouge ou ne soit pas rouge, nous construirons un diagramme en arbre donnant tous les résultats possibles de cette expérience aléatoire en fonction des événements R : «la boule pigée est rouge» et R' : «la boule pigée n'est pas rouge».



a) $P(A) = P(\text{piger deux boules rouges}) = P(\text{piger une boule rouge à la 1^{re} pige et à la 2^e pige})$

$$\begin{aligned} P(A) &= P(R_1 \cap R_2) \\ &= P(R_1) P(R_2 | R_1) \\ &= \frac{4}{10} \times \frac{3}{9} = 13,3 \% \end{aligned}$$

b) $P(B) = P(\text{ne piger aucune boule rouge}) = P(\text{ne pas piger une boule rouge à la 1^{re} pige et ne pas piger une boule rouge à la 2^e pige})$

$$\begin{aligned} P(B) &= P(R'_1 \cap R'_2) \\ &= P(R'_1) P(R'_2 | R'_1) \\ &= \frac{6}{10} \times \frac{5}{9} = 33,3 \% \end{aligned}$$

c) $P(C) = P(\text{piger au moins une boule rouge}) = P(\text{tous les cas possibles}) - P(\text{ne piger aucune boule rouge})$

$$\begin{aligned} P(C) &= P(S) - P(B) \\ &= 100 \% - 33,3 \% \\ &= 66,7 \% \end{aligned}$$

EXEMPLE 2

Le dernier recensement révèle que 55 % des couples canadiens de même sexe sont de sexe masculin. Le pourcentage de couples avec des enfants (E) est de 4 % chez les couples de sexe masculin (M) et de 17 % chez les couples de sexe féminin (F).

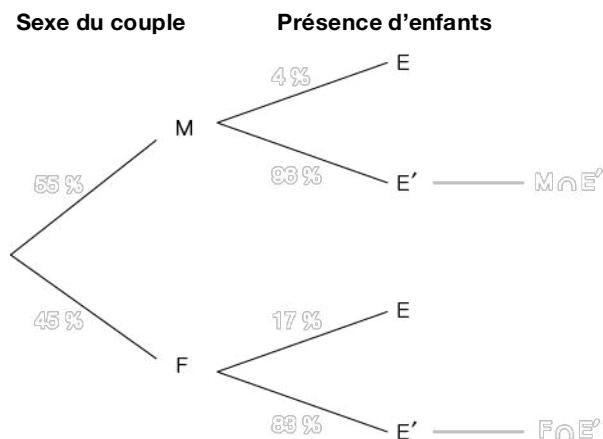
Source: Statistique Canada. *Recensement 2011.*

En se basant sur ces statistiques, calculer la probabilité qu'un couple de même sexe :

- a) n'ait pas d'enfants s'il est de sexe masculin.
- b) ait des enfants et soit de sexe féminin.
- c) n'ait pas d'enfants.
- d) soit de sexe masculin s'il n'a pas d'enfants.

Solution

Comme la donnée du problème fournit des probabilités conditionnelles (par exemple, $P(E | M) = 4 \%$), il est approprié de construire un diagramme en arbre pour représenter la situation. À la 1^{re} épreuve, nous nous intéresserons au sexe du couple et à la 2^e épreuve, à la présence d'enfants dans le couple.



EXEMPLE 3

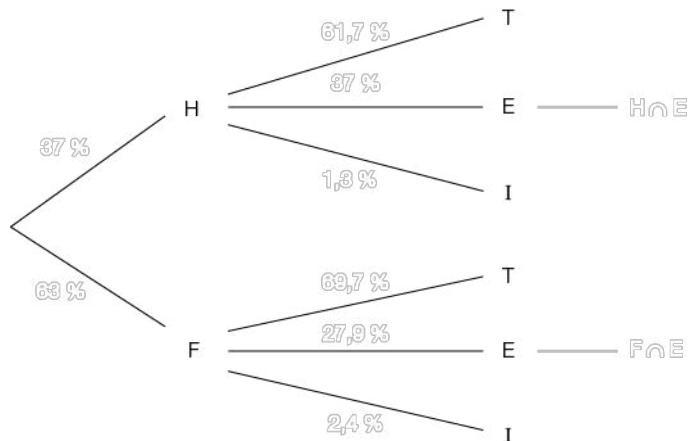
Une étude indique que 37 % des détenteurs d'un DEC en formation technique sont des hommes. Dans l'année suivant l'obtention de leur diplôme, 61,7 % des diplômés masculins sont sur le marché du travail, 37 % poursuivent des études et 1,3 % sont inactifs (ni aux études ni sur le marché du travail). Chez les diplômés féminins, ces pourcentages sont 69,7 %, 27,9 % et 2,4 % respectivement.

Source: Ministère de l'Enseignement supérieur. *La relance au collégial en formation technique – 2013. La situation d'emploi de personnes diplômées. Enquêtes de 2011/2012/2013*, 2014.

Représenter la situation par un diagramme en arbre, puis répondre aux questions suivantes.

- Quelle est la probabilité qu'une personne poursuive ses études après l'obtention d'un DEC en formation technique ?
- La probabilité qu'une personne poursuive ses études après l'obtention d'un DEC en formation technique dépend-elle du sexe du diplômé ?
- Sachant qu'une personne a poursuivi ses études après son DEC en formation technique, quelle est la probabilité que ce soit un homme ?

Solution



EXERCICE DE COMPRÉHENSION | 2.4

Les travailleurs indépendants² consacrent-ils davantage d'heures à leur travail que les employés ?

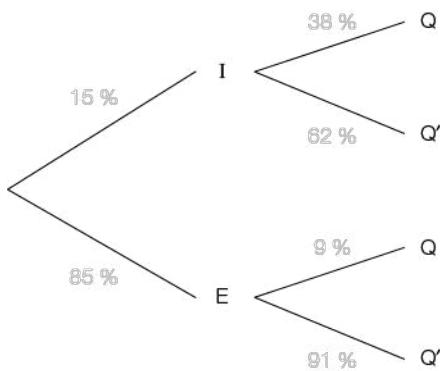
Une étude révèle que 15 % des travailleurs canadiens qui occupent un emploi sont des travailleurs indépendants (I). Parmi ces derniers, 38 % consacrent plus de 40 heures par semaine (Q) à leur travail alors que ce pourcentage est de 9 % chez les employés (E).

Source: Industrie Canada. *Principales statistiques relatives aux petites entreprises – juillet 2012.*

Représenter la situation dans le diagramme en arbre ci-dessous, puis répondre aux questions suivantes.

- Quelle est la probabilité qu'un travailleur consacre plus de 40 heures par semaine à son travail ?
- Quelle est la probabilité qu'un travailleur consacre 40 heures par semaine ou moins à son travail, si c'est un employé ?
- Quelle est la probabilité qu'un travailleur soit un employé qui consacre 40 heures par semaine ou moins à son travail ?
- Sachant qu'un travailleur consacre plus de 40 heures par semaine à son travail, quelle est la probabilité que ce soit un travailleur indépendant ?
- Les événements Q et I sont-ils indépendants ? Justifier et interpréter la réponse.

Solution



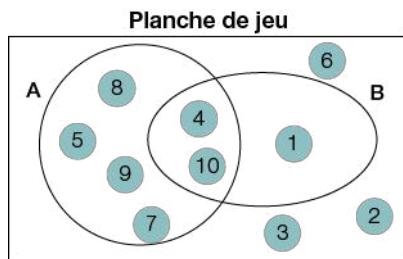
2. Un travailleur indépendant est une personne qui tire directement son revenu de l'exploitation de son entreprise ou de l'exercice de son métier ou de sa profession au lieu de recevoir un salaire d'un employeur.

Faut-il construire un tableau ou un diagramme en arbre ?

- Si l'expérience aléatoire est simple et que les statistiques données dans le problème portent sur l'intersection de deux événements, $n(A \cap B)$ ou $P(A \cap B)$, on construit un tableau.
- Si l'expérience aléatoire est décomposable (la donnée du problème contient des probabilités conditionnelles $P(A | B)$), on construit un diagramme en arbre.

EXERCICES 2.2

1. Dans un pot contenant des jetons numérotés de 1 à 10, on pique au hasard un jeton, puis on le place sur le cercle de la planche de jeu qui contient le numéro.



Soit les deux événements suivants :

- A : «piger un jeton de la zone A»
B : «piger un jeton de la zone B».

- a) Calculer les probabilités suivantes :
- | | |
|------------------|------------------------------|
| i) $P(A B)$ | iv) $P(B A')$ |
| ii) $P(B A)$ | v) $P(A A \cup B)$ |
| iii) $P(A B')$ | vi) $P(A \cap B A \cup B)$ |
- b) i) Les événements A et B sont-ils incompatibles ? Justifier la réponse.
ii) Les événements A et B sont-ils indépendants ? Justifier la réponse.
2. Répartition des diplômés universitaires selon le sexe et le grade obtenu, Québec, promotion 2010

Sexe	Grade obtenu			
	Baccalauréat	Maîtrise	Doctorat	Total
Femmes	21 155	5 706	766	27 627
Hommes	13 011	4 914	875	18 800
Total	34 166	10 620	1 641	46 427

Source: Ministère de l'Enseignement supérieur. *Indicateur de l'éducation, édition 2012, 2013.*

On choisit un diplômé au hasard.

- Quelles sont les chances que ce soit une femme ?
 - Quelles sont les chances que ce soit un homme détenteur d'une maîtrise ?
 - Quelles sont les chances que cette personne détienne un baccalauréat si c'est une femme ?
 - Si la personne détient un doctorat ou une maîtrise, quelles sont les chances que ce soit un homme ?
 - Selon les données du tableau, existe-t-il un lien entre le fait que le diplôme obtenu soit un doctorat et le sexe du diplômé ? Justifier et interpréter la réponse.
3. On lance un dé.
- Quelle est la probabilité d'obtenir un nombre divisible par 3 sachant que le nombre obtenu est pair ?
 - Les événements A : «obtenir un nombre divisible par 3» et B : «obtenir un nombre pair» sont-ils indépendants ? Justifier et interpréter la réponse.
4. On pique une carte d'un jeu de 52 cartes. Soit les événements V : «piger un valet» et F : «piger une figure (roi, dame, valet)».
- Calculer $P(V | F)$.
 - Calculer $P(F | V)$.
 - Les événements V et F sont-ils indépendants ?
5. Un homme divorcé a-t-il plus de chances de se remarier avec une femme célibataire ou avec une femme divorcée ? Voici quelques statistiques à ce sujet.

Répartition des mariages¹ selon l'état matrimonial des époux, Québec, 2011

Époux	Épouse			Total
	Célibataire (F_C)	Veuve (F_V)	Divorcée (F_D)	
Célibataire (H_C)	65,7 %	0,4 %	7,9 %	74,0 %
Veuf (H_V)	0,6 %	0,7 %	1,1 %	2,4 %
Divorcé (H_D)	10,4 %	1,0 %	12,2 %	23,6 %
Total	76,7 %	2,1 %	21,2 %	100,0 %

1. Mariages de conjoints de sexe opposé.

Source: Institut de la statistique du Québec. *Mariages selon l'état matrimonial des époux, 2005-2011*.

Exprimer la question en langage mathématique avant d'y répondre.

- Quelle est la probabilité qu'un mariage unisse deux personnes célibataires ?
 - Quelle est la probabilité que le marié soit célibataire ?
 - Quelle est la probabilité que le marié ou la mariée soient célibataires ?
 - Si le marié est veuf, quelle est la probabilité que la mariée soit veuve ?
 - Si la mariée est veuve, quelle est la probabilité que le marié soit veuf ?
 - La probabilité que le marié soit divorcé dépend-elle de l'état matrimonial de la mariée ? Justifier et interpréter la réponse.
6. Quel est le taux de syndicalisation dans les entreprises québécoises ?

Une étude révèle qu'en 2012, 49,8 % des employés travaillent pour des petites et moyennes entreprises (PME). Le pourcentage de travailleurs syndiqués est de 27,3 % dans les PME alors qu'il est de 56,2 % dans les grandes entreprises (GE).

Source: Statistique Canada. *Enquête sur la population active, 2012*, adapté par l'Institut de la statistique du Québec, dans *Flash-info*, vol. 13, n° 3, septembre 2013.

- Construire un diagramme en arbre représentant les résultats de l'étude.
- Sachant qu'une personne travaille pour une PME, quelle est la probabilité qu'elle ne soit pas syndiquée ?
- Quelle est la probabilité qu'un travailleur ne soit pas syndiqué et travaille pour une grande entreprise ?
- Calculer la probabilité qu'un travailleur soit syndiqué.

e) La probabilité qu'un travailleur soit syndiqué diminue-t-elle si celui-ci travaille pour une PME ? Si oui, de combien de points de pourcentage ?

f) Quelle est la probabilité qu'un travailleur québécois ne soit pas syndiqué ?

g) Sachant qu'un travailleur est syndiqué, calculer la probabilité qu'il travaille pour une grande entreprise.

7. Un constructeur de véhicules automobiles possède trois usines : 45 % de la production se fait à l'usine A, 35 % à l'usine B et le reste à l'usine C. On sait que 3 % des automobiles fabriquées à l'usine A doivent faire l'objet d'au moins un rappel, tandis que cette proportion est respectivement de 5 % et de 6 % pour les usines B et C.

- Calculer la probabilité qu'une automobile fasse l'objet d'au moins un rappel.
- Si un véhicule fait l'objet d'un rappel, quelle est la probabilité qu'il ait été fabriqué à l'usine C ?

8. En 2012, les ventes d'enregistrements audio au Québec se répartissent ainsi : 70 % sont sur support physique (CD ou disque vinyle) et 30 % sont sur support numérique. La proportion d'enregistrements en français est de 40 % pour la musique sur support physique et de 34 % pour la musique sur support numérique.

Source: Institut de la statistique du Québec. *Optique culture, n° 24*, mai 2013.

- Quelle est la probabilité de vendre de la musique en français sur support physique au Québec ?
- Quelle est la probabilité qu'un enregistrement audio vendu au Québec ne soit pas en français ?
- Sachant qu'une personne a acheté de la musique en français, quelle est la probabilité que ce soit de la musique numérique ?

9. Quelle proportion des petites et moyennes entreprises (PME) canadiennes ont un site Web ?

Répartition des PME, par nombre d'employés, selon leur présence sur Internet, Canada, 2011

Nombre d'employés	Possède un site Web		Total
	Oui	Non	
Moins de 20	63,3 %	36,7 %	100,0 %
De 20 à 99	81,0 %	19,0 %	100,0 %
De 100 à 499	96,9 %	3,1 %	100,0 %
Total	69,8 %	30,2 %	100,0 %

Source: CEFARIO. *NetPME 2011 : L'utilisation des TIC par les PME canadiennes et québécoises*, octobre 2011.

- a) Au Canada, 51 % des PME comptent moins de 20 employés. On choisit une PME au hasard.
- Quelle est la probabilité qu'elle compte moins de 20 employés et ait un site Web?
 - Sachant que l'entreprise choisie ne possède pas de site Web, quelle est la probabilité qu'elle ait moins de 20 employés?
- b) Peut-on dire que la probabilité qu'une PME possède un site Web dépend de son nombre d'employés?
- Justifier et interpréter la réponse.
- c) On pige avec remise trois PME dans la liste des PME canadiennes.
- Quelle est la probabilité que les trois PME pigées aient un site Web?
 - Quelle est la probabilité qu'au moins une des trois PME pigées ait un site Web?
10. Une personne peut se rendre au travail en voiture, en métro ou en train. Elle a l'habitude de prendre sa voiture un jour sur cinq, le métro trois jours sur cinq et le train un jour sur cinq. La probabilité qu'elle soit en retard au travail est de 10 % si elle prend sa voiture, de 2 % si elle prend le métro et de 3 % si elle prend le train.
- Quelle est la probabilité que la personne arrive en retard au travail?
 - Si la personne arrive en retard au travail, quelle est la probabilité qu'elle ait pris sa voiture?
11. Le tableau suivant est tiré d'une étude portant sur les entreprises privées canadiennes.

Répartition des entreprises selon la taille et le secteur d'activité, Canada, 2012

Taille de l'entreprise	Secteur d'activité		
	Services	Biens	Total
Petites et moyennes ¹	74 %	26 %	100 %
Grandes ²	76 %	24 %	100 %
Total	75 %	25 %	100 %

1. Moins de 500 employés.

2. 500 employés et plus.

Source: Industrie Canada. *Principales statistiques relatives aux entreprises – juillet 2012.*

En 2012, les PME représentent 48 % des entreprises privées au Canada. Considérer ce fait pour répondre aux questions suivantes.

- Calculer la probabilité qu'une entreprise soit une PME du secteur des services.
 - Calculer la probabilité qu'une entreprise soit une grande entreprise du secteur des biens.
 - Si une entreprise est dans le secteur des services, quelle est la probabilité que ce soit une PME?
 - Sachant qu'une entreprise est dans le secteur des biens, quelle est la probabilité que ce soit une grande entreprise?
 - Peut-on dire que la probabilité qu'une entreprise canadienne soit dans le secteur des services dépend de sa taille? Justifier et interpréter la réponse.
 - Supposons que l'on pige, avec remise, deux entreprises dans la liste des entreprises canadiennes.
 - Calculer la probabilité de piger deux PME.
 - Calculer la probabilité de piger au moins une PME.
12. On pige successivement trois cartes sans remise dans un jeu de 52 cartes. À chaque tirage, on s'intéresse à l'événement C: «piger un cœur».
- Construire un arbre représentant tous les résultats possibles de cette expérience aléatoire.
 - Calculer la probabilité de piger trois cartes de cœur.
 - Calculer la probabilité de ne piger aucun cœur.
 - À l'aide de la probabilité de l'événement contraire, calculer la probabilité de piger au moins un cœur.

2.9 L'analyse combinatoire

Quelles sont les chances de gagner le gros lot à la loterie 6/49 ?

Quels sont les risques qu'une personne mal intentionnée découvre le code d'accès de votre téléphone intelligent ?

Quelles sont les chances de piger un échantillon particulier de 100 personnes dans une population de 3 000 personnes ?

Pour calculer ces probabilités, nous devons dénombrer tous les résultats possibles de ces expériences aléatoires. Or, il est impensable de le faire en dressant la liste de tous ces résultats : il y en a trop ! L'analyse combinatoire fournit des outils pour dénombrer rapidement les résultats possibles de ce type d'expériences.

La présente section ne vise pas à faire une étude exhaustive de l'analyse combinatoire. Elle a plutôt pour objectif de présenter brièvement les techniques de dénombrement nécessaires à la compréhension de la loi binomiale qui sera étudiée au chapitre 3.

2.9.1 Le principe de multiplication

Le principe de multiplication est à la base des techniques de dénombrement en analyse combinatoire. Il s'énonce ainsi :

«Le nombre de résultats engendrés par une suite d'opérations est le produit du nombre de choix possibles à chacune des opérations.»

EXEMPLE 1

Un restaurant offre en table d'hôte 4 choix pour l'entrée, 8 choix pour le plat principal et 3 choix pour le dessert. Combien de tables d'hôte différentes peut-on composer avec ces choix ?

Solution

Pour composer une table d'hôte, il faut d'abord choisir une entrée, puis un plat principal et finalement un dessert. On indique ci-dessous le nombre de choix offerts à chaque étape du processus :

Entrée	Plat principal	Dessert
4 choix	8 choix	3 choix

Nombre de tables d'hôte possibles = $4 \times 8 \times 3 = 96$

EXEMPLE 2

Si l'on emploie les chiffres 1, 2, 3, 4, combien de dossiers peut-on identifier avec une cote de quatre chiffres ?

Solution

1 ^{er} chiffre	2 ^e chiffre	3 ^e chiffre	4 ^e chiffre
4 choix	4 choix	4 choix	4 choix

Nombre de cotes possibles = $4 \times 4 \times 4 \times 4 = 256$

Exemples de cotes : 1111, 2111, 3111, 4111, 1211, 1311, etc.

2.9.2 Les permutations et les arrangements

On s'intéresse ici plus particulièrement au nombre de résultats possibles d'expériences qui n'admettent pas la répétition d'éléments et où l'ordre des éléments est important.

Notation factorielle

L'application du principe de multiplication amène souvent à multiplier plusieurs entiers consécutifs : par exemple, $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$. Pour alléger l'écriture de ce type de produit, on a recours à la **notation factorielle**, notée $n!$ (se lit n factorielle). Le produit précédent correspond à $8!$ et sa valeur est 40 320. La notation factorielle se définit ainsi :

$$n! = n(n-1)(n-2) \times \cdots \times 3 \times 2 \times 1 \quad \text{pour tout entier } n \text{ positif}$$

$$0! = 1 \text{ par convention}$$

EXEMPLE

- a) Donner la valeur de $6!$. b) Simplifier $\frac{80!}{78!}$. c) Simplifier $\frac{70!}{3!(70-3)!}$.

Solution

a) $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

b) $\frac{80!}{78!} = \frac{80 \times 79 \times 78!}{78!} = 80 \times 79 = 6\,320$

c) $\frac{70!}{3!(70-3)!} = \frac{70!}{3!67!} = \frac{70 \times 69 \times 68 \times 67!}{3!67!} = \frac{70 \times 69 \times 68}{3 \times 2 \times 1} = 54\,740$

NOTE

Certaines calculatrices ont un bouton ($n!$) qui permet de calculer des factorielles. Toutefois, la valeur maximale que l'on peut calculer est $69!$.

Permutations

Une **permutation** est une disposition ordonnée de n éléments différents. Par exemple, les 6 permutations possibles des lettres A, B, C sont : ABC, ACB, BAC, BCA, CAB et CBA. On dénombre toutes les permutations possibles de n éléments avec le principe de multiplication.

EXEMPLE 1

Si l'on emploie les chiffres 1, 2, 3, 4, combien de dossiers peut-on identifier avec une cote de quatre chiffres distincts ?

Solution

1 ^{er} chiffre	2 ^e chiffre	3 ^e chiffre	4 ^e chiffre
4 choix	3 choix	2 choix	1 choix

Nombre de cotés possibles = $4! = 24$

Exemples de cotés : 1234, 2143, 4321, 2413, etc.

Ces cotés correspondent aux 24 permutations possibles des chiffres 1, 2, 3, 4.

EXEMPLE 2

Combien d'anagrammes peut-on faire en permutant les 5 lettres du mot BRAVO ?

Solution

1 ^{re} lettre	2 ^e lettre	3 ^e lettre	4 ^e lettre	5 ^e lettre
5 choix	4 choix	3 choix	2 choix	1 choix

Nombre de permutations = $5! = 120$ anagrammes

Arrangements

Un **arrangement** est une disposition ordonnée de x éléments différents choisis parmi n éléments différents. Par exemple, en sélectionnant 2 lettres parmi les lettres A, B, C, on obtient les 6 arrangements suivants : AB, AC, BA, BC, CA et CB. À noter que si l'on sélectionnait les 3 lettres, on obtiendrait les permutations des lettres A, B, C. Le principe de multiplication permet de dénombrer tous les arrangements possibles.

EXEMPLE 1

On emploie les chiffres 1, 2, 3, 4 pour coter des dossiers. Combien de dossiers peut-on identifier avec une cote de deux chiffres distincts ?

Solution

1 ^{er} chiffre	2 ^e chiffre
4 choix	3 choix

Nombre de cotes possibles = $4 \times 3 = 12$

Les cotes possibles sont : 12, 13, 14, 21, 23, 24, 31, 32, 34, 41, 42, 43.

Ces 12 cotes correspondent à tous les arrangements de 2 chiffres différents choisis parmi 1, 2, 3, 4.

EXEMPLE 2

On choisit un président et un secrétaire d'assemblée parmi cinq candidats.

- Combien d'arrangements possibles y a-t-il pour ces postes ?
- Exprimer le nombre d'arrangements possibles à l'aide de factorielles.

Solution

a)

Président	Secrétaire
5 choix	4 choix

Nombre d'arrangements possibles = $5 \times 4 = 20$

- Par un artifice de calcul, on peut écrire :

$$5 \times 4 = \frac{5 \times 4}{1} \times \frac{3!}{3!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = \frac{5!}{3!}$$

2.9.3 Les combinaisons

On s'intéressera ici au nombre de résultats possibles d'expériences qui n'admettent pas la répétition d'éléments et où l'ordre des éléments n'est pas important.

Le Lotto 6/49 consiste à choisir 6 nombres compris entre 1 et 49. Comme l'ordre des nombres choisis n'a pas d'importance, on dit que les 6 nombres forment une combinaison. Par exemple, les sélections {12, 25, 32, 33, 40, 45} et {25, 12, 32, 33, 40, 45} forment une seule et même combinaison, car elles se composent des mêmes nombres, l'inversion de la position des nombres 12 et 25 n'ayant pas d'importance pour cette loterie.

Il en est de même pour la sélection sans remise d'un échantillon de personnes : l'ordre de sélection des personnes qui constituent l'échantillon n'a aucune importance. Deux échantillons seront différents si au moins une personne est différente. Un échantillon prélevé sans remise est une combinaison de x personnes choisies parmi les n personnes de la population. Comment peut-on dénombrer tous les échantillons possibles ?

La mise en situation suivante va répondre à cette question. Auparavant, définissons la notion de combinaison.

Une **combinaison** est une disposition non ordonnée de x éléments différents choisis parmi n éléments différents.

MISE EN

SITUATION

On prélève sans remise un échantillon aléatoire de 3 personnes dans une population composée de Anne, Bianca, Carl et David. Combien d'échantillons possibles y a-t-il pour cette expérience aléatoire ?

La population étant petite, il est facile ici d'énumérer les 4 échantillons possibles :

{Anne, Bianca, Carl}, {Anne, Bianca, David}, {Anne, Carl, David}, {Bianca, Carl, David}

Voyons comment on peut dénombrer ces échantillons sans les énumérer.

Techniquement, pour sélectionner l'échantillon, on pique une première personne, puis une deuxième et une troisième parmi les 4 personnes de la population. En procédant ainsi, on forme un arrangement, puisque l'on tient compte de l'ordre. En vertu du principe de multiplication, le nombre d'arrangements possibles est :

$$4 \times 3 \times 2 = 24 \text{ arrangements possibles}$$

En identifiant chaque personne par l'initiale de son prénom, on obtient les 24 arrangements ci-dessous. On observe que pour chaque combinaison de 3 personnes, on peut associer 6 arrangements en permutant l'ordre de sélection des 3 personnes : il y a donc 6 fois plus d'arrangements que de combinaisons.

Échantillons (combinaisons)	Arrangements
ABC	ABC, ACB, BCA, BAC, CAB, CBA
ABD	ABD, ADB, BAD, BDA, DAB, DBA
ACD	ACD, ADC, CAD, CDA, DAC, DCA
BCD	BCD, BDC, CBD, CDB, DBC, DCB

Pour obtenir le nombre d'échantillons (combinaisons) possibles, il faut donc diviser le nombre d'arrangements de 3 personnes choisies parmi 4 personnes par le nombre de permutations des 3 personnes.

$$\text{Nombre d'échantillons (combinaisons)} = \frac{4 \times 3 \times 2}{3!} = \frac{24}{6} = 4$$

EXEMPLE 1

- Combien d'échantillons possibles y a-t-il si l'on pige sans remise 4 personnes dans une population de 10 personnes ?
- Exprimer le nombre d'échantillons possibles à l'aide de factorielles.
- De façon générale, combien de combinaisons de x éléments choisis parmi n éléments y a-t-il ?

Solution

a) Nombre d'échantillons (combinaisons) = $\frac{10 \times 9 \times 8 \times 7}{4!} = \frac{5\,040}{24} = 210$

b) Par un artifice de calcul, on peut écrire :

$$\frac{10 \times 9 \times 8 \times 7}{4!} = \frac{10 \times 9 \times 8 \times 7}{4!} \times \frac{6!}{6!} = \frac{10!}{4!6!} \quad \text{ou, sous une autre forme, } \frac{10!}{4!(10-4)!}$$

c) En utilisant le résultat précédent comme modèle, le nombre de combinaisons de x éléments choisis parmi n correspond à l'expression factorielle suivante :

$$\text{Nombre de combinaisons} = \frac{n!}{x!(n-x)!}$$

Le nombre de combinaisons de x éléments choisis parmi n éléments, noté $\binom{n}{x}$, s'obtient ainsi :

Nombre de combinaisons de x éléments choisis parmi n

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad \text{où } 0 \leq x \leq n$$

EXEMPLE 2

Combien de combinaisons possibles y a-t-il au Lotto 6/49 ? Si vous achetez un billet, quelles sont vos chances de gagner le gros lot ?

Solution

$$\binom{49}{6} = \frac{49!}{6!(49-6)!} = \frac{49 \times 48 \times 47 \times 46 \times 45 \times 44 \times 43!}{6!43!} = \frac{49 \times 48 \times 47 \times 46 \times 45 \times 44}{6 \times 5 \times 4 \times 3 \times 2 \times 1}$$

= _____ combinaisons possibles

La probabilité de gagner le gros lot avec un billet =

EXEMPLE 3

On pige sans remise un échantillon de 3 personnes dans une population de 5 hommes et 4 femmes.

- Combien d'échantillons possibles y a-t-il ?
- Quelle est la probabilité que l'échantillon soit composé de 3 hommes ?
- Quelle est la probabilité que l'échantillon soit composé de 2 hommes et 1 femme ?

Solution

a) Nombre d'échantillons possibles :

b) Pour obtenir un échantillon composé de 3 hommes, il faut piger 3 hommes parmi les 5 hommes.

Soit l'événement A : «piger 3 hommes».

$$n(A) =$$

$$P(A) = \frac{n(A)}{n(S)} =$$

c) Soit l'événement B : «piger 2 hommes et 1 femme». Pour que B se réalise, il faut piger 2 hommes parmi les 5 hommes, puis 1 femme parmi les 4 femmes. En vertu du principe de multiplication, on a :

3 personnes	
2 hommes	1 femme
$\binom{5}{2}$	$\binom{4}{1}$

On note que la somme des nombres supérieurs de la notation $(5 + 4)$ donne 9, la taille de la population, et que la somme des nombres inférieurs $(2 + 1)$ donne 3, la taille de l'échantillon.

$$n(B) = \frac{5!}{2!3!} \times \frac{4!}{1!3!} = \frac{5 \times 4 \times 3!}{2!3!} \times \frac{4 \times 3!}{1!3!} = 10 \times 4 = 40$$

$$P(B) = \frac{n(B)}{n(S)} =$$

EXEMPLE 4

Daniel a été élu président de sa classe par 60 % des 30 élèves. Si l'on prélève sans remise un échantillon aléatoire de 5 élèves, quelle est la probabilité que 60 % d'entre eux aient appuyé Daniel?

Solution

L'échantillon qui nous intéresse comprend 3 élèves qui ont appuyé Daniel ($60\% \times 5$), et donc 2 élèves qui ne l'ont pas appuyé. Pour obtenir cet échantillon, il faut piger 3 élèves parmi les 18 ($60\% \times 30$) qui ont voté pour Daniel et 2 élèves parmi les 12 qui n'ont pas voté pour lui.

En posant A : «piger 3 élèves qui ont voté pour Daniel et 2 élèves qui n'ont pas voté pour lui», on calcule la probabilité demandée ainsi :

$$P(A) = \frac{n(A)}{n(S)} = \frac{\binom{18}{3} \binom{12}{2}}{\binom{30}{5}} = \frac{\frac{18!}{3!15!} \times \frac{12!}{2!10!}}{\frac{30!}{5!25!}} = \frac{816 \times 66}{142\,506} = 37,8\%$$

EXERCICES DE COMPRÉHENSION | 2.5

1. a) Combien d'anagrammes peut-on faire en permutant les lettres du mot CHAT ?

b) Combien d'arrangements de 2 lettres peut-on faire avec les lettres du mot CHAT ?
2. Combien de véhicules peut-on immatriculer avec une plaque d'immatriculation composée de 3 lettres suivies de 3 chiffres ?
3. On pige sans remise un échantillon de 3 personnes dans une population comptant 6 ouvriers et 4 techniciens. Quelles sont les chances de piger un échantillon comprenant 2 ouvriers et 1 technicien ?

EXERCICES 2.3

1. Évaluer les expressions suivantes.
a) $7!$ b) $\frac{102!}{100!}$ c) $\frac{10!}{8!3!}$ d) $\frac{8!}{4!(8-4)!}$
2. Le bar laitier Chocolats favoris offre 6 saveurs de crème glacée molle (vanille, chocolat, marbré, orange, fraise, érable) et 12 saveurs d'enrobage au chocolat (classique lait, classique noir, fondant noisette, blanc-bec, pétillant, pain d'épices, orange éclatant, café crème, *dulce de leche*, piment intense, et 2 saveurs-surprises). Si vous achetez un cornet de crème glacée molle par jour, combien de jours vous faudra-t-il pour goûter à toutes les associations de saveurs de crème glacée et de chocolat ?
3. Avec 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, vous devez créer un code d'accès de 4 chiffres pour votre téléphone intelligent.
a) Combien de codes possibles y a-t-il ?

b) Quels sont les risques qu'une personne mal intentionnée découvre votre code par hasard et du premier coup ?
4. Le service de l'Opération Nez rouge est assuré par une équipe de trois personnes : le chauffeur et le partenaire, qui prennent place dans la voiture du client, et l'escorte motorisée qui suit la voiture du client. Un groupe de 9 amis, 5 hommes et 4 femmes, se sont portés bénévoles pour une soirée.
 - a) En tenant compte des rôles attribués aux membres de l'équipe de raccompagnement, combien d'équipes de 3 personnes peut-on constituer avec ces 9 bénévoles ?
 - b) Dans combien d'équipes le rôle d'escorte motorisée est-il attribué à un homme ?
 - c) Dans combien d'équipes le chauffeur et le partenaire sont-ils des femmes ?

- d) Combien d'équipes comptent une femme comme escorte motorisée, un homme comme chauffeur et une femme comme partenaire ?
5. Une bibliothèque scolaire cote les livres avec 2 lettres suivies de 3 chiffres. Combien de livres peut-on coter avec ce système :
- si l'on peut répéter les lettres et les chiffres ?
 - si l'on peut répéter les lettres, mais pas les chiffres ?
6. Pour tout achat de 80 \$ et plus, une librairie offre au client de choisir gratuitement 3 livres de poche dans une liste contenant 10 titres pour enfants et 20 titres pour adultes. Si un client bénéficie de cette offre :
- combien de combinaisons de livres peut-il faire ?
 - combien de combinaisons peut-il faire s'il opte pour 3 livres pour enfants ?
 - combien de combinaisons peut-il faire s'il opte pour 2 livres pour adultes et 1 livre pour enfants ?
7. Une population comprend 10 Québécois, 6 Européens et 4 Américains. On prélève, au hasard et sans remise, un échantillon de 3 personnes dans cette population.
- Combien d'échantillons possibles y a-t-il ?
 - Quelles sont les chances de prélever un échantillon qui comprend 3 Européens ?
 - Quelles sont les chances de prélever un échantillon qui comprend 2 Québécois et 1 Européen ?
 - Quelles sont les chances de prélever un échantillon qui comprend exactement 2 Québécois ?
- e) Quelles sont les chances de prélever un échantillon qui comprend 1 Québécois, 1 Européen et 1 Américain ?
8. En 2013, 52 % des 50 albums musicaux les plus vendus au Québec sont québécois. On pige sans remise un échantillon de 6 albums parmi ces 50 albums. Quelle est la probabilité de piger un échantillon comprenant 50 % d'albums québécois ?
- Source:** Observatoire de la culture et des communications du Québec. *Les ventes d'enregistrements sonores en 2013*, avril 2014.
9. Un billet de loterie La Mini comporte 6 chiffres dont le 1^{er} chiffre est différent de zéro. Le tirage a lieu une fois par semaine. Vous achetez un billet au coût de 50 ¢. Le numéro inscrit sur le billet est 125 488.
- Source:** Loto-Québec. Site Internet, 2014.
- Combien de billets différents Loto-Québec peut-elle mettre en circulation chaque semaine pour cette loterie ?
 - Pour gagner le gros lot de 50 000 \$, il faut que le numéro du billet soit identique au numéro tiré par Loto-Québec. Quelles sont les chances de gagner le gros lot si tous les billets sont vendus ?
 - Pour gagner 5 000 \$, il faut que les 5 derniers chiffres du billet, soit 25 488, soient identiques à ceux du numéro tiré par Loto-Québec. Combien de billets ont cette caractéristique ? Quelles sont alors les chances de gagner ce lot ?
 - Pour gagner 250 \$, il faut que les 4 derniers chiffres du billet, soit 5 488, soient identiques à ceux du numéro tiré par Loto-Québec. Combien de billets ont cette caractéristique ? Quelles sont alors les chances de gagner ce lot ?

Calcul de la probabilité d'un événement A à l'aide de la définition

- Classique : $P(A) = \frac{n(A)}{n(S)} = \frac{\text{nombre de résultats favorables à l'événement A}}{\text{nombre de résultats possibles de l'expérience aléatoire}}$
- Empirique : $P(A) = \frac{n_A}{n} = \frac{\text{nombre de fois que l'événement A se produit}}{\text{nombre de répétitions de l'expérience aléatoire}} = \text{fréquence relative}$

Calcul de la probabilité d'un événement particulier

- Événement contraire A'
 $P(A') = 1 - P(A)$
- Union de deux événements $A \cup B$
 - Si l'on peut dénombrer les résultats, on utilise la définition :
$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} \text{ ou } P(A \cup B) = \frac{n_{A \cup B}}{n}$$
 - Si l'on ne peut pas dénombrer les résultats, on utilise la règle d'addition :
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
- Intersection de deux événements $A \cap B$
 - Si l'expérience aléatoire comporte une seule épreuve, on utilise la définition :
$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} \text{ ou } P(A \cap B) = \frac{n_{A \cap B}}{n}$$
 - Si l'expérience aléatoire est décomposable en épreuves successives, on utilise la règle de multiplication :
$$P(A \cap B) = P(A) P(B | A)$$

Calcul d'une probabilité conditionnelle

Probabilité d'obtenir A sachant que B s'est déjà réalisé : $P(A | B) = \frac{n(A \cap B)}{n(B)}$ ou $P(A | B) = \frac{P(A \cap B)}{P(B)}$

Événements indépendants

A et B sont indépendants si $P(A) = P(A | B)$, ce qui signifie que la probabilité de l'événement A reste la même que l'événement B se réalise ou non : l'événement B n'influence pas la probabilité de réalisation de l'événement A.

Démarche de résolution de problèmes de probabilité

1. Assigner une lettre majuscule à chacun des événements en cause.
2. Si la situation décrite semble complexe, rassembler l'information :
 - dans un tableau, si l'expérience aléatoire est simple ;
 - dans un diagramme en arbre, si l'expérience aléatoire est décomposable.
3. Exprimer la question à l'aide des symboles suivants :
 P : «la probabilité de...» ou «les chances d'avoir...»
 \cap : «et» \cup : «ou» (inclusif) $|$: «si» ou «sachant que» ou «étant donné»
4. Calculer la probabilité demandée en utilisant la définition ou les propriétés.

Analyse combinatoire

On peut dénombrer les résultats d'une expérience en utilisant l'une ou l'autre des techniques suivantes :

- Le principe de multiplication

Nombre de résultats d'une suite d'opérations = produit du nombre de choix à chaque opération

À utiliser, entre autres, pour dénombrer des **permutations** (dispositions ordonnées de n éléments différents) et des **arrangements** (dispositions ordonnées de x éléments différents choisis parmi n éléments différents).

- La formule $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ où $0 \leq x \leq n$

À utiliser pour dénombrer des **combinaisons** (dispositions non ordonnées de x éléments différents choisis parmi n éléments différents).

EXERCICES RÉCAPITULATIFS

1. IL LANCE ET COMPTE !

Durant la saison régulière 2013-2014, les Canadiens de Montréal ont gagné 23 des 41 matchs disputés à l'extérieur et ont perdu 36 des 82 matchs de la saison.

Source: Le site officiel des Canadiens de Montréal. 2014.

Exprimer la question en langage mathématique avant d'y répondre. On choisit un match au hasard parmi les 82 matchs disputés.

- Quelle est la probabilité que le match ait été joué à domicile ?
- Si l'équipe n'a pas gagné le match, quelle est la probabilité qu'il ait été disputé à domicile ?
- Quelle est la probabilité que l'équipe ait gagné le match s'il a été disputé à domicile ?
- Quelle est la probabilité que ce soit un match gagné à domicile ?
- Quelle est la probabilité que l'équipe n'ait ni gagné le match ni joué le match à l'extérieur ?
- Quelle est la probabilité que l'équipe ait gagné le match ou ait joué le match à l'extérieur ?
- La probabilité que l'équipe gagne le match dépend-elle de l'endroit où il se joue ? Justifier et interpréter la réponse.

- Une enseignante de 1^{re} année du primaire utilise le jeu du sac à lettres pour favoriser l'apprentissage de l'alphabet. Le sac contient 26 jetons identifiés par une lettre de l'alphabet (6 voyelles et 20 consonnes). À tour de rôle, chaque enfant pige 3 jetons de façon successive et lit à haute voix les lettres qui y sont inscrites.

a) Si la pige s'effectue sans remise, calculer la probabilité que le 1^{er} enfant :

- prélève 3 voyelles.
- prélève d'abord 2 voyelles et ensuite 1 consonne.
- prélève au moins une voyelle.

b) Si la pige s'effectue avec remise, calculer la probabilité que le 1^{er} enfant :

- prélève 3 voyelles.
- prélève d'abord 2 voyelles et ensuite 1 consonne.

c) Si la pige est sans remise et que le 1^{er} enfant a déjà prélevé 2 voyelles et 1 consonne du sac, quelle est la probabilité que le 2^e enfant pige d'abord 2 voyelles, puis 1 consonne parmi les jetons restants ?

d) Si le 1^{er} enfant pige les 3 jetons en une fois parce qu'il n'a pas compris la consigne de piges successives, quelle est la probabilité qu'il prélève 2 voyelles et 1 consonne ?

3. Une enquête révèle les statistiques suivantes sur les Québécois de 15 ans et plus :

- 51,7 % n'ont aucun problème de santé de longue durée ;
- 4,9 % ont au moins un problème de santé de longue durée et n'ont pas de médecin de famille ;
- 78,7 % ont un médecin de famille.

Source: Institut de la statistique du Québec. *Enquête québécoise sur l'expérience de soins 2010-2011, volume 1 et volume 2, mars 2013.*

On choisit un Québécois de 15 ans et plus au hasard.

- a) Calculer la probabilité qu'il n'ait pas de médecin de famille.
 - b) Calculer la probabilité qu'il n'ait ni problème de santé de longue durée ni médecin de famille.
 - c) Sachant qu'il a au moins un problème de santé de longue durée, calculer la probabilité qu'il ait un médecin de famille.
 - d) Calculer la probabilité qu'il n'ait pas de médecin de famille s'il n'a aucun problème de santé de longue durée.
 - e) Le tableau suivant est tiré de l'enquête. Ces statistiques permettent-elles de conclure que la probabilité qu'un Québécois ait un médecin de famille dépend de son âge? Justifier mathématiquement et interpréter la réponse.

Si une tendance se dégage de ces statistiques, l'énoncer.

Proportion des personnes ayant un médecin de famille selon l'âge, Québec, 2010-2011

Âge	Proportion
De 15 à 24 ans (A)	68,7 %
De 25 à 49 ans (B)	70,6 %
De 50 à 64 ans (C)	86,4 %
De 65 à 74 ans (D)	92,6 %
75 ans et plus (E)	95,8 %
Ensemble	78,7 %

4. Quelle est la probabilité qu'un jeune Québécois obtienne un diplôme d'études secondaires avant ses 20 ans?

En 2010, pour une cohorte d'étudiants de même génération composée à 48,7 % de filles, la probabilité d'obtenir un diplôme d'études secondaires avant l'âge de 20 ans était de 80,2 % chez les filles et de 67,1 % chez les garçons.

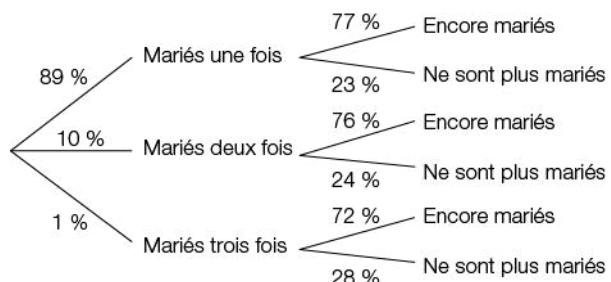
Source: Ministère de l'Éducation, du Loisir et du Sport, et Ministère de l'Enseignement supérieur. *Indicateurs de l'éducation – Édition 2012-2013*.

- a) Quelle est la probabilité qu'un jeune obtienne un diplôme d'études secondaires avant l'âge de 20 ans ?
 - b) Si un jeune obtient un diplôme d'études secondaires :
 - i) quelle est la probabilité que ce soit un garçon ?
 - ii) quelle est la probabilité que ce soit une fille ?
 - c) Si on émet l'hypothèse que la probabilité d'obtention d'un diplôme d'études secondaires ne dépend pas du sexe, quelles devraient être les chances :

- i) qu'une fille obtienne un diplôme d'études secondaires?
 - ii) qu'un garçon obtienne un diplôme d'études secondaires?

5. OUI, JE LE VEUX... LE MARIAGE EST-IL ENCORE POPULAIRE?

Une étude révèle que 80 % des Canadiens de 25 ans et plus se sont mariés au moins une fois. Le diagramme suivant indique quelle était leur situation matrimoniale au moment de l'étude.



Source: Statistique Canada. *Tendances sociales canadiennes*, été 2006.

- a) Si une personne de 25 ans et plus s'est mariée trois fois, quelle est la probabilité qu'elle ne vive plus avec son conjoint ?
 - b) Quelle est la probabilité qu'une personne de 25 ans et plus qui s'est mariée au moins une fois ne vive plus avec son conjoint ?
 - c) Si une personne de 25 ans et plus ne vit plus avec son conjoint, quelle est la probabilité qu'elle se soit mariée deux fois ?
 - d) Compléter le tableau suivant à l'aide des données du diagramme en arbre.

Répartition des Canadiens de 25 ans et plus mariés au moins une fois selon le nombre de mariages et la situation matrimoniale au moment de l'étude

		Situation matrimoniale		
Nombre de mariages		Encore mariés	Ne sont plus mariés	Total
Mariés une fois				
Mariés deux fois				
Mariés trois fois				
Total				100 %

- e) Compléter le tableau de distribution conditionnelle suivant.

Répartition des Canadiens de 25 ans et plus mariés au moins une fois, par nombre de mariages, selon la situation matrimoniale au moment de l'étude

Nombre de mariages	Situation matrimoniale		
	Encore mariés	Ne sont plus mariés	Total
Mariés une fois			100 %
Mariés deux fois			100 %
Mariés trois fois			100 %
Total			100 %

6. a) Combien d'échantillons possibles y a-t-il si l'on prélève sans remise un échantillon aléatoire de 5 personnes dans un groupe composé de 10 locataires et 6 propriétaires ?

- b) Quelle est la probabilité de piger un échantillon qui comprend 3 locataires et 2 propriétaires ?

7. Un système informatique est protégé par un mot de passe composé de 5 caractères différents : 3 lettres suivies de 2 chiffres. Une personne mal intentionnée cherche à pénétrer dans le système informatique. Quelle est la probabilité qu'elle découvre le code en un seul essai :

- a) par hasard ?
 b) si elle sait que les 3 lettres sont des voyelles et que le 1^{er} chiffre est 0 ?
 c) si elle sait que les 3 lettres forment une des permutations des lettres C, D, E ?

PRÉPARATION À L'EXAMEN

Pour préparer votre examen, assurez-vous d'avoir les compétences suivantes :

		Si vous avez la compétence, cochez.
Probabilité		
• Décrire un espace échantillonnaux et un événement en compréhension et en extension.	<input type="radio"/>	
• Calculer une probabilité à l'aide de la définition classique ou empirique.	<input type="radio"/>	
• Calculer une probabilité à l'aide des propriétés des probabilités.	<input type="radio"/>	
Probabilité conditionnelle		
• Calculer une probabilité conditionnelle.	<input type="radio"/>	
• Déterminer la dépendance ou l'indépendance de deux événements, et interpréter le résultat.	<input type="radio"/>	
• Résoudre un problème de probabilité en appliquant la règle de multiplication.	<input type="radio"/>	
Analyse combinatoire		
• Appliquer le principe de multiplication pour dénombrer des résultats.	<input type="radio"/>	
• Dénombrer des combinaisons.	<input type="radio"/>	
• Calculer une probabilité en utilisant les techniques de dénombrement.	<input type="radio"/>	



3

Chapitre

Les lois de probabilité



OBJECTIFS DU CHAPITRE

- Construire et analyser la distribution d'une variable aléatoire.
- Utiliser les lois de probabilité binomiale, de Poisson et normale pour résoudre des problèmes concrets.

OBJECTIF DU LABORATOIRE

Le laboratoire 3 vise à apprendre à utiliser Excel pour calculer une probabilité à l'aide des lois binomiale, de Poisson ou normale.



Le présent chapitre est consacré à l'étude de trois lois de probabilité particulièrement importantes en inférence statistique : la loi binomiale, la loi de Poisson et la loi normale.

3.1 Les variables aléatoires

Dans cette section, nous examinons les expériences aléatoires dont les résultats peuvent être associés à des valeurs numériques. La mise en situation ci-dessous, basée sur l'idée d'un sondage d'opinion politique, illustre ce type d'expérience aléatoire.

Bien que, généralement, un sondage d'opinion s'effectue auprès d'un échantillon d'environ 1 000 personnes, nous travaillerons pour le moment avec des échantillons de petite taille. Progressivement, au fil de l'acquisition des connaissances, nous serons en mesure d'augmenter la taille des échantillons pour atteindre, finalement, 1 000.

MISE EN

SITUATION

Supposons que, lors d'une élection au Québec, les votes se sont répartis comme suit :

Parti libéral (L) : 55 % Parti québécois (Q) : 40 % Autres partis (A) : 5 %

Six mois après l'élection, on effectue un sondage pour connaître, entre autres choses, le taux de satisfaction des électeurs à l'égard du nouveau gouvernement en fonction du parti qu'ils ont appuyé lors du dernier scrutin.

Pour ce faire, un échantillon aléatoire de 3 électeurs est prélevé avec remise parmi l'ensemble des électeurs. Après avoir demandé aux répondants d'indiquer leur satisfaction à l'égard du nouveau gouvernement, on leur a posé la question suivante :

« Pour quel parti avez-vous voté aux dernières élections ? »

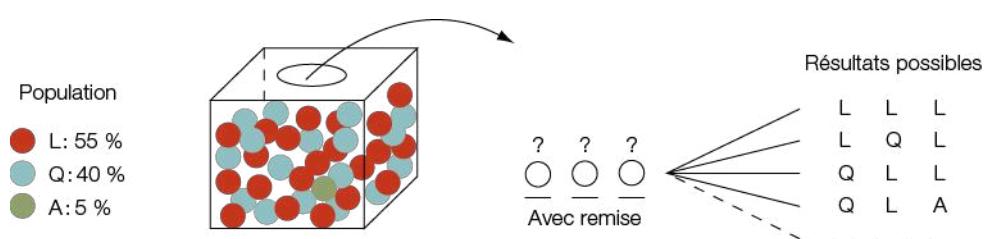
Voici des exemples de réponses possibles si les 3 électeurs sondés ont répondu à la question :

- les 3 électeurs ont voté pour le Parti libéral (L L L) ;
- le 1^{er} électeur pigé a voté pour le Parti libéral, le 2^e, pour le Parti québécois et le 3^e, pour le Parti libéral (L Q L) ;
- le 1^{er} et le 2^e électeur pigés ont voté pour le Parti québécois et le 3^e, pour le Parti libéral (Q Q L) ;
- etc.

En respectant l'ordre des réponses, les résultats possibles de cette expérience aléatoire sont :

$$S = \{(L L L), (L Q L), (Q L L), (Q L A), \dots\}$$

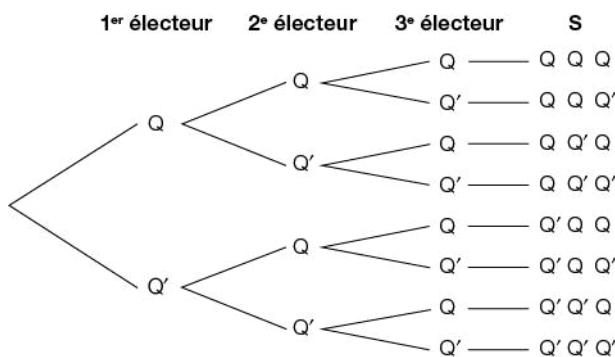
$$[n(S) = 3 \times 3 \times 3 = 27 \text{ résultats possibles}]$$



L'intérêt de ce type d'expérience aléatoire ne tient pas tant à l'ordre des réponses obtenues [(Q Q L), (Q L Q) ou (L Q Q)] qu'à la possibilité de déterminer le nombre d'électeurs qui ont voté pour chacun des partis. En effet, il est plus intéressant de connaître la probabilité que deux électeurs sur trois aient voté pour le Parti québécois que la probabilité que ces deux électeurs soient, par exemple, les deux premières personnes pigées.

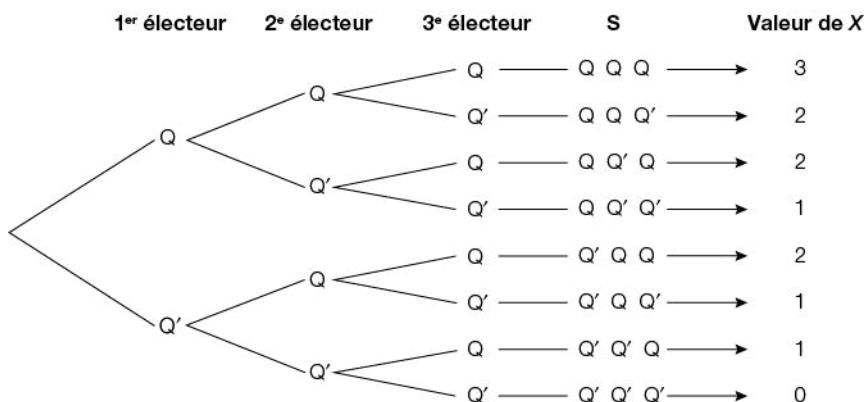
3.1.1 Le concept de variable aléatoire

Dans le contexte de la mise en situation, supposons que nous voulions déterminer si l'électeur pigé a voté pour le Parti québécois ou non. Le diagramme en arbre suivant présente tous les résultats possibles de l'expérience aléatoire en fonction des événements Q: «l'électeur a voté pour le Parti québécois» et Q': «l'électeur n'a pas voté pour le Parti québécois».



Pour faciliter l'analyse des résultats possibles, associons une valeur numérique à chaque élément de S. Cette valeur, notée X , est définie ainsi :

X : «nombre d'électeurs ayant voté pour le Parti québécois dans un échantillon de 3 électeurs».



On observe que la variable X peut prendre les valeurs 0, 1, 2, 3.

Comme la valeur obtenue pour X résulte d'un tirage, on dit que X est une variable aléatoire.

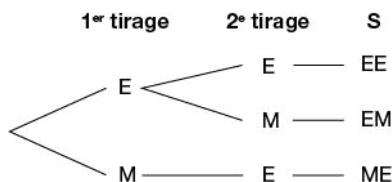
Variable aléatoire

Une **variable aléatoire** est une fonction qui associe une valeur numérique à chacun des résultats possibles d'une expérience aléatoire. On emploie les majuscules X , Y ou Z pour désigner une variable aléatoire et les minuscules x , y ou z pour représenter une valeur de la variable.

Une variable aléatoire est dite **continue** si elle peut en théorie prendre n'importe laquelle des valeurs contenues dans un intervalle donné de nombres réels (par exemple, X : «la taille d'un étudiant choisi au hasard dans la classe»); sinon, elle est dite **discrète** (comme dans la mise en situation).

EXEMPLE

Un groupe est composé de 3 élèves de 10 ans et de 1 moniteur âgé de 40 ans. On pique sans remise 2 personnes dans le groupe. Le diagramme en arbre présente les résultats possibles de cette expérience aléatoire en fonction des événements E : «la personne pigée est un élève» et M : «la personne pigée est un moniteur».



Soit les variables aléatoires suivantes :

X : «le nombre d'élèves parmi les 2 personnes pigées»;

Y : «le nombre de moniteurs parmi les 2 personnes pigées»;

Z : «la moyenne d'âge des personnes pigées».

Déterminer la valeur des variables aléatoires X , Y et Z associée à chaque résultat de l'expérience aléatoire, et dire si la variable aléatoire est discrète ou continue.

Solution

1 ^e tirage	2 ^e tirage	S	Valeur de X	Valeur de Y	Valeur de Z
E	E	EE →	—	—	—
E	M	EM →	—	—	—
M	E	ME →	—	—	—

Type de variable aléatoire : _____

3.1.2 La distribution de probabilité

MISE EN

SITUATION (suite)

Reprendons la mise en situation qui consiste à piger, avec remise, un échantillon de 3 personnes dans une population d'électeurs ayant voté à 55 % pour le Parti libéral, à 40 % pour le Parti québécois et à 5 % pour un autre parti.

Considérons de nouveau la variable aléatoire X : «nombre d'électeurs ayant voté pour le Parti québécois dans un échantillon de 3 électeurs». On sait que les valeurs possibles pour X sont 0, 1, 2, 3.

Considérant que 40 % des électeurs de la population ont voté pour le Parti québécois aux élections, selon vous, laquelle des valeurs de X a le plus de chances d'être obtenue ?

- 0 ; aucun électeur (0 % de l'échantillon) n'a voté pour le Parti québécois.
- 1 ; un seul électeur (33 % de l'échantillon) a voté pour le Parti québécois.
- 2 ; deux électeurs (67 % de l'échantillon) ont voté pour le Parti québécois.
- 3 ; tous les électeurs (100 % de l'échantillon) ont voté pour le Parti québécois.

Intuitivement, on peut penser que la valeur 1 est celle qui a plus de chances d'être obtenue. Est-ce le cas ?

La seule façon de confirmer cette intuition est de calculer la probabilité d'obtenir chacune des valeurs de la variable aléatoire X . La fonction de probabilité définie ci-dessous va permettre d'effectuer ces calculs.

Fonction de probabilité

La **fonction de probabilité**, notée $f(x)$, associe à chaque valeur x de la variable aléatoire X la probabilité d'obtenir cette valeur, soit $P(X = x)$. On définit cette fonction ainsi :

Fonction de probabilité

$$f(x) = P(X = x) \quad \text{où} \quad x = 0, 1, 2, 3, \dots$$

EXEMPLE 1

Calculer $f(1)$ pour la variable aléatoire X définie dans la mise en situation.

Solution

Puisque $f(1) = P(X = 1)$, il faut calculer cette dernière probabilité.

Le diagramme en arbre présenté à la page 133 montre que la variable aléatoire X prend la valeur 1 si l'on prélève avec remise l'un ou l'autre des échantillons suivants : $Q Q' Q'$, $Q' Q Q'$ ou $Q' Q' Q$.

Il semble logique d'additionner les probabilités respectives de ces échantillons pour obtenir la probabilité que $X = 1$.

Soit l'événement $A = \{Q Q' Q', Q' Q Q', Q' Q' Q\}$. On dit que A est l'événement antécédent de 1, car il contient les résultats de l'expérience aléatoire associés à la valeur 1. On peut donc écrire :

$$f(1) = P(X = 1) = P(A) \quad \text{où } A \text{ est l'événement antécédent de 1.}$$

Calculons $P(A)$ en appliquant la règle de multiplication. Il faut tenir compte du fait que les tirages sont indépendants, puisqu'ils sont effectués avec remise.

$$\begin{aligned} P(A) &= P(Q \cap Q' \cap Q') + P(Q' \cap Q \cap Q') + P(Q' \cap Q' \cap Q) \\ &= (0,4 \times 0,6 \times 0,6) + (0,6 \times 0,4 \times 0,6) + (0,6 \times 0,6 \times 0,4) \\ &= 0,144 + 0,144 + 0,144 \\ &= 3 \times 0,144 \\ &= 43,2 \% \end{aligned}$$

$$f(1) = P(X = 1) = P(A) = 43,2 \%$$

Événement antécédent de x

L'**événement antécédent de x** est un événement qui contient tous les résultats d'une expérience aléatoire associés à la valeur x d'une variable aléatoire X .

Distribution de probabilité

Une **distribution de probabilité** est un tableau qui donne la fonction de probabilité $f(x)$ associée à chaque valeur x d'une variable aléatoire X .

EXEMPLE 1 (suite)

Construire la distribution de probabilité de la variable aléatoire X définie dans la mise en situation.

Solution

Dans l'exemple précédent, on a calculé $f(1)$. La même procédure s'applique pour évaluer $f(0)$, $f(2)$ et $f(3)$, soit déterminer la probabilité de réalisation de l'événement antécédent associé à chaque valeur.

	x	$f(x) = P(X=x)$	Probabilité de l'événement antécédent de x
Événement antécédent de 0 Q' Q' Q'	0	21,6 %	$0,6 \times 0,6 \times 0,6$
Événement antécédent de 1 Q Q' Q' Q' Q Q' Q' Q' Q	1	43,2 %	$0,4 \times 0,6 \times 0,6$ $0,6 \times 0,4 \times 0,6$ $0,6 \times 0,6 \times 0,4$
Événement antécédent de 2 Q Q Q' Q Q' Q Q' Q Q	2	28,8 %	$0,4 \times 0,4 \times 0,6$ $0,4 \times 0,6 \times 0,4$ $0,6 \times 0,4 \times 0,4$
Événement antécédent de 3 Q Q Q	3	6,4 %	$0,4 \times 0,4 \times 0,4$

On présente ainsi la distribution de probabilité de la variable aléatoire X :

Distribution de probabilité de X : «nombre d'électeurs ayant voté pour le Parti québécois dans un échantillon de 3 électeurs»

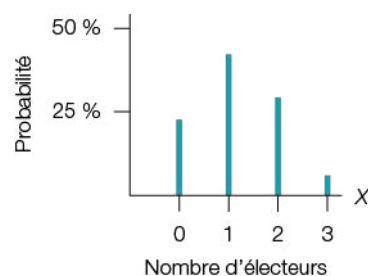
x	0	1	2	3	Total
$f(x)$	21,6 %	43,2 %	28,8 %	6,4 %	100,0 %

On peut représenter graphiquement la distribution de probabilité d'une variable aléatoire discrète par un diagramme en bâtons.

Analyse de la distribution de probabilité

Si l'on pique avec remise 3 électeurs dans une population où 40 % des électeurs ont voté pour le Parti québécois, le résultat le plus probable (43 %) est qu'un seul des 3 électeurs ait voté pour ce parti. Le résultat le moins probable (6 %) est que les 3 électeurs aient voté pour ce parti.

Distribution de probabilité de X : «nombre d'électeurs ayant voté pour le Parti québécois dans un échantillon de 3 électeurs»



Propriétés d'une fonction de probabilité

Une fonction de probabilité possède les propriétés suivantes :

$$0 \leq f(x) \leq 1 \quad [\text{Exprimé en pourcentage : } 0 \% \leq f(x) \leq 100 \%]$$

$$\sum f(x) = 1 \quad [\text{Exprimé en pourcentage : } \sum f(x) = 100 \%]$$

Prélude au test d'hypothèse

Lors d'un test d'hypothèse (notion qui sera abordée au chapitre 5), c'est sur la base de la distribution de probabilité qu'on prend la décision de rejeter ou de ne pas rejeter une hypothèse. L'exemple suivant illustre le raisonnement qu'on tient dans ce cas.

Un an après les élections où le Parti québécois a remporté 40 % des votes, on prélève avec remise un échantillon aléatoire de 8 électeurs. Si 6 des 8 électeurs, soit 75 %, déclarent qu'ils voteront pour le Parti québécois s'il y avait des élections le jour même, doit-on en conclure que le pourcentage d'électeurs qui appuient le Parti québécois dans la population a augmenté ?

Dans l'hypothèse où le pourcentage d'appuis au Parti québécois est encore de 40 % dans la population au moment du prélèvement de l'échantillon, on a la distribution de probabilité suivante pour la variable aléatoire X : «nombre de sympathisants du Parti québécois dans un échantillon de 8 électeurs» :

Distribution de probabilité de X : «nombre de sympathisants du Parti québécois dans un échantillon de 8 électeurs»

x	0	1	2	3	4	5	6	7	8	Total
$f(x)$	1,7 %	9,0 %	20,9 %	27,9 %	23,2 %	12,4 %	4,1 %	0,8 %	0,1 %	100,1 %

Théoriquement, selon cette hypothèse, il y a seulement 4 % de chances de prélever un échantillon comptant 6 électeurs qui appuient le Parti québécois. Une probabilité aussi faible conduit à l'une ou l'autre des conclusions suivantes :

- ou bien le hasard a donné un échantillon très rare pour une population qui compte 40 % de sympathisants du Parti québécois ;
- ou bien le pourcentage de sympathisants du Parti québécois dans la population a augmenté et il n'est plus de 40 %.

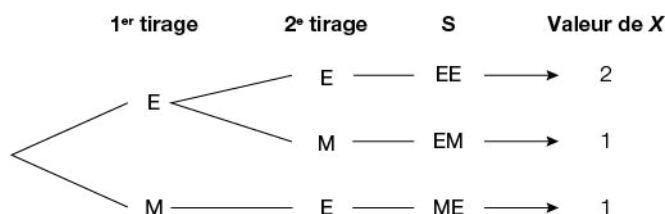
La deuxième déduction semble la plus plausible.

Il va de soi qu'un échantillon de plus grande taille permettrait de décider avec une plus grande certitude si le pourcentage de sympathisants du Parti québécois dans la population a augmenté de façon significative.

EXEMPLE 2

On pige, sans remise, 2 personnes dans un groupe formé de 3 élèves et de 1 moniteur. Soit la variable aléatoire X : «nombre d'élèves parmi les 2 personnes pigées». Construire le tableau de distribution de probabilité de X .

Solution



Calculons la fonction de probabilité pour chaque valeur de X .

Soit les événements E : «piger un élève» et M : «piger un moniteur».

- L'événement antécédent de 1 est A = {EM, ME} ; donc,

$$f(1) = P(X=1) = P(A)$$

$$= P(E \cap M) + P(M \cap E)$$

$$= P(E) P(M | E) + P(M) P(E | M) \quad [\text{tirages dépendants, car pige sans remise}]$$

$$= \left(\frac{3}{4} \times \frac{1}{3}\right) + \left(\frac{1}{4} \times \frac{3}{3}\right) = \frac{3}{12} + \frac{3}{12} = \frac{6}{12} \quad [\text{selon la règle de multiplication}]$$

$$= 0,5 = 50 \%$$

- L'événement antécédent de 2 est B = {EE} ; donc,

$$f(2) = P(X=2) = P(B)$$

$$= P(E \cap E)$$

$$= \frac{3}{4} \times \frac{2}{3} = \frac{6}{12} = \frac{1}{2}$$

$$= 0,5 = 50 \%$$

**Distribution de probabilité de X:
«nombre d'élèves parmi les 2 personnes pigées»**

x	1	2	Total
f(x)	50 %	50 %	100 %

Analyse de la distribution de probabilité

Si l'on pige sans remise 2 personnes dans une population constituée de 3 élèves et de 1 moniteur, il y a autant de chances de piger 1 élève que d'en piger 2.

EXEMPLE 3

En 2012, 44 % des nouveau-nés sont de rang 1 (1^{er} enfant de la mère). On note le rang de naissance de 2 enfants nés la même journée dans un hôpital. Compléter la distribution de probabilité de la variable aléatoire X: «nombre d'enfants de rang 1 parmi 2 nouveau-nés».

Source: Institut de la statistique du Québec. *Naissances selon le rang de naissance, 2012*.

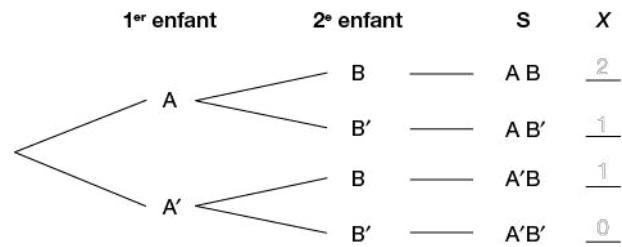
Solution

**Distribution de probabilité de X:
«nombre d'enfants de rang 1 parmi 2 nouveau-nés»**

x		Total
f(x)		

Soit les événements A : «le 1^{er} enfant est de rang 1» et B : «le 2^e enfant est de rang 1».

Comme le rang du 2^e nouveau-né ne peut pas dépendre du rang du 1^{er} nouveau-né, les événements A et B sont indépendants.



EXERCICE DE COMPRÉHENSION | 3.1

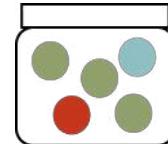
Un bocal contient 5 boules : 3 vertes, 1 rouge, 1 bleue. Un jeu consiste à prélever, sans remise, une boule jusqu'à ce que l'on obtienne une boule verte. Construire la distribution de probabilité de la variable aléatoire suivante :

X : «nombre de tirages nécessaires pour obtenir une boule verte».

Solution

Distribution de probabilité de X : «nombre de tirages nécessaires pour obtenir une boule verte»

x		Total
f(x)		



3.1.3 L'espérance et l'écart type d'une variable aléatoire

Lorsque l'on étudie une variable en statistique descriptive, on construit le tableau de distribution, puis on calcule la moyenne et l'écart type de la distribution. On procède de même pour une variable aléatoire : après avoir construit la distribution de probabilité, on calcule la moyenne et l'écart type de la distribution de probabilité. La mise en situation présentée au début de la section sert également à illustrer le calcul et l'interprétation de ces deux mesures.

MISE EN

SITUATION (suite)

Le tableau suivant donne la distribution de probabilité du nombre d'électeurs ayant voté pour le Parti québécois parmi 3 électeurs pigés avec remise dans une population où 40 % des électeurs ont voté pour ce parti aux dernières élections.

Distribution de probabilité de X : «nombre d'électeurs ayant voté pour le Parti québécois dans un échantillon de 3 électeurs»

x	0	1	2	3	Total
$f(x)$	21,6 %	43,2 %	28,8 %	6,4 %	100,0 %

Si l'on répète cette expérience aléatoire un très grand nombre de fois, par exemple 1 000 fois, et que l'on note chaque fois la valeur de la variable aléatoire X , on obtient 1 000 nombres dont la valeur est soit 0, 1, 2 ou 3. Deux questions se posent plus particulièrement :

- Quel pourcentage de 0, de 1, de 2 ou de 3 peut-on espérer trouver parmi ces 1 000 nombres ?
- Si l'on calcule la moyenne et l'écart type des 1 000 nombres notés, quelles valeurs peut-on espérer trouver ?

Puisque la distribution de probabilité de X indique qu'il y a 21,6 % de chances de piger un échantillon de trois électeurs où $X = 0$, il est logique d'espérer trouver environ 21,6 % de 0 parmi les 1 000 nombres notés. En appliquant le même raisonnement pour les autres valeurs de X , on est théoriquement justifié d'espérer que la répartition des 1 000 échantillons selon les valeurs de X soit semblable à la distribution de probabilité de X , ce qui donne le tableau suivant.

Répartition espérée de 1 000 échantillons de 3 électeurs selon le nombre de sympathisants du Parti québécois

Nombre de sympathisants	0	1	2	3	Total
Pourcentage espéré	21,6 %	43,2 %	28,8 %	6,4 %	100,0 %
Effectif espéré	216	432	288	64	1 000

À partir de ce tableau, il est facile de calculer la moyenne et l'écart type de la distribution des valeurs de X en utilisant le mode statistique de la calculatrice ou les formules présentées au chapitre 1. Comme la distribution repose sur un modèle probabiliste, la moyenne est appelée **espérance** et on la note $E(X)$ ou μ , alors que l'**écart type** est noté σ . Voici comment on calcule ces mesures :

- Espérance

$$E(X) = 0 \times 21,6 \% + 1 \times 43,2 \% + 2 \times 28,8 \% + 3 \times 6,4 \% = 1,2$$

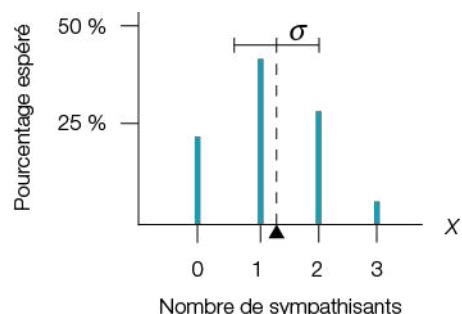
- Écart type

$$\sigma = \sqrt{(0 - 1,2)^2 \times 21,6 \% + (1 - 1,2)^2 \times 43,2 \% + (2 - 1,2)^2 \times 28,8 \% + (3 - 1,2)^2 \times 6,4 \%} = 0,8$$

Interprétation des mesures

Si l'on répète l'expérience aléatoire un grand nombre de fois, on peut espérer trouver en moyenne 1,2 sympathisant du Parti québécois par échantillon avec un écart type de 0,8 sympathisant. Donc, la plupart des échantillons compteront 1 ou 2 sympathisants du Parti québécois. (Les nombres 1 et 2 sont les entiers compris dans l'intervalle $[E(X) - \sigma; E(X) + \sigma] = [0,4; 2]$.)

Répartition espérée de 1 000 échantillons de 3 électeurs selon le nombre de sympathisants du Parti québécois (modèle théorique)



Espérance et écart type d'une variable aléatoire

$$\text{Espérance : } E(X) = \sum x f(x)$$

$$\text{Écart type : } \sigma = \sqrt{\sum [x - E(X)]^2 f(x)}$$

Interprétation de l'espérance et de l'écart type

Si l'on répète une expérience aléatoire un grand nombre de fois, on peut espérer une moyenne égale à $E(X)$ et un écart type égal à σ pour les valeurs de la variable aléatoire X . Donc, pour la plupart des expériences aléatoires on obtiendra une valeur de X comprise entre $[E(X) - \sigma]$ et $[E(X) + \sigma]$.

EXEMPLE

En 2014, 45 % des ménages québécois sont équipés d'une tablette numérique. On prélève avec remise un échantillon aléatoire de 7 ménages. À l'aide des données du tableau suivant, calculer et interpréter l'espérance et l'écart type de la variable aléatoire X : «nombre de ménages équipés d'une tablette numérique dans un échantillon de 7 ménages».

Source: CEFRIQ. NETendances 2014: Équipement et branchement Internet des foyers québécois, vol. 5, n° 2.

Distribution de probabilité de X : «nombre de ménages équipés d'une tablette numérique dans un échantillon de 7 ménages»

x	0	1	2	3	4	5	6	7	Total
$f(x)$	1,5 %	8,7 %	21,4 %	29,2 %	23,9 %	11,7 %	3,2 %	0,4 %	100,0 %

Solution

En utilisant le mode statistique de la calculatrice, on obtient :

Espérance : $E(X) = 3,2$ ménages

Écart type : $\sigma = 1,3$ ménage

Interprétation

Si l'on étudie un grand nombre d'échantillons de 7 ménages, on peut espérer trouver en moyenne 3,2 ménages équipés d'une tablette numérique avec un écart type de 1,3 ménage. La plupart des échantillons comprendront 2, 3 ou 4 ménages équipés d'une tablette numérique. (Les nombres 2, 3 et 4 sont les entiers compris dans l'intervalle $[1,9; 4,5]$.)

EXERCICE DE COMPRÉHENSION | 3.2

Calculer et interpréter l'espérance et l'écart type de la distribution de la variable aléatoire X , si celle-ci est définie comme étant le nombre de tirages sans remise nécessaires pour piger une boule verte dans un bocal contenant 5 boules : 3 vertes, 1 rouge, 1 bleue.

Distribution de probabilité de X : «nombre de tirages sans remise nécessaires pour obtenir une boule verte»

x	1	2	3	Total
$f(x)$	60 %	30 %	10 %	100,0 %

Solution

Interprétation

Si l'on répète l'expérience un grand nombre de fois, il faudra en moyenne _____ tirage pour obtenir une boule verte, avec un écart type de _____ tirage. Donc, lors de la plupart des expériences, il faut _____ ou _____ tirages pour obtenir une boule verte.

3.1.4 L'espérance et les jeux de hasard

Les exemples suivants illustrent l'application de l'espérance à des jeux de hasard. On utilise également ce concept pour calculer l'espérance de vie et déterminer des primes d'assurance.

EXAMPLE 1

Une loterie émet chaque semaine 1 000 billets vendus 0,50 \$ chacun. Il y a trois prix à gagner: 50 \$, 20 \$ et 10 \$. Calculer et interpréter l'espérance de la variable aléatoire X : «gain net d'un joueur qui achète un billet».

Solution

Pour déterminer l'espérance de X , il faut construire la distribution de probabilité de X .

Distribution de probabilité de X : «gain net d'un joueur qui achète un billet»

x					Total
$f(x)$					1

$$E(X) = \underline{\hspace{2cm}}$$

Interprétation de l'espérance de gain net

Un joueur qui participe souvent à cette loterie perdra en moyenne 0,42 \$ par billet: s'il participe 100 fois, il perdra en moyenne 42 \$, soit 100 fois $E(X)$. Cette loterie n'est pas équitable; en fait, aucune ne l'est. Lorsqu'un jeu est équitable, $E(X) = 0$.

EXEMPLE 2

Un comité d'étudiants organise le jeu de hasard suivant pour le financement du bal de fin d'études.

Chaque joueur lance un dé. Si le résultat est 6, il gagne 2 \$; dans les autres cas, il doit payer une somme telle que l'espérance du gain net soit favorable au comité. Quelle somme doit payer chaque perdant si l'espérance de gain net par partie est fixée à 0,50 \$ en faveur du comité ?

Solution

Soit la variable aléatoire X : «gain net du comité par partie». Soit k \$, la somme qu'un joueur doit payer quand il perd. Comme l'événement antécédent de -2 \$ est {6} et que celui de k \$ est {1, 2, 3, 4, 5}, on a la distribution de probabilité suivante :

Distribution de probabilité de X : «gain net du comité par partie»

x	-2 \$	k \$	Total
$f(x)$	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{6}{6}$

$$E(X) = -2 \times \frac{1}{6} + k \times \frac{5}{6} = 0,50 \text{ \$}$$

$$\frac{-2 + 5k}{6} = 0,50 \text{ \$}$$

$$-2 + 5k = 3 \text{ \$}$$

$$5k = 5 \text{ \$}$$

$$k = 1 \text{ \$}$$

Interprétation

Chaque joueur qui tire un autre résultat que 6 doit payer 1 \$ pour que l'espérance de gain net par partie soit de 0,50 \$ en faveur du comité. Si 400 étudiants jouent à ce jeu, le comité peut espérer amasser la somme de 200 \$ pour le bal de fin d'études :

$$\text{Gain net espéré pour 400 parties} = 400 E(X) = 400 \times 0,50 = 200 \text{ \$}$$

EXEMPLE 3

Un marchand d'appareils électroniques offre des garanties prolongées sur certains produits. Par exemple, à l'achat d'un téléviseur, il en coûte 100 \$ de plus pour prolonger de deux ans la garantie d'un an déjà offerte par le fabricant. L'expérience démontre que seulement 8 % des clients font une réclamation durant la période de garantie supplémentaire et que le coût moyen des réparations est de 275 \$. Pour chaque garantie prolongée vendue, quelle est l'espérance de gain net du marchand ? Compléter l'interprétation de la réponse.

Solution

Distribution de probabilité de X : «gain net du marchand par client»

x		Total
$f(x)$		1

Interprétation

Si le marchand vend un grand nombre de garanties prolongées à ses clients, il peut espérer un gain net de _____ \$ par client en moyenne.

EXERCICE DE COMPRÉHENSION | 3.3

Un bocal contient 3 boules vertes, 1 boule rouge et 1 boule bleue. Le jeu consiste à piger sans remise une boule dans le bocal jusqu'à ce qu'on tire une boule verte. Chaque joueur doit payer 2 \$ pour participer au jeu ; on lui remet 1 \$ s'il pige une boule verte en deux tirages, et 4 \$ s'il pige une boule verte en trois tirages. Calculer et interpréter l'espérance de gain net d'un joueur.

Solution

Distribution de probabilité de X : «gain net du joueur»

x				Total
$f(x)$				100 %

Interprétation

Si l'on joue plusieurs fois à ce jeu, il peut arriver que l'on gagne quelquefois, mais, en moyenne, on aura perdu _____ par partie ; si l'on joue 100 fois, les pertes seront en moyenne de _____.

EXERCICES 3.1

1. Un groupe est composé de 3 hommes et 2 femmes. Pour chacune des expériences aléatoires suivantes, construire la distribution de probabilité de X : «nombre de femmes pigées».
 - a) On pige 1 personne dans le groupe.
 - b) On pige sans remise 2 personnes dans le groupe.
 - c) On pige avec remise 2 personnes dans le groupe.
 2. Dans un groupe composé de 2 hommes, 2 femmes et 1 enfant, on pige sans remise une personne jusqu'à ce que l'on obtienne une femme. Construire la distribution de probabilité de la variable aléatoire X : «nombre de tirages nécessaires pour obtenir une femme».
 3. En 2013, 68 % des Québécois de 18 à 24 ans possèdent un téléphone intelligent. On pige avec remise 3 jeunes dans ce groupe d'âge et l'on s'intéresse au nombre de jeunes qui ont un téléphone intelligent.

Source: CEFRIQ. NETendances 2013: La mobilité au Québec: une montée en flèche, vol. 4, n° 7.

 - a) Définir la variable aléatoire et donner ses valeurs.
 - b) Construire la distribution de probabilité de la variable aléatoire.
- c) Quelle est la probabilité que 2 des 3 jeunes aient un téléphone intelligent?
- d) Quelle est la probabilité que les 3 jeunes aient un téléphone intelligent?
- e) Calculer l'espérance et l'interpréter.
4. Calculer l'espérance et l'écart type de la variable aléatoire définie au numéro 2, et interpréter ces mesures.
5. Un homme et une femme gagnent leur vie en vendant des abonnements à un magazine. Ils reçoivent un salaire de base de 200 \$ par semaine auquel s'ajoute un montant de 6 \$ par abonnement vendu. Voici, pour l'homme et la femme, la distribution de probabilité de la variable aléatoire X : «nombre d'abonnements vendus par semaine».
- Pour la femme
- Distribution de probabilité de X : «nombre d'abonnements vendus par semaine»
- | x | [15; 20[| [20; 25[| [25; 30[| [30; 35[| Total |
|--------|----------|----------|----------|----------|-------|
| $f(x)$ | 8 % | 12 % | 42 % | 38 % | 100 % |

Pour l'homme

Distribution de probabilité de X :
«nombre d'abonnements vendus par semaine»

x	[15; 20[[20; 25[[25; 30[[30; 35[Total
$f(x)$	14 %	15 %	38 %	33 %	100 %

- a) Comparer les chances que chacun a de vendre au moins 25 abonnements par semaine.
- b) Calculer et interpréter l'espérance et l'écart type de la variable aléatoire X :
- i) pour la femme.
 - ii) pour l'homme.
- c) i) Quel revenu moyen la femme peut-elle espérer gagner par semaine ?
- ii) Quel revenu moyen l'homme peut-il espérer gagner par semaine ?
6. On lance simultanément un dé rouge et un dé blanc.
- a) Décrire en extension l'espace échantillonnaux S de cette expérience aléatoire.
- b) Indiquer quelles valeurs peut prendre la variable aléatoire X : «total des points obtenus».

c) Déterminer la valeur de $f(7)$.

d) Construire la distribution de probabilité de X .

e) Calculer et interpréter l'espérance de X .

f) Indiquer quelles valeurs peut prendre la variable Y : «le plus élevé des deux nombres obtenus ou le nombre de points par dé s'il s'agit d'une paire».

g) Construire le distribution de probabilité de Y .

7. Un jeu de hasard coûte 3 \$ par partie. Il consiste à piger une carte dans un jeu de 52 cartes. Si l'on tire un cœur, on gagne 5 \$, et si l'on tire un carreau, on gagne 4 \$. Le tirage d'une carte noire ne rapporte rien. Calculer et interpréter l'espérance de la variable X : «gain net d'un joueur».

8. Un jeu consiste à lancer simultanément deux pièces de monnaie. On gagne 3 \$ si l'on obtient deux fois face et 1 \$ si l'on obtient une seule face. Par contre, il faut débourser k \$ si l'on n'obtient aucune face. Si l'on veut que le jeu soit équitable, c'est-à-dire que $E(X) = 0$ \$ où X est le gain net d'un joueur, quelle doit être la valeur de k ?

3.2 La loi binomiale

Dans la présente section, nous étudions les variables aléatoires discrètes dont la distribution de probabilité s'apparente à un modèle binomial.

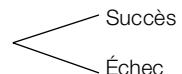
3.2.1 Le contexte d'une expérience aléatoire binomiale

Dans plusieurs expériences aléatoires, les résultats possibles sont de nature dichotomique : on obtient un «succès» ou un «échec». Voici quelques exemples.

Si l'on choisit une personne au hasard :

- elle est de sexe «masculin» ou de sexe «féminin»;
- elle «a voté» ou elle «n'a pas voté» pour le parti A aux dernières élections ;
- elle «est mariée» ou elle «n'est pas mariée» ;
- elle «a des enfants» ou elle «n'a pas d'enfants».

Résultats de l'épreuve



Ces situations décrivent un contexte où le modèle binomial est approprié pour le calcul des probabilités. Voici les critères auxquels doit satisfaire une expérience pour être considérée comme une expérience aléatoire binomiale.

Contexte d'une expérience aléatoire binomiale

Une expérience aléatoire est dite **expérience binomiale** si elle satisfait aux critères suivants :

1. L'expérience est constituée d'une suite de n épreuves indépendantes.
Cela signifie que le résultat d'une épreuve ne dépend aucunement du résultat de l'épreuve précédente.
2. Deux résultats possibles sont associés à chaque épreuve : « succès » ou « échec ».
 - La probabilité d'un succès, notée p , est la même à chaque épreuve.
 - La probabilité d'un échec, notée q , est la même à chaque épreuve : $q = 1 - p$.
3. La variable aléatoire X correspond au nombre de succès en n épreuves.

Si l'expérience aléatoire satisfait aux trois critères, on dit que la variable aléatoire X suit une loi binomiale de paramètres n et p , ce qui s'écrit :

$$X \text{ suit une } B(n; p) \quad \text{où} \quad X = 0, 1, 2, 3, \dots, n$$

EXEMPLE

Les expériences aléatoires suivantes sont-elles des expériences binomiales ?

- a) Une expérience aléatoire consiste à lancer six fois une pièce de monnaie. On étudie la variable aléatoire X : « nombre de faces obtenues ».

Solution

1. On effectue six lancers (ou épreuves) indépendants : $n = 6$.
2. À chaque lancer, il y a deux résultats possibles :
 - succès : « obtenir face », dont la probabilité est $p = 0,5$;
 - échec : « ne pas obtenir face », dont la probabilité est $q = 0,5$.
3. La variable aléatoire X correspond au nombre de succès en six épreuves.

Donc, X suit une $B(6; 0,5)$ où $X = 0, 1, 2, 3, 4, 5, 6$.

- b) On prélève avec remise un échantillon aléatoire de 4 personnes dans un groupe de 20 étudiants parmi lesquels 15 ont un permis de conduire. La variable aléatoire X est définie comme le nombre d'étudiants de l'échantillon qui détiennent un permis de conduire.

Solution

- c) On prélève sans remise un échantillon aléatoire de 4 personnes parmi 10 filles et 8 garçons. On s'intéresse au nombre de filles dans l'échantillon.

Solution

- d) On pige avec remise 4 logements dans un immeuble de 25 logements. On s'intéresse au nombre de personnes qui habitent le logement pigé.

Solution

1. $n = 4$ tirages indépendants (tirage avec remise).
2. À chaque tirage :
 - succès : impossible à définir.

La variable X : «nombre de personnes habitant le logement pigé» ne suit pas une binomiale, car le deuxième critère n'est pas respecté.

Tirage sans remise dans une grande population

En pratique, un tirage sans remise d'un petit nombre d'unités statistiques dans une grande population est considéré comme un tirage avec remise. Statistiquement, si la taille de la population est plus grande ou égale à 20 fois la taille de l'échantillon ($N \geq 20n$), la probabilité d'un événement est presque la même que l'on tienne compte ou non des unités déjà prélevées.

Par exemple, si l'on prélève un échantillon de 50 personnes dans une population de 2 000 personnes comprenant 800 femmes, la probabilité que la 21^e personne pigée soit une femme sachant que l'on a déjà pigé 10 femmes est presque la même, que le tirage soit fait avec ou sans remise :

$$\text{Tirage sans remise : } P(F_{21e} \mid \text{on a déjà 10 femmes}) = \frac{790}{1980} = 0,399$$

$$\text{Tirage avec remise : } P(F_{21e} \mid \text{on a déjà 10 femmes}) = \frac{800}{2000} = 0,400$$

On observe une différence négligeable de 0,001.

3.2.2 La fonction de probabilité d'une loi binomiale

La mise en situation suivante sert à illustrer une méthode de calcul rapide pour déterminer la probabilité associée à chaque valeur d'une variable aléatoire binomiale.

MISE EN SITUATION

Les statistiques d'un magasin indiquent que 25 % des clients paient leurs achats avec une carte de débit. Quatre clients font la file à la caisse. Déterminer l'événement qui a le plus de chances de se produire : que 0, 1, 2, 3 ou 4 clients paient avec une carte de débit.

Pour répondre à cette question, il faut construire la distribution de probabilité de la variable aléatoire X : «nombre de clients qui paient avec une carte de débit parmi 4 clients».

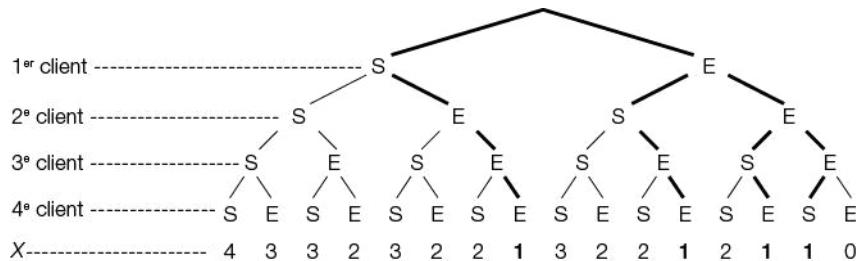
- Vérifions que cette expérience aléatoire est de type binomial.
1. L'expérience est constituée de 4 épreuves indépendantes (le mode de paiement d'un client n'est pas influencé par le mode de paiement du client précédent) : $n = 4$.
 2. Chaque épreuve peut avoir 2 résultats possibles :
 - succès : «le client paie avec une carte de débit» avec $p = 0,25$;
 - échec : «le client ne paie pas avec une carte de débit» avec $q = 0,75$.
 3. La variable aléatoire X correspond au nombre de succès en 4 épreuves.

Donc, X suit une $B(4; 0,25)$ où $X = 0, 1, 2, 3, 4$.

- Construisons la distribution de probabilité de la binomiale $B(4; 0,25)$.

x	0	1	2	3	4	Total
$f(x)$?	?	?	?	?	100 %

Le diagramme en arbre suivant donne tous les résultats possibles de cette expérience aléatoire.



- Déterminons $f(1)$.

Quatre chemins du diagramme mènent à la valeur 1 : le succès peut être obtenu avec le 1^{er}, le 2^e, le 3^e ou le 4^e client. Il y a donc 4 combinaisons possibles pour obtenir 1 succès en 4 épreuves. La formule suivante permet d'obtenir ce nombre :

$$\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4!}{1!3!} = 4 \text{ façons}$$

L'événement antécédent de 1 est $A = \{\text{SEEE}, \text{ESEE}, \text{EESE}, \text{EEES}\}$.

Pour chacun des chemins menant à la valeur 1, la probabilité est la suivante :

$$P(\text{SEEE}) = P(S \cap E \cap E \cap E) = (0,25)(0,75)(0,75)(0,75) = (0,25)(0,75)^3$$

$$P(\text{ESEE}) = P(E \cap S \cap E \cap E) = (0,75)(0,25)(0,75)(0,75) = (0,25)(0,75)^3$$

$$P(\text{EESE}) = P(E \cap E \cap S \cap E) = (0,75)(0,75)(0,25)(0,75) = (0,25)(0,75)^3$$

$$P(\text{EEES}) = P(E \cap E \cap E \cap S) = (0,75)(0,75)(0,75)(0,25) = (0,25)(0,75)^3$$

Ainsi, les éléments de l'événement antécédent de 1 sont tous équiprobables. Donc,

$$\begin{aligned}
 f(1) &= P(X=1) = P(A) \\
 &= P(\text{SEEE}) + P(\text{ESEE}) + P(\text{EESE}) + P(\text{EEES}) \\
 &= 4(0,25)(0,75)^3 \\
 &= \binom{4}{1}(0,25)(0,75)^3 \\
 &= 42,2 \%
 \end{aligned}$$

- Déterminons $f(2)$.

Combien y a-t-il de façons d'obtenir 2 succès en 4 épreuves ?

?

L'événement antécédent de 2 est $B = \{\text{SSEE, ESSE, EESS, SEES, SESE, ESES}\}$. Calculer la probabilité d'un élément quelconque de l'événement B.

?

Calculer $f(2)$.

- Déterminons $f(0), f(3)$ et $f(4)$.

?

Compléter la distribution de probabilité de la variable aléatoire X .

Distribution de probabilité de la binomiale $B(4; 0,25)$

Événement antécédent de x	x	$f(x) = P(X = x)$
{EEEE}	0	$f(0) =$
{SSEE, ESEE, EESE, EES}	1	$f(1) = \binom{4}{1} (0,25)(0,75)^3 = 4(0,25)(0,75)^3 = 42,2\%$
{SSEE, SESE, SEES, ESSE, ESES, EESS}	2	$f(2) = \binom{4}{2} (0,25)^2 (0,75)^2 = 6(0,25)^2 (0,75)^2 = 21,1\%$
{SSSE, SSES, SESS, ESSS}	3	$f(3) =$
{SSSS}	4	$f(4) =$

?

Déduire du tableau la formule permettant d'évaluer $f(x)$ pour une binomiale $B(4 ; 0,25)$.

$$f(x) = P(X = x) =$$

En généralisant, on obtient la formule de la fonction de probabilité $f(x)$ pour une binomiale $B(n ; p)$:

Fonction de probabilité d'une $B(n ; p)$

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{(n-x)} \quad \text{où} \quad x = 0, 1, 2, \dots, n$$

EXEMPLE 1

Calculer la probabilité d'obtenir 60 % à un test objectif de 10 questions, comportant chacune 4 choix de réponse, si l'on coche les réponses au hasard. La variable aléatoire X est le nombre de bonnes réponses données. Vérifier d'abord qu'il s'agit d'un contexte binomial.

Solution

EXEMPLE 2

On sait par expérience que 20 % des personnes sollicitées par un jeune vendeur de tablettes de chocolat en achètent. Calculer la probabilité que, parmi les six prochaines personnes sollicitées par cet enfant, au moins une achète une tablette. Vérifier d'abord qu'il s'agit d'un contexte binomial.

Solution

EXERCICES 3.2

1. Pour les expériences aléatoires suivantes, définir la variable aléatoire X et vérifier si elle suit une loi binomiale. Si oui, donner les paramètres n et p ; sinon, dire pourquoi X ne suit pas une loi binomiale.
 - a) On s'intéresse au nombre d'enfants de rang 2 (2^e enfant de la mère) parmi 10 nouveau-nés choisis au hasard. Selon l'Institut de la statistique du Québec, 36 % des nouveau-nés sont des enfants de rang 2.
 - b) On note le poids à la naissance de 10 nouveau-nés.
 - c) On s'intéresse au nombre de fumeurs dans un échantillon de 8 personnes prélevées sans remise dans un groupe de 30 personnes comprenant 12 fumeurs.
 - d) On s'intéresse au nombre de fumeurs dans un échantillon de 8 personnes prélevées avec remise dans un groupe de 30 personnes comprenant 12 fumeurs.
 - e) On s'intéresse au nombre de cégepiens qui ont échoué 2 cours ou plus dans un échantillon aléatoire de 30 cégepiens prélevés parmi les cégepiens du Québec. On sait que le tiers des cégepiens échouent 2 cours ou plus par session.
 - f) On pige 5 familles et on s'intéresse au nombre d'enfants dans la famille.
 - g) On pige, avec remise, une carte d'un jeu jusqu'à ce que l'on tire un cœur. On s'intéresse au nombre de tirages nécessaires pour obtenir le premier cœur.
2. Une expérience comporte 5 épreuves indépendantes, et la variable aléatoire X suit une $B(5; 0,75)$.
 - a) Énumérer les valeurs de la variable aléatoire X .
 - b) Combien y a-t-il de façons d'obtenir 4 succès en 5 épreuves?
 - c) Énumérer les éléments de l'événement antécédent de 4.
 - d) Calculer la probabilité d'un élément quelconque de l'événement antécédent de 4.
3. Le dernier recensement révèle que 21 % des enfants québécois de 17 ans ou moins vivent dans une famille monoparentale. On prélève sans remise un échantillon de 4 enfants dans ce groupe d'âge. Soit X : «le nombre d'enfants de l'échantillon qui vivent dans une famille monoparentale».

Source: Statistique Canada. *Tableau 111-0022, CANSIM*, octobre 2013.

 - a) Construire la distribution de probabilité de X .
 - b) Calculer la probabilité:
 - i) que 2 des 4 enfants vivent dans une famille monoparentale.
 - ii) que moins de 3 enfants vivent dans une famille monoparentale.
 - iii) qu'au moins 2 enfants vivent dans une famille monoparentale.
 - iv) qu'au plus 3 enfants vivent dans une famille monoparentale.
 4. Lors d'un contrôle de qualité, on pige avec remise 10 pièces dans un lot de 100 pièces. Si le lot contient 12 pièces défectueuses, quelle est la probabilité:
 - a) qu'aucune des 10 pièces tirées ne soit défectueuse?
 - b) qu'au moins une des 10 pièces tirées soit défectueuse?
 5. On lance un dé 5 fois. Quelle est la probabilité d'obtenir 4 fois un nombre supérieur à 4?
 6. En 2009, 24,5 % des titulaires d'un permis de conduire ont accumulé au moins 2 points d'inaptitude sur une période de 2 ans. On prélève un échantillon de 50 titulaires d'un permis de conduire. Quelle est la probabilité que huit d'entre eux aient accumulé au moins deux points d'inaptitude?

Source: Société de l'assurance automobile du Québec. *Les infractions et les sanctions reliées à la conduite d'un véhicule routier*, décembre 2011.

3.2.3 L'espérance et l'écart type d'une loi binomiale

Dans la section 3.1, nous avons appris à calculer l'espérance et l'écart type d'une distribution de probabilité. La mise en situation suivante présente une méthode plus rapide pour déterminer ces mesures dans le cas d'une distribution binomiale.

En considérant un magasin où 25 % des clients paient leurs achats avec une carte de débit, nous avons établi que la probabilité qu'un échantillon de 4 clients compte 0, 1, 2, 3 ou 4 clients qui paient de cette façon est donnée par la distribution de probabilité suivante :

Distribution de probabilité de la binomiale $B(4; 0,25)$

x	$f(x) = P(X = x)$
0	$f(0) = \binom{4}{0} (0,25)^0 (0,75)^4 = 31,6\%$
1	$f(1) = \binom{4}{1} (0,25)(0,75)^3 = 42,2\%$
2	$f(2) = \binom{4}{2} (0,25)^2 (0,75)^2 = 21,1\%$
3	$f(3) = \binom{4}{3} (0,25)^3 (0,75) = 4,7\%$
4	$f(4) = \binom{4}{4} (0,25)^4 (0,75)^0 = 0,4\%$

Analyse de la distribution de probabilité

Dans un magasin où 25 % des clients paient leurs achats avec une carte de débit, sur 4 clients, le résultat le plus probable (42 %) est qu'un seul d'entre eux utilise sa carte de débit. Le résultat le moins probable (0,4 %) est que les 4 clients paient de cette façon.

Complétons cette analyse en calculant l'espérance et l'écart type de la distribution de probabilité.



- Si l'on réalise souvent cette expérience aléatoire, sachant que 25 % des clients paient avec une carte de débit, combien de clients en moyenne vont utiliser ce mode de paiement parmi les 4 clients de l'échantillon ? Exprimer le raisonnement en langage mathématique.

$$E(X) = \underline{\hspace{2cm}} \text{ client(s)}$$

- Utiliser le mode statistique de la calculatrice pour vérifier que l'espérance de la distribution de probabilité de la binomiale $B(4 ; 0,25)$ est bien égale au résultat précédent, puis déterminer l'écart type de la distribution.

$$E(X) = \underline{\hspace{2cm}} \quad \sigma = \underline{\hspace{2cm}}$$

- Vérifier que $\sigma = \sqrt{npq} = \underline{\hspace{2cm}}$

Donc, si X suit une binomiale $B(4 ; 0,25)$, alors $E(X) = 1$ et $\sigma = 0,9$.

Interprétation

Sur un grand nombre d'échantillons de 4 clients, on peut espérer trouver en moyenne 1 client par échantillon qui paie avec une carte de débit, l'écart type étant de 0,9 client. Donc, dans la plupart des échantillons, au plus 2 clients utiliseront ce mode de paiement.

En généralisant les résultats de la mise en situation à une $B(n ; p)$, on obtient les formules suivantes :

Espérance et écart type d'une binomiale $B(n; p)$

Espérance : $E(X) = np$

Écart type : $\sigma = \sqrt{npq}$

Utilisation d'une table binomiale

Vous trouverez en annexe (voir la page 337) une table donnant la distribution de probabilité d'une variable aléatoire binomiale $B(n; p)$ pour $n \leq 20$. Par exemple, on obtient les valeurs de probabilités $f(x)$ de la binomiale $B(4; 0,25)$ par l'une ou l'autre des méthodes suivantes :

Avec la fonction de probabilité		Avec la table binomiale	
x	$f(x) = P(X = x)$	x	$f(x) = P(X = x)$
0	$1 (0,25)^0 (0,75)^4 = 0,3164$	0	0,3164
1	$4 (0,25)^1 (0,75)^3 = 0,4219$	1	0,4219
2	$6 (0,25)^2 (0,75)^2 = 0,2109$	2	0,2109
3	$4 (0,25)^3 (0,75)^1 = 0,0469$	3	0,0469
4	$1 (0,25)^4 (0,75)^0 = 0,0039$	4	0,0039

Attention !

Quand la valeur de n est dans la table, mais pas celle de p , on utilise la fonction de probabilité pour calculer $f(x)$. À titre d'exemple, c'est ce qu'il faut faire avec la $B(12; 0,23)$.

EXEMPLE 1

Dans une entreprise, une machine appose 4 000 étiquettes par heure sur des contenants. Habituellement, la production de cette machine comporte au plus 5 % d'étiquettes défectueuses (déchirées, mal collées, etc.), ce qui est acceptable. Afin de s'assurer de ne pas dépasser ce pourcentage, on prélève toutes les heures, sans remise, un échantillon aléatoire de 20 contenants dont on inspecte les étiquettes.

- Si la machine fonctionne bien, combien d'étiquettes défectueuses en moyenne peut-on espérer trouver dans un échantillon de 20 contenants ? Quel est l'écart type ? Interpréter ces mesures.

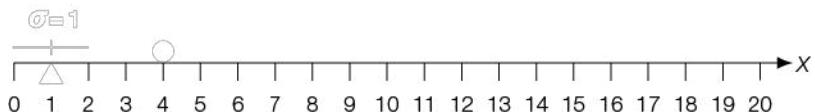
Solution

Interprétation

Si l'on prélève un grand nombre d'échantillons de 20 contenants, on s'attend à trouver en moyenne _____ étiquette défectueuse par échantillon avec un écart type de _____ étiquette. La plupart des échantillons contiendront soit 0, 1 ou 2 contenants dont l'étiquette est défectueuse.

- Si la personne chargée du contrôle trouve 4 étiquettes défectueuses dans un même échantillon, devrait-elle faire vérifier le réglage de la machine ? Représenter graphiquement la situation, puis justifier la réponse à l'aide de la valeur de la cote z .

Solution



Conclusion

Dans l'hypothèse où la production contient 5 % d'étiquettes défectueuses, une cote z de 3 est très rare. On peut en tirer l'une ou l'autre des conclusions suivantes :

- ou bien on a prélevé par hasard un échantillon qui a très peu de chances d'être obtenu ;
- ou bien le pourcentage d'étiquettes défectueuses est en fait supérieur à 5 %.

La seconde conclusion semble la plus vraisemblable : il faut vérifier la machine.

- c) Selon la table binomiale, quelles sont les chances de prélever un échantillon qui compte 4 étiquettes défectueuses, selon l'hypothèse que la machine fonctionne bien, c'est-à-dire qu'elle ne produit pas plus de 5 % d'étiquettes défectueuses ?

Solution

Bien sûr, on peut se tromper en décidant de faire vérifier la machine, mais le risque n'est que de 1,3 % : il y a seulement 1,3 % de chances d'obtenir 4 étiquettes défectueuses dans un échantillon quand la machine fonctionne bien. Cet exemple illustre un des buts du cours, soit d'apprendre à calculer les risques lors d'une prise de décision portant sur une situation où le hasard intervient.

- d) Suggérer un plan de contrôle qui permettrait de s'assurer du bon fonctionnement de la machine. Pour qu'il soit facile à appliquer, ce plan doit avoir la forme suivante.

Plan de contrôle

Vérifier l'ajustement de la machine si le nombre d'étiquettes défectueuses dans un échantillon de 20 étiquettes est _____.

Selon votre plan de contrôle, quels sont les risques de faire vérifier inutilement la machine ?

Solution

EXEMPLE 2

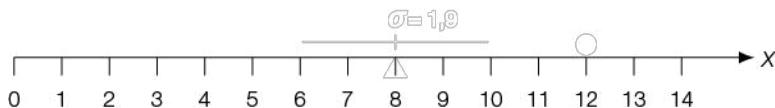
- a) En 2011, 57 % des détenteurs d'un baccalauréat en psychologie ont poursuivi leurs études après l'obtention de leur diplôme¹. En 2015, un échantillon de 14 détenteurs d'un baccalauréat en psychologie est prélevé parmi les diplômés de la promotion 2014. Dans l'hypothèse où les statistiques de 2011 sont encore valables en 2015, en moyenne, combien de diplômés encore aux études peut-on espérer trouver dans l'échantillon ?

Source: Ministère de l'Enseignement supérieur. *La relance à l'université – 2013. La situation d'emploi des personnes diplômées.*

Solution

- b) En se basant sur la valeur de la cote z , doit-on conclure que les statistiques de 2011 ne sont plus valables en 2015 si 12 des 14 diplômés de l'échantillon ont poursuivi leurs études après l'obtention de leur baccalauréat en psychologie ? Quelles sont les chances de piger un tel échantillon dans l'hypothèse où les statistiques de 2011 sont encore valables en 2015 ? Représenter la situation sur l'axe ci-dessous.

Solution



- c) Le pourcentage de titulaires d'un baccalauréat en psychologie qui poursuivent leurs études a-t-il augmenté ou diminué depuis 2011 ? _____

EXERCICES 3.3

1. Sur chaque table d'un restaurant, un présentoir contient un questionnaire destiné à sonder le niveau de satisfaction des clients. Les statistiques indiquent que seulement 10 % des clients remplissent le questionnaire. On prélève un échantillon aléatoire de 10 clients et l'on s'intéresse à la variable aléatoire X : «nombre de clients qui ont rempli le questionnaire».
 - a) Calculer et interpréter l'espérance et l'écart type de X .
 - b) Quelle est la probabilité que plus de deux clients aient rempli le questionnaire?
2. Selon des écologistes, 50 % des Québécois sont contre l'exploitation du gaz de schiste, 30 % sont pour et 20 % n'ont pas d'opinion.
 - a) On effectue un sondage auprès d'un échantillon aléatoire de 10 Québécois. Calculer la probabilité que l'échantillon comprenne au moins 5 répondants favorables à l'exploitation du gaz de schiste.
 - b) On effectue un sondage auprès de 1 000 Québécois. Calculer et interpréter l'espérance et l'écart type du nombre de répondants favorables à l'exploitation du gaz de schiste.

1. Pour exercer la profession de psychologue, il faut détenir un doctorat en psychologie.

- c) Devriez-vous douter de l'affirmation des écolos selon laquelle 30 % des Québécois sont pour l'exploitation du gaz de schiste si, sur les 1 000 Québécois de l'échantillon :
- on en compte 250 qui sont favorables à l'exploitation du gaz de schiste ?
 - on en compte 315 qui sont favorables à l'exploitation du gaz de schiste ?
- d) Devriez-vous douter de l'affirmation voulant que 30 % des Québécois sont pour l'exploitation des gaz de schiste :
- si, dans un échantillon de 10 Québécois, 2 sont pour l'exploitation du gaz de schiste ?
 - si, dans un échantillon de 1 000 Québécois, 200 sont pour l'exploitation du gaz de schiste ?
3. Une étude révèle que 20 % des PME cessent leurs activités moins d'un an après leur création.
- Source:** Industrie Canada. *Principales statistiques relatives aux petites entreprises – Août 2013*.
- Pour un échantillon aléatoire de 6 nouvelles PME :
 - quelle est la probabilité qu'exactement 2 PME cessent leurs activités la 1^{re} année ?
 - quelle est la probabilité que plus de 2 PME cessent leurs activités la 1^{re} année ?
 - quelle est la probabilité que toutes survivent à leur 1^{re} année ?
 - Pour 100 nouvelles PME, en moyenne, combien cesseront leurs activités la 1^{re} année ?
4. Une étude révèle que 35 % des jeunes de 18 à 24 ans ont un baladeur numérique.
- Source:** CEFARIO. *NETendances 2013: La mobilité au Québec: une montée en flèche*, vol. 4, n° 7.
- Dans un échantillon de 12 jeunes de ce groupe d'âge :
 - quelles sont les chances qu'un seul jeune n'ait pas de baladeur numérique ?
 - quelles sont les chances que 8 des 12 jeunes aient un baladeur numérique ?
 - Calculer et interpréter l'espérance et l'écart type de cette loi binomiale.
5. Un test à choix multiples comprend 10 questions, et chaque question comporte 5 choix. Si l'on répond au hasard à chaque question :
- quelle est la probabilité de n'avoir aucune bonne réponse sur 10 ?
 - quelle est la probabilité d'obtenir 100 % à ce test ?
 - quelle est la probabilité d'obtenir la note de passage, qui est de 60 % ?
 - quelle est la probabilité d'obtenir au moins la note de passage ?
 - quelle est la probabilité d'obtenir moins de 60 % ?
6. Les participants à un congrès peuvent payer les frais d'inscription à l'avance ou sur place à l'ouverture du congrès. On estime que 20 % des congressistes paieront à l'ouverture. Une personne est chargée d'accueillir les participants et de diriger ceux qui n'ont pas payé vers trois guichets réservés à cette fin. Quelle est la probabilité que le nombre de guichets ne soit pas suffisant si 10 congressistes se présentent en même temps sur les lieux du congrès ?
7. Léo doit répondre par vrai ou faux à un test éclair comportant 15 questions sur les lectures obligatoires d'un cours. N'ayant pas lu les documents, il décide de répondre au hasard en lançant une pièce de monnaie. Il répond vrai s'il obtient face, et faux s'il obtient pile. La note de passage est de 60 %. Quelle est la probabilité que Léo obtienne au moins la note de passage ?
8. Une variable aléatoire X suit une loi binomiale $B(8; 0,75)$.
- Utiliser la table binomiale pour déterminer la probabilité d'obtenir 6 succès.

Conseil

Comme la table ne donne pas la distribution de probabilité pour une binomiale où $p = 0,75$, il faut aborder le problème sous un autre angle. Avoir 6 succès en 8 épreuves avec $p = 0,75$, cela équivaut à avoir 2 échecs en 8 épreuves avec $q = 0,25$. En définissant la variable aléatoire Y : «nombre d'échecs en 8 épreuves», on a $P(X = 6)$ pour une $B(8; 0,75)$ est égal à $P(Y = 2)$ pour une $B(8; 0,25)$.
 - Calculer la probabilité d'obtenir au moins 7 succès.
9. En 2012, le pourcentage d'internautes qui utilisent des sites de réseautage social, tels que Facebook et Twitter, est de 70 % chez les femmes et de 64 % chez les hommes.
- Source:** Statistique Canada. «Utilisation d'Internet et du commerce électronique par les particuliers, 2012», *Le Quotidien*, octobre 2013.
- On préleve un échantillon de 10 utilisatrices d'Internet. Utiliser la table binomiale pour trouver les probabilités demandées.
- Quelle est la probabilité que 8 des 10 femmes utilisent des sites de réseautage social ?
 - Quelle est la probabilité que moins de 8 femmes utilisent des sites de réseautage social ?

3.3 La loi de Poisson

La loi de Poisson, ou loi des événements rares, est un modèle de distribution de probabilité qui s'applique à une variable aléatoire discrète. Cette loi est utilisée, entre autres, dans les domaines suivants : contrôle de la qualité, vérification comptable, gestion de files d'attente, assurance, démographie et télécommunication. On s'en sert aussi comme approximation d'une distribution binomiale dans le cas de petites probabilités.



Siméon Denis Poisson (1781-1840)

À 17 ans, Siméon Denis Poisson termine ses études à l'École Polytechnique de Paris, se classant au premier rang. À 18 ans, il voit publier son mémoire sur les différences finies grâce à l'intervention du mathématicien Legendre, qui remarque son talent. Il enseigne ensuite les mathématiques pendant quelques années. Un autre de ses articles est publié en 1837, dans lequel il analyse ce que l'on connaît aujourd'hui comme la distribution de Poisson. Ce géomètre, mathématicien et physicien français a grandement contribué à l'avancement des mathématiques et de la physique.

3.3.1 Le contexte d'une expérience aléatoire de Poisson

Une variable aléatoire discrète suit un modèle de Poisson si elle correspond à un nombre de succès par région. Le terme **région** désigne un intervalle de temps, une longueur, une surface, un volume, etc. Pour définir la fonction de probabilité d'une expérience aléatoire de Poisson, il faut connaître la moyenne de succès pour la région considérée ; on note cette moyenne par λ (lire lambda). La fonction de probabilité d'une variable aléatoire X qui suit une loi de Poisson de moyenne λ , notée $\text{Po}(\lambda)$, est définie ainsi :

Fonction de probabilité de la loi de Poisson $\text{Po}(\lambda)$

$$f(x) = \text{P}(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{où } x = 0, 1, 2, 3, \dots$$

$$\text{Espérance : } E(X) = \lambda$$

$$\text{Écart type : } \sigma = \sqrt{\lambda}$$

La lettre e de l'expression est une constante qui vaut approximativement 2,7183. Pour faciliter le calcul des probabilités, on fait appel à la table de Poisson (*voir la page 342*).

Par exemple, on détermine la valeur $f(0)$ pour une variable aléatoire X qui suit une $\text{Po}(0,1)$ de la façon suivante :

Dans la table : $f(0) = \text{P}(X = 0) = 0,9048$ (nombre à l'intersection de la colonne $\lambda = 0,1$ et de la ligne 0).

$$\text{Avec la formule, on a : } f(0) = \text{P}(X = 0) = \frac{e^{-0,1}(0,1)^0}{0!} = \frac{0,9048 \times 1}{1} = 0,9048.$$

EXEMPLE 1

Un fabricant de tissu doit s'assurer régulièrement que le nombre de défauts (éraflures, fils tirés, taches, etc.) n'est pas trop élevé. La technologie utilisée pour la fabrication permet de fixer le standard de qualité à une moyenne de 2,3 défauts par 50 m^2 de tissu. Au dernier contrôle de qualité, on a trouvé 6 défauts dans un échantillon de 50 m^2 de tissu.

- a) Devrait-on en conclure que le standard de qualité n'est plus respecté? Justifier la réponse à l'aide de la cote z et représenter la situation sur l'axe gradué ci-dessous.

Solution

- Vérifions d'abord qu'il s'agit d'une expérience aléatoire de Poisson.

X : «nombre de défauts sur une surface de 50 m^2 ».

Succès : _____

Région : _____

Moyenne λ pour la région considérée : _____

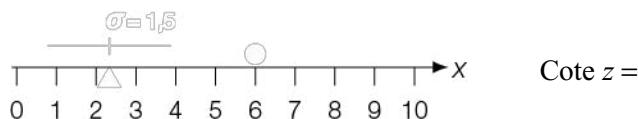
Donc, X suit une _____ où $X =$ _____

- Calculons la cote z de $X = 6$.

– Déterminons l'espérance et l'écart type de X .

Pour 50 m^2 de tissu, on a: $E(X) = \lambda =$ _____ et $\sigma = \sqrt{\lambda} =$ _____.

– Représentons ces mesures sur l'axe, situons $X = 6$ défauts, puis calculons la cote z de 6.



Cote $z =$ _____

Conclusion

Selon l'hypothèse qu'il y a 2,3 défauts en moyenne par 50 m^2 de tissu, il est assez rare d'obtenir 6 défauts, donc une cote z de 2,5. On peut tirer l'une ou l'autre des conclusions suivantes :

- ou bien on a prélevé un échantillon qu'on a peu de chances d'obtenir;
- ou bien le standard de qualité n'est pas respecté.

La seconde conclusion semble plus vraisemblable : le standard de qualité n'est pas respecté.

- b) Selon l'hypothèse que le standard de qualité est respecté, quelles sont les chances de prélever un échantillon de 50 m^2 de tissu qui contient 6 défauts?

Solution

Avec une cote z de 2,5, on s'attend à ce que la probabilité d'obtenir $X = 6$ pour une $\text{Po}(2,3)$ soit petite. En effet, la table de la loi de Poisson donne :

$$f(6) = P(X = 6) = \text{_____}.$$

On risque peu de se tromper en pariant sur le non-respect du standard de qualité : il y a seulement 2,1 % de chances d'obtenir un échantillon avec 6 défauts quand le standard de qualité est respecté.

EXEMPLE 2

On a observé qu'en moyenne 36 personnes se présentent à un guichet automatique situé près d'une station de métro entre 17 h et 18 h. Estimer la probabilité que moins de 2 personnes se présentent au guichet sur une période quelconque de 2 minutes entre 17 h et 18 h.

Solution

X : «nombre de personnes qui se présentent au guichet sur une période de 2 minutes».

Succès : une personne se présente au guichet.

Région : période de 2 minutes.

Moyenne pour 2 minutes : $\lambda = ?$

On sait qu'il y a en moyenne 36 personnes par 60 minutes. En supposant une distribution uniforme dans le temps, on obtient le nombre moyen λ de personnes par 2 minutes ainsi :

$$\frac{36 \text{ personnes}}{60 \text{ min}} = \frac{\lambda}{2 \text{ min}} \Rightarrow \lambda = \frac{2 \times 36}{60} = 1,2 \text{ personne}$$

X suit une $\text{Po}(1,2)$ où $X = 0, 1, 2, \dots$

$$\begin{aligned} P(X < 2) &= f(0) + f(1) \\ &= 0,3012 + 0,3614 = 0,6626 \\ &= 66,3 \% \end{aligned}$$

On peut estimer qu'il y a 66 % de chances que moins de 2 personnes se présentent au guichet automatique sur une période de 2 minutes entre 17 h et 18 h.

EXEMPLE 3

Vous faites du bénévolat auprès d'un organisme qui offre des vêtements à prix réduits aux personnes défavorisées. En moyenne, 4 personnes se présentent toutes les 20 minutes. Vous désirez vous absenter cinq minutes pour prendre une collation et personne ne peut vous remplacer. Quels sont les risques qu'au moins une personne se présente pendant votre absence ?

Solution

3.3.2 L'approximation de la loi binomiale par la loi de Poisson

La loi de Poisson permet de faire une bonne approximation d'une loi binomiale lorsque n est grand et que p est petit, c'est-à-dire en pratique si $n > 20$, $p \leq 0,10$ et $np \leq 5$.

Approximation de la loi binomiale par la loi de Poisson

Si $n > 20$, $p \leq 0,10$ et $np \leq 5$, alors

$$B(n; p) \approx Po(\lambda) \quad \text{où} \quad \lambda = np$$

NOTE

L'approximation est d'autant meilleure que n est grand et que p est petit. Il est à souligner que les conditions sur n et np varient d'un auteur à l'autre, ce qui porte parfois à confusion. Certains posent $n \geq 30$ et $np \leq 5$, d'autres $n \geq 50$ et $np \leq 10$, ou bien $n > 20$, $p \leq 0,10$ et $np \leq 5$. Comme la plupart des tables binomiales limitent la valeur de n à 20, nous avons retenu ce dernier critère.

EXEMPLE 1

On a constaté que 3 % des appareils fabriqués par une entreprise sont défectueux. Calculer la probabilité qu'un lot de 60 appareils contienne au plus 5 appareils défectueux.

Solution

X : « le nombre d'appareils défectueux » suit une $B(60; 0,03)$.

On cherche $P(X \leq 5)$. Pour calculer cette probabilité :

- on ne peut pas utiliser la table binomiale, car la valeur $n = 60$ n'est pas dans la table.
- on peut toujours utiliser la formule binomiale, mais les calculs sont longs :

$$P(X \leq 5) = \binom{60}{5}(0,03)^5(0,97)^{55} + \binom{60}{4}(0,03)^4(0,97)^{56} + \cdots + \binom{60}{0}(0,03)^0(0,97)^{60}$$

- Vérifions si l'on peut utiliser la loi de Poisson pour faire une approximation de la $B(60; 0,03)$.

$$\text{On a : } n = 60 > 20 \quad p = 0,03 \leq 0,10 \quad np = (60)(0,03) = 1,8 \leq 5.$$

$$\text{Donc, } B(60; 0,03) \approx Po(\lambda) \quad \text{où} \quad \lambda = np = 1,8.$$

En examinant le nombre de valeurs possibles pour une $Po(1,8)$, on constate qu'il y a moins de probabilités à additionner si l'on calcule la probabilité cherchée à l'aide de la probabilité de l'événement contraire :

$$\begin{aligned} P(X \leq 5) &= 1 - P(X > 5) \\ &= 1 - [f(6) + f(7) + f(8) + f(9)] \\ &= 1 - (0,0078 + 0,0020 + 0,0005 + 0,0001) \\ &= 0,9896 = 99,0 \% \end{aligned}$$

Pour l'hypothèse selon laquelle 3 % des appareils fabriqués sont défectueux, il y a 99 % de chances qu'un lot de 60 appareils en contienne au plus 5 défectueux.

EXEMPLE 2

La garantie accompagnant un produit stipule qu'un client insatisfait dispose de quinze jours pour obtenir un remboursement. L'expérience montre que seulement 2,5 % des clients se prévalent de ce droit. Sur un échantillon de 80 clients qui ont acheté le produit :

- Combien de clients en moyenne réclameront un remboursement dans un délai de quinze jours suivant l'achat ? Quelle devrait être la valeur de l'écart type ?
- Quelle est la probabilité que moins de 4 clients demandent un remboursement ?

Solution

X : «nombre de clients sur 80 qui réclameront un remboursement dans un délai de 15 jours».

EXERCICES 3.4

- Les statistiques révèlent qu'une entreprise subit en moyenne 9 pannes informatiques par année. En supposant que le modèle de Poisson s'applique à la distribution du nombre de pannes.
 - Quelle est la probabilité qu'il n'y ait aucune panne sur une période de 6 mois ?
 - Quelle est la probabilité qu'il se produise au moins 1 panne sur une période de 2 mois ?
 - Quel est le nombre de pannes le plus probable sur une période de 2 mois et quelle est sa probabilité ?
- Une étude sur l'achalandage du site Web d'un cégep dans la semaine précédant le début des cours révèle qu'en moyenne 2,4 personnes par minute visitent le site entre 12 h et 21 h.
 - Déterminer l'écart type du nombre de visiteurs par minute entre 12 h et 21 h et interpréter cette mesure.
 - Quelle est la probabilité que le réseau subisse un ralentissement entre 12 h et 21 h si ce phénomène se produit lorsque plus de 9 personnes visitent le site dans la même minute ?

- c) Calculer la probabilité que plus de 3 personnes par minute visitent le site Web entre 12 h et 21 h.
- d) Calculer la probabilité que personne ne visite le site Web durant une période de 30 secondes entre 12 h et 21 h.
3. La directrice d'une usine vient de signer un contrat avec un sous-traitant qui doit fournir une pièce utilisée dans le montage de moteurs. Le contrat stipule que le pourcentage de pièces défectueuses ne doit jamais être supérieur à 1 %. L'usine vient de recevoir un lot de 1 000 pièces, et on décide d'en vérifier la qualité en prélevant un échantillon aléatoire de 40 pièces.
- En moyenne, combien de pièces défectueuses s'attend-on à trouver dans l'échantillon si le contrat est respecté? Quel devrait être l'écart type? Interpréter ces mesures.
 - i) Si l'on trouve plus de 2 pièces défectueuses dans l'échantillon, en se basant sur la cote z , devrait-on décider de retourner le lot au sous-traitant?
 - ii) Dans l'hypothèse où le sous-traitant respecte le contrat quant au pourcentage de pièces défectueuses, quelles sont les chances que l'échantillon contienne plus de 2 pièces défectueuses?
4. **TERRITOIRE EXCEPTIONNEL POUR LA CHASSE AU CERF**
- Le cerf de Virginie jouit d'un habitat idéal sur l'île d'Anticosti: la nourriture y est abondante et, exception faite de l'homme, il n'est la proie d'aucun prédateur. Afin de fixer le quota annuel de chasse, on recense régulièrement la population de cerfs. Selon l'étude, il y a en moyenne 20 cerfs par kilomètre carré sur l'île. En se basant sur cette donnée, calculer la probabilité qu'un chasseur trouve de 5 à 10 cerfs dans une zone d'un quart de kilomètre carré.
- Source:** Ministère des Ressources naturelles et de la Faune du Québec. Janvier 2014.
5. Dans une entreprise, il se produit en moyenne un accident de travail par période de quatre semaines.
- Quelle est la probabilité qu'il ne se produise aucun accident en quatre semaines?
 - Quelle est la probabilité qu'il se produise moins de 3 accidents en deux semaines?
6. Une entreprise estime que 1,5 % des jouets qu'elle fabrique présentent des défauts. Pour un échantillon de 100 jouets tirés au hasard, quelles sont les chances:
- qu'exactement 2 jouets soient défectueux?
 - que pas plus de 2 jouets soient défectueux?
 - qu'au moins 2 jouets soient défectueux?
7. **LES MARIAGES DE CONJOINTS DE MÊME SEXE SONT-ILS NOMBREUX?**
- Les mariages de conjoints de même sexe sont permis au Québec depuis 2004. Sur les 23 491 mariages célébrés en 2012, on compte 530 mariages entre conjoints de même sexe. On prélève un échantillon aléatoire de 150 mariages parmi les 23 491 mariages célébrés cette année-là.
- Source:** Institut de la statistique du Québec. *Mariages et unions civiles selon le sexe des conjoints*, juin 2013.
- Quelle est la probabilité qu'il y ait 3 mariages de conjoints de même sexe dans l'échantillon?
 - Quelle est la probabilité qu'il y ait moins de 5 mariages de conjoints de même sexe dans l'échantillon?
8. L'incendie représente 4 % des réclamations en assurance habitation au Québec. Dans un échantillon de 120 réclamations, quelle est la probabilité qu'il y ait au moins quatre réclamations pour incendie?
- Source:** Bureau d'assurance du Canada. 2012.

3.4 La loi normale

La loi normale est la loi la plus importante en statistique. Bien que cette loi soit définie pour une variable aléatoire continue, on l'utilise parfois pour une variable aléatoire discrète quand celle-ci compte un grand nombre de valeurs. La loi normale peut aussi servir à faire une approximation de la loi binomiale dans certaines conditions.



Carl Friedrich Gauss (1777-1855)

Enfant, Gauss possède très tôt une compréhension remarquable des mathématiques. On raconte qu'il apprend seul à lire et à compter à l'âge de trois ans. Au fil de sa carrière, ce mathématicien, astronome et physicien allemand fait des découvertes dans de nombreux domaines, dont l'algèbre, l'optique et l'électromagnétisme, mais on l'associe principalement au domaine des probabilités pour sa loi normale ou loi de Gauss, une loi statistique continue représentée par une courbe en forme de cloche. On le surnomme le « prince des mathématiciens ».

3.4.1 Un rappel

Avant d'entreprendre l'étude de la loi normale, il est bon de rappeler une notion qui nous sera très utile.

En statistique descriptive, on a appris à construire un histogramme pour représenter la distribution d'une variable quantitative continue. On sait que la construction d'un tel histogramme repose sur la proportionnalité entre la surface des rectangles et le pourcentage de données des classes, ce qui permet de déterminer le pourcentage de données dans une classe à l'aide d'un rapport d'aires :

$$\text{Pourcentage de données dans une classe} = \frac{\text{aire du rectangle de la classe}}{\text{aire totale de l'histogramme}} \times 100 \%$$

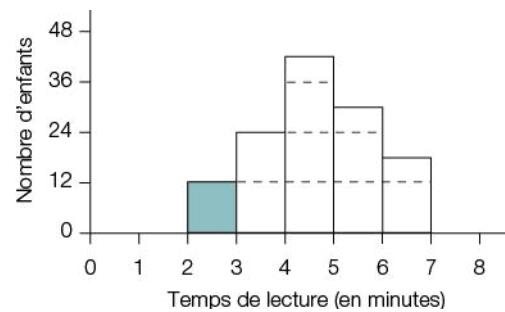
EXEMPLE

On a mesuré le temps que mettent 126 enfants de même âge pour lire un texte. Le tableau et le graphique suivants présentent les résultats.

Répartition des enfants selon le temps pris pour lire un texte

Temps (en min)	Nombre d'enfants	Pourcentage
$2 \leq X < 3$	12	9,5 %
$3 \leq X < 4$	24	19,0 %
$4 \leq X < 5$	42	33,3 %
$5 \leq X < 6$	30	23,8 %
$6 \leq X < 7$	18	14,3 %
Total	126	99,9 %

Répartition des enfants selon le temps pris pour lire un texte



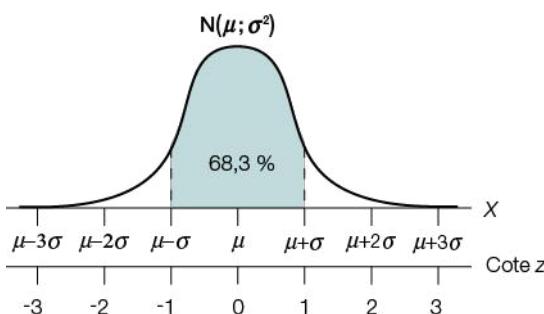
En prenant l'aire du rectangle bleu comme unité de mesure, on calcule le pourcentage d'enfants qui ont pris entre 4 et 5 minutes pour lire le texte ainsi :

$$\% \text{ d'enfants ayant pris entre 4 et 5 minutes} = \frac{\text{aire du rectangle de la 3^e classe}}{\text{aire totale de l'histogramme}} \times 100 \% = \frac{3,5}{10,5} \times 100 \% = 33,3 \%$$

3.4.2 La courbe normale ou courbe de Gauss

Chaque fois que l'on prend des mesures analogues sur des sujets semblables (par exemple la taille d'enfants de même âge, le volume de jus dans des contenants de même grandeur, etc.), on obtient une distribution ayant la forme d'une cloche. On donne le nom de **courbe normale**, ou **courbe de Gauss**, à ce type de graphique où le pourcentage de données est élevé autour de la moyenne et de plus en plus faible à mesure que l'on s'éloigne de celle-ci, dans un sens ou dans l'autre.

On désigne une courbe normale par la notation $N(\mu; \sigma^2)$ où μ représente la moyenne et σ , l'écart type de la distribution.



Équation de la courbe normale

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Bien qu'il existe un très grand nombre de courbes normales, elles présentent toutes les caractéristiques suivantes :

1. La courbe a la forme d'une cloche parfaitement symétrique par rapport à la moyenne : le mode, la médiane et la moyenne ont la même valeur. Théoriquement, la courbe s'étend indéfiniment de chaque côté de la moyenne.
2. L'aire totale comprise entre la courbe et l'axe des x est égale à 1.
3. La surface entre la courbe et l'axe des x se répartit comme suit :
 - 68,3 % de la surface totale est comprise entre la moyenne moins un écart type et la moyenne plus un écart type, soit entre les bornes de l'intervalle $[\mu - \sigma; \mu + \sigma]$. Les données appartenant à cet intervalle ont une cote z comprise entre -1 et 1.
 - 99,7 % de la surface totale est comprise entre la moyenne moins trois écarts types et la moyenne plus trois écarts types, soit entre les bornes de l'intervalle $[\mu - 3\sigma; \mu + 3\sigma]$. Dans une distribution normale, presque toutes les données ont une cote z comprise entre -3 et 3.

EXEMPLE

La courbe normale $N(50; 100)$ représentée possède les caractéristiques suivantes :

- Sa moyenne est $\mu = 50$ et son écart type $\sigma = 10$.
- 68,3 % de la surface sous la courbe est comprise entre 40 et 60 :

$$\mu - \sigma = 50 - 10 = 40$$

$$\mu + \sigma = 50 + 10 = 60$$

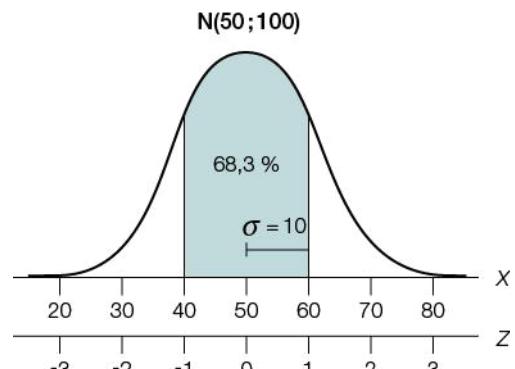
Dans cette zone, les cotes z sont comprises entre -1 et 1.

- 99,7 % de la surface sous la courbe est comprise entre 20 et 80 :

$$\mu - 3\sigma = 50 - 3 \times 10 = 20$$

$$\mu + 3\sigma = 50 + 3 \times 10 = 80$$

Dans cette zone, les cotes z sont comprises entre -3 et 3.



3.4.3 Le calcul d'une probabilité pour une loi normale

Dans les sections précédentes, nous avons vu comment calculer la probabilité pour une variable aléatoire discrète. On ne peut pas utiliser la même approche pour une variable aléatoire continue, comme l'illustre la mise en situation suivante.

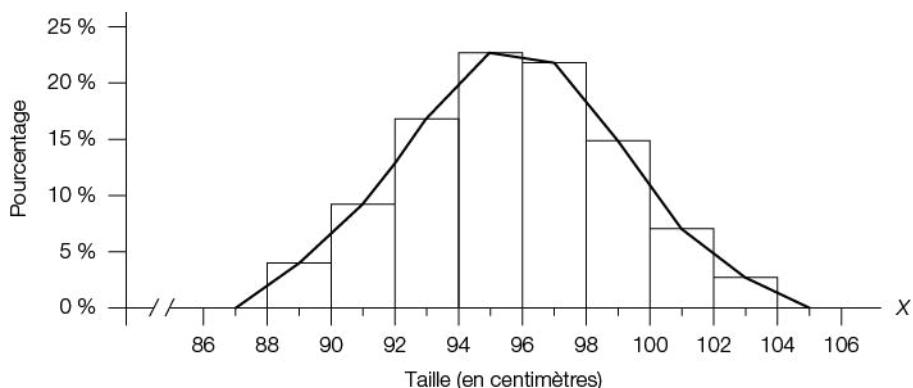
MISE EN SITUATION

Afin d'établir une courbe de croissance des enfants québécois, des chercheurs ont suivi le développement de 500 garçons et de 500 filles de leur naissance jusqu'à l'âge de 12 ans. Avec les données recueillies, on a pu construire la distribution de la taille et du poids des enfants à divers âges. Voici la distribution de la taille des 500 garçons à l'âge de 3 ans ainsi que la représentation graphique de cette distribution.

Répartition des garçons selon leur taille à l'âge de 3 ans

Taille à 3 ans	Nombre de garçons	Pourcentage
Moins de 90 cm	22	4,4 %
[90 cm; 92 cm[47	9,4 %
[92 cm; 94 cm[84	16,8 %
[94 cm; 96 cm[113	22,6 %
[96 cm; 98 cm[109	21,8 %
[98 cm; 100 cm[75	15,0 %
[100 cm; 102 cm[35	7,0 %
102 cm et plus	15	3,0 %
Total	500	100,0 %

Répartition des garçons selon leur taille à l'âge de 3 ans



Puisque le polygone de fréquences a la forme d'une cloche, nous pouvons utiliser la loi normale comme modèle mathématique de la distribution. Cela signifie que nous serons en mesure, à l'aide de la loi normale, d'estimer des pourcentages qui n'apparaissent pas dans le tableau de la mise en situation, tel le pourcentage de garçons qui ont une taille comprise entre 95 cm et 97,5 cm à l'âge de 3 ans.

Quelle moyenne et quel écart type doit avoir la loi normale qui donnerait une bonne approximation de la distribution ?

On peut penser qu'en prenant la moyenne et l'écart type de la distribution on atteindra notre objectif: c'est ce que nous ferons. Pour les 500 garçons de 3 ans étudiés, la taille moyenne est de 95,7 cm et l'écart type de 3,3 cm ; on choisit donc la loi normale $N(95,7; 3,3^2)$ comme modèle mathématique de la distribution (voir le graphique à la page suivante).

Comment calculer une probabilité pour une variable aléatoire continue ?

À l'aide de l'expérience aléatoire qui suit, nous verrons en quoi le calcul d'une probabilité pour une variable aléatoire continue diffère de celui d'une variable aléatoire discrète et nous apprendrons à utiliser la loi normale pour calculer une probabilité.

Une expérience aléatoire consiste à piger un garçon parmi les 500 garçons étudiés. Soit la variable aléatoire X : «la taille du garçon pigé».

❓ Quelles sont les chances que le garçon pigé mesure 92,832 cm ?

$$P(X = 92,832) = \underline{\hspace{2cm}}$$

Contrairement à une variable aléatoire discrète, la probabilité qu'une variable aléatoire continue soit égale à un nombre précis est pratiquement nulle: la probabilité qu'un garçon mesure exactement 92,832000 cm est négligeable. Comme une variable continue peut prendre une infinité de valeurs, n'importe quelle valeur particulière a un poids nul et donc, une probabilité nulle. Ce n'est toutefois pas le cas d'un intervalle.

❓ Quelles sont les chances que le garçon pigé mesure 102 cm ou plus ?

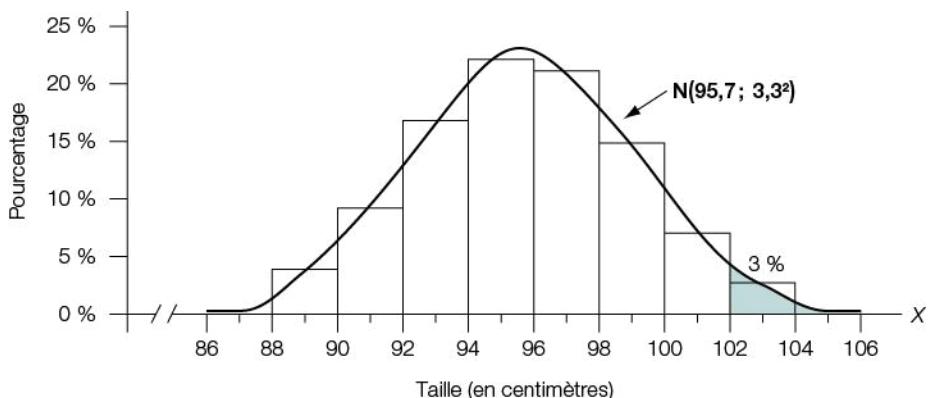
$$P(X \geq 102 \text{ cm}) = \underline{\hspace{2cm}}$$

NOTE

On considère que $P(X \geq 102 \text{ cm}) = P(X > 102 \text{ cm})$, puisque $P(X = 102 \text{ cm}) = 0$.

Voyons maintenant comment on peut utiliser la loi normale $N(95,7 ; 3,3^2)$ pour obtenir une approximation du pourcentage de garçons qui mesurent 102 cm ou plus, c'est-à-dire $P(X \geq 102 \text{ cm})$.

Répartition des garçons selon leur taille à l'âge de 3 ans



$$P(X \geq 102 \text{ cm}) = 3,0 \%$$

$$= \frac{\text{aire du rectangle de la classe } [102 ; 104]}{\text{aire totale de l'histogramme}} \times 100 \%$$

$$\approx \frac{\text{aire sous la courbe normale pour } X \geq 102 \text{ cm}}{\text{aire totale sous la courbe normale}} \times 100 \%$$

$$\approx \frac{\text{aire sous la courbe normale pour } X \geq 102 \text{ cm}}{1} \times 100 \%$$

$$\approx \text{aire sous la courbe normale } N(95,7 ; 3,3^2) \text{ pour } X \geq 102 \text{ cm} \times 100 \%$$

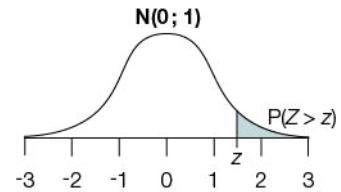
Comment trouver l'aire sous la courbe normale pour $X \geq 102$ cm ?

Bien sûr, on ne peut pas imaginer trouver une table mathématique qui donne, pour chaque courbe normale qui existe, l'aire d'une surface sous la courbe. Par contre, on a montré que lorsqu'une distribution suit un modèle normal, la distribution des cotes z de ses données suit également un modèle normal de moyenne 0 et d'écart type 1. On donne le nom de **loi normale centrée réduite** à cette loi, que l'on note $N(0; 1)$, et l'on désigne sa variable par la lettre Z (pour cote z) et ses valeurs par z . C'est en utilisant la table construite pour la loi $N(0; 1)$ et la cote z des données d'une distribution que l'on pourra trouver l'aire d'une surface sous une courbe normale quelconque.

La table de la loi normale centrée réduite $N(0; 1)$

La table de la loi normale centrée réduite (*voir la page 347*) donne l'aire sous la courbe $N(0; 1)$ située à la droite d'une valeur z et, par conséquent, le pourcentage de données ayant une cote z plus grande que cette valeur z .

On utilise la notation symbolique $P(Z > z)$ pour désigner ce pourcentage.



Pour estimer avec la loi normale le pourcentage de garçons qui mesurent 102 cm et plus, il faut calculer la cote z de 102, puis chercher dans la table $N(0; 1)$ l'aire sous la courbe se situant à droite de la cote z :

$$\begin{aligned} P(X \geq 102 \text{ cm}) &= P(Z \geq \text{cote } z \text{ de } 102) \\ &= P\left(Z \geq \frac{102 - 95,7}{3,3}\right) \\ &= P(Z > 1,91) \end{aligned}$$

Pour trouver l'aire associée à l'intervalle $z \geq 1,91$, on repère, dans la 1^{re} colonne, l'entier et la 1^{re} décimale de la cote z (soit 1,9) et, sur la 1^{re} ligne, la 2^e décimale (soit 0,01); l'aire cherchée se trouve à l'intersection de cette ligne et de cette colonne. On remarquera que la plus petite valeur possible pour z dans la table est 0.

Donc, $P(X \geq 102 \text{ cm}) = P(Z \geq 1,91) = 0,0281 \approx 2,8\%$.

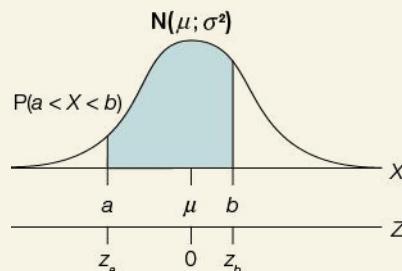
Il est à noter que l'écart n'est que de 0,2 point de pourcentage par rapport à la valeur de 3 % donnée dans le tableau de distribution.

Table $N(0; 1)$

z	0,00	0,01	0,02
1,7	0,0446	0,0436	0,0427
1,8	0,0359	0,0351	0,0344
1,9	0,0287	0,0281	0,0274
2,0	0,0228	0,0222	0,0217

Ce qui précède permet d'écrire la définition suivante.

Probabilité pour une variable aléatoire continue qui suit une normale $N(\mu; \sigma^2)$



$P(a < X < b) \approx$ aire sous la courbe normale $N(\mu; \sigma^2)$ entre $X = a$ et $X = b$

\approx aire sous la courbe normale $N(0; 1)$ entre la cote z de a et la cote z de b

3.4.4 La loi normale centrée réduite $N(0; 1)$

La mise en situation nous a donné un aperçu de l'utilité de la loi $N(0; 1)$ pour estimer le pourcentage de données d'une distribution normale plus grandes qu'une certaine valeur. Mais comment faudrait-il procéder pour estimer le pourcentage de données entre deux valeurs, par exemple le pourcentage de garçons qui mesurent entre 98 et 100 cm, avec une table qui ne donne que la surface à droite d'une cote z ?

La présente section est consacrée au développement de stratégies qui permettront de trouver l'aire de n'importe quelle surface sous la courbe $N(0; 1)$, et même de trouver une cote z associée à une aire donnée.

Recherche d'une aire sous une $N(0; 1)$

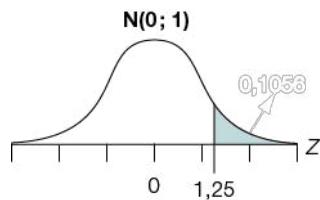
La recherche d'une aire sous la courbe $N(0; 1)$ associée à des cotes z sera grandement facilitée si l'on prend l'habitude de se poser la question suivante :

« Quelle aire peut facilement être trouvée dans la table à partir des cotes z données? »

EXEMPLES

- Quelles sont les chances qu'une donnée ait une cote z supérieure à 1,25 dans une distribution normale ?

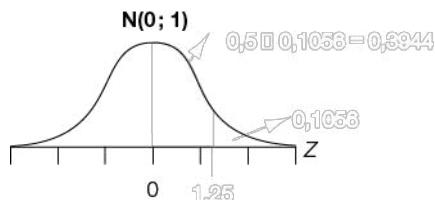
Solution



$$P(Z > 1,25) =$$

- Quelle est la probabilité qu'une donnée ait une cote z comprise entre 0 et 1,25 dans une distribution normale ?

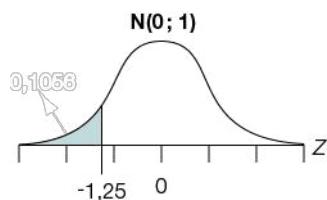
Solution



$$P(0 < Z < 1,25) =$$

- Quel pourcentage de données ont une cote z inférieure à -1,25 dans une distribution normale ?

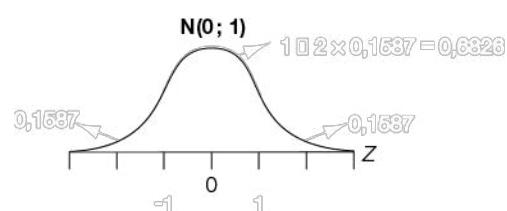
Solution



$$P(Z < -1,25) =$$

- Quel pourcentage de données d'une distribution normale ont une cote z comprise entre -1 et 1 ?

Solution



$$P(-1 < Z < 1) =$$

5. Quel pourcentage des données d'une distribution normale ont une cote z comprise entre 1 et 1,5 ?

Solution

NOTE

Il est bon de savoir qu'il existe deux autres types de table $N(0; 1)$: l'une donne l'aire sous la courbe entre 0 et une valeur z positive (par exemple, la valeur 0,3944 de la surface représentée à l'exemple 3 précédent) et l'autre, utilisée dans le logiciel Excel, donne l'aire sous la courbe de $-\infty$ à une valeur z quelconque. La marche à suivre pour résoudre un problème avec ces tables est identique à celle qui est décrite précédemment : on construit la solution à partir de la représentation graphique de l'aire cherchée et de l'aire qui peut facilement être obtenue dans la table utilisée.

Recherche d'une cote z pour une $N(0; 1)$

EXEMPLE 1

- a) Dans une distribution normale, 20 % des données ont une cote z plus grande qu'une certaine valeur. Laquelle ?

Solution

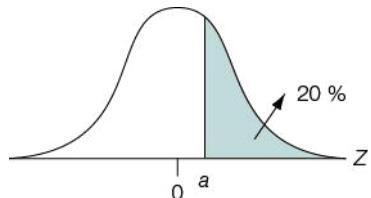
On cherche un nombre a tel que

$$P(Z > a) = 0,20$$

Dans la table, pour une aire de 0,20, on a :

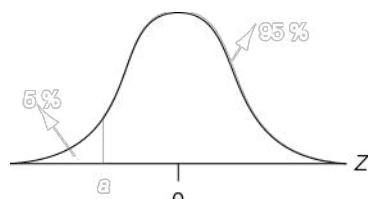
$$a = 0,84$$

Dans une distribution normale, 20 % des données ont une cote z supérieure à 0,84.



- b) Dans une distribution normale, 95 % des données ont une cote z plus grande que quelle valeur ?

Solution

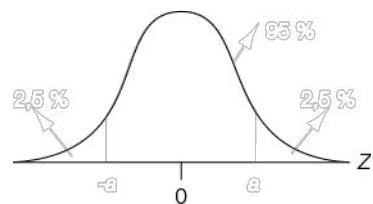


NOTE

Nous conviendrons de prendre, dans la table $N(0; 1)$, la cote z correspondant à la valeur la plus près de l'aire cherchée. Dans le cas où nous trouvons deux valeurs qui sont aussi près l'une que l'autre de l'aire cherchée, nous conviendrons de prendre la valeur se situant au centre des cotes z correspondant à ces deux valeurs.

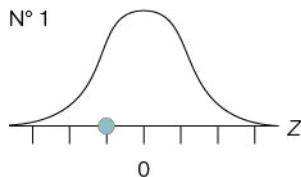
EXEMPLE 2

Pour une distribution normale, trouver une valeur a telle que le pourcentage de données qui ont une cote z comprise entre $-a$ et a soit de 95 %.

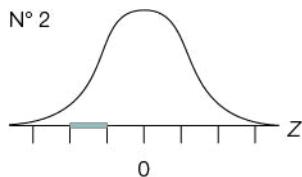
Solution**EXERCICES DE COMPRÉHENSION | 3.4**

1. Associer les expressions suivantes à une des représentations graphiques :

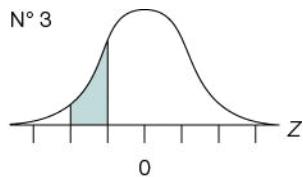
a) $-2 < Z < -1$ _____



b) $P(-2 < Z < -1)$ _____

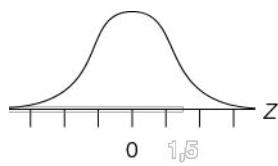


c) $Z = -1$ _____

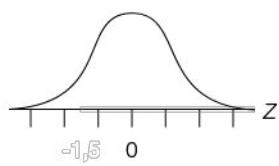


2. Représenter les intervalles suivants sur l'axe des Z de la $N(0 ; 1)$.

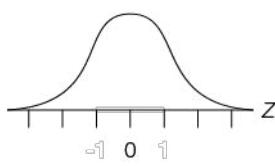
a) $Z < 1,5$



b) $Z > -1,5$

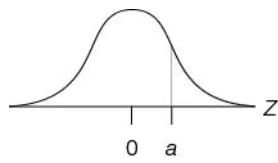


c) $-1 < Z < 1$

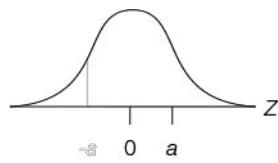


3. Représenter les surfaces dont l'aire correspond aux probabilités suivantes, sachant que $a > 0$.

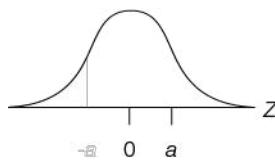
a) $P(Z > a)$



b) $P(Z > -a)$



c) $P(Z < -a)$



► 4. Utiliser l'information fournie par le graphique pour déterminer le pourcentage de données d'une distribution normale :

a) qui ont une cote z plus petite que a .

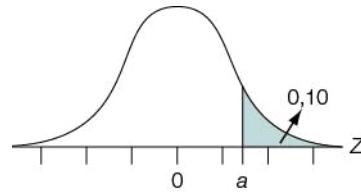
$$P(Z < a) = \underline{\hspace{2cm}}$$

b) qui ont une cote z plus petite que $-a$.

$$P(Z < -a) = \underline{\hspace{2cm}}$$

c) qui ont une cote z comprise entre $-a$ et a .

$$P(-a < Z < a) = \underline{\hspace{2cm}}$$



5. Pour la distribution $N(0 ; 1)$, trouver la valeur a telle que $P(Z > a) = 0,995$.

6. Pour la loi $N(0 ; 1)$, évaluer :

a) $P(Z = 2,35)$

b) $P(Z < 2,35)$

EXERCICES 3.5

1. Illustrer la surface correspondant à la probabilité demandée sur la courbe $N(0 ; 1)$ et la calculer.

a) $P(Z < 1,16)$

b) $P(0 < Z < 1,23)$

c) $P(Z = 1,5)$

d) $P(1,04 \leq Z < 2,12)$

e) $P(-1,96 < Z < 1,96)$

f) $P(Z < -0,97)$

g) $P(-1,28 \leq Z < -0,75)$

2. Sachant que $P(Z > 0,44) = 0,33$, situer sur la $N(0 ; 1)$ chacun des éléments suivants de cette expression :

a) $Z > 0,44$

b) 0,44

c) 0,33

3. Pour une $N(0 ; 1)$, trouver une valeur a telle que :

a) $P(Z > a) = 0,01$

b) $P(-a < Z < a) = 0,90$

c) $P(-a < Z < a) = 0,99$

4. Chercher la partie manquante des expressions suivantes et indiquer si elle correspond à une probabilité ou à la borne d'un intervalle sur l'axe Z de la $N(0; 1)$.
- $P(Z < ?) = 0,10$
 - $P(Z > 0) = ?$
 - $P(1,3 < Z < 1,9) = ?$
 - $P(Z > ?) = 0,09$
5. Vrai ou faux ? Pour une $N(0; 1)$:
- $0,52 = 0,3015$
 - $P(2,6) = 0,0047$
 - $P(Z > 2,35) = 0,0094$

3.4.5 La loi normale comme modèle mathématique

Maintenant que nous avons développé des habiletés à travailler avec la loi normale centrée réduite, poursuivons l'étude de la loi normale en tant que modèle mathématique d'une distribution de moyenne μ et d'écart type σ .

MISE EN

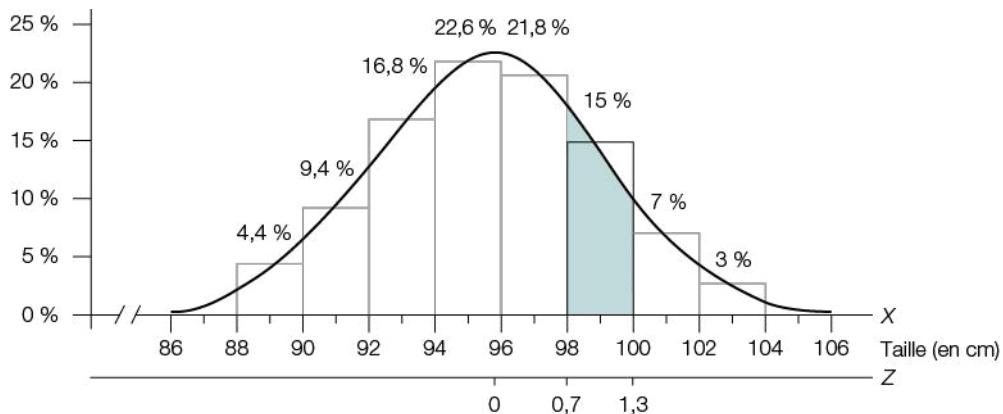
SITUATION (suite)

Nous savons que la loi normale $N(95,7; 3,3^2)$ peut servir de modèle mathématique pour la distribution de la taille des garçons âgés de 3 ans. Voyons comment cette loi normale peut nous aider à résoudre des problèmes portant sur la taille des garçons de cet âge.

Recherche d'une aire sous une $N(\mu; \sigma^2)$

La distribution ci-dessous indique que 15 % des garçons mesurent entre 98 et 100 cm à l'âge de 3 ans. Utilisons la loi normale $N(95,7; 3,3^2)$ pour estimer ce pourcentage.

Répartition des garçons selon leur taille à l'âge de 3 ans



En déterminant l'aire sous la courbe normale, entre 98 et 100 cm, nous obtiendrons une estimation du pourcentage cherché.

$$\begin{aligned}
 P(98 < X < 100) &\approx P(Z_{98} < Z < Z_{100}) \\
 &\approx P\left(\frac{98 - 95,7}{3,3} < Z < \frac{100 - 95,7}{3,3}\right) \\
 &\approx P(0,70 < Z < 1,30) \\
 &\approx 0,2420 - 0,0968 = 0,1452 \\
 &\approx 14,5 \%
 \end{aligned}$$

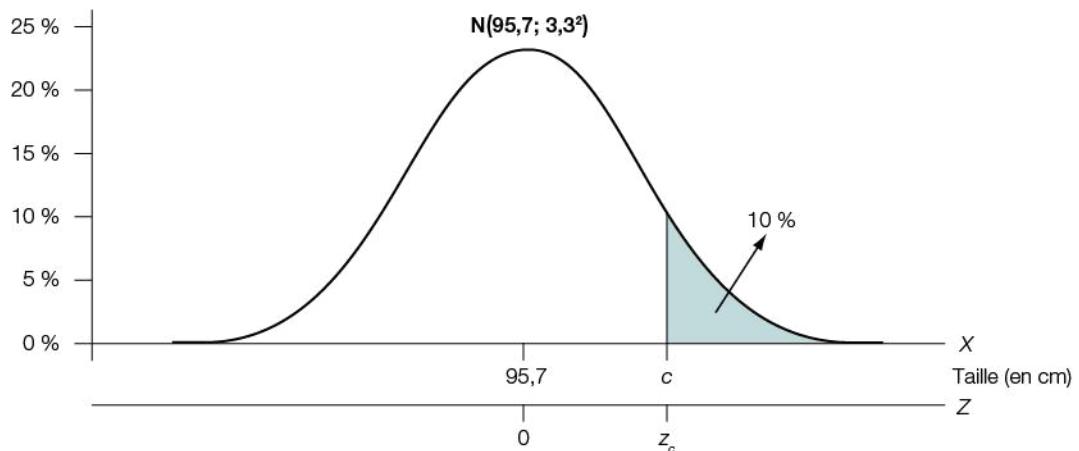
L'écart n'est que de 0,5 point de pourcentage par rapport à la valeur de 15 % obtenue pour la distribution.

NOTE

Plus la forme de l'histogramme (ou du polygone de fréquences) d'une distribution se rapproche d'une courbe normale, plus l'approximation effectuée en prenant la loi normale comme modèle mathématique est précise. Au chapitre 6, nous présentons le test d'ajustement du khi-deux qui sert à vérifier, en cas de doute, l'ajustement d'une distribution à une loi normale.

Recherche d'une valeur X pour une $N(\mu; \sigma^2)$

Samuel est un petit garçon de 3 ans particulièrement grand : on a dit à ses parents que seulement 10 % des garçons de son âge sont plus grands que lui. Sachant que la distribution de la taille d'un garçon de 3 ans suit une loi normale de moyenne 95,7 cm et d'écart type 3,3 cm, estimer la taille de Samuel.



Comme l'illustre le graphique, il s'agit de trouver une valeur c sur l'axe de la taille (X) telle que 10 % des garçons aient une taille plus grande que c ; par conséquent, la cote z de la taille de ces garçons sera plus grande que la cote z de c , que l'on note z_c .

- On cherche c , une taille, telle que $P(X > c) = 0,10$.
- En cote z on a : $P(Z > z_c) = 0,10$.

Selon la table, $z_c = 1,28$.

- La taille c se situe donc à 1,28 écart type à droite de la moyenne :

$$c = \text{moyenne} + 1,28 \times \text{écart type}$$

$$c = 95,7 \text{ cm} + 1,28 \times 3,3 \text{ cm}$$

$$c = 99,9 \text{ cm}$$

Selon le modèle normal, on peut estimer que Samuel mesure près de 100 cm.

Il est à noter que la distribution (voir la page 165) indique qu'il y a effectivement 10 % (7 % + 3 %) des garçons qui ont une taille de 100 cm et plus à l'âge de 3 ans.

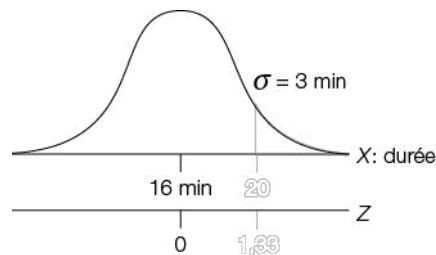
EXEMPLE

Louis prend tous les jours le traversier Québec-Lévis pour se rendre au travail. En moyenne, il fait le trajet entre chez lui et le quai d'embarquement en 16 minutes avec un écart type de 3 minutes. La distribution de la durée du trajet suit un modèle normal.

- a) Si le traversier quitte le quai à 8 h et que Louis quitte la maison à 7 h 40, quels sont les risques qu'il manque le bateau de 8 h?

Solution

Distribution de la durée du trajet (en min)
pour se rendre au traversier



- b) À quelle heure Louis doit-il quitter la maison s'il veut réduire le risque de manquer le traversier de 8 h à moins de 1 %?

Solution

EXERCICE DE COMPRÉHENSION | 3.5

La loi normale en gestion de stock

Une épicerie reçoit la livraison de son stock d'œufs une fois par semaine. Pour offrir un produit plus frais aux clients et diminuer la quantité d'œufs à entreposer, on décide d'augmenter la fréquence de livraison aux deux jours. Les statistiques de ventes de la dernière année indiquent que l'épicerie a vendu en moyenne 300 douzaines d'œufs aux deux jours avec un écart type de 25 douzaines. On a aussi observé que la distribution des ventes suivait un modèle normal.

- a) Si l'on décide de commander 325 douzaines d'œufs aux deux jours, quels sont les risques que cette quantité ne soit pas suffisante pour répondre à la demande?

- b) Si l'on veut réduire les risques d'être en rupture de stock à 5 %, combien de douzaines d'œufs doit-on commander au fournisseur ?

3.4.6 L'approximation de la loi binomiale par la loi normale

Dans le présent chapitre, nous avons élaboré plusieurs stratégies pour calculer la probabilité d'une variable aléatoire binomiale. Par exemple, pour une binomiale $B(n; p)$:

- Si $n \leq 20$, on utilise la table quand on peut y trouver la valeur de p , comme pour la binomiale $B(15; 0,2)$; sinon on utilise la fonction de probabilité, comme pour la binomiale $B(15; 0,23)$.
- Si $n > 20$, $p \leq 0,10$ et $np \leq 5$, on utilise la table de la loi de Poisson pour faire une approximation de la probabilité demandée, comme pour la $B(200; 0,01)$.

Mais, ces stratégies ne permettent pas de calculer $P(X \geq 60)$ pour une $B(200; 0,4)$.

En effet, avec $n > 20$ et $p > 0,10$, on ne peut utiliser l'approximation par la loi de Poisson, et il est impensable d'utiliser la fonction de probabilité, les calculs seraient trop longs. Que faire ?

Nous utiliserons la loi normale pour obtenir une approximation de la probabilité cherchée. Toutefois, les conditions suivantes sur np et nq doivent être respectées pour que l'on puisse utiliser la loi normale.

Approximation de la loi binomiale par la loi normale

Si $np \geq 5$ et $nq \geq 5$, alors

$$B(n; p) \approx N(\mu; \sigma^2) \quad \text{où} \quad \mu = np \text{ et } \sigma = \sqrt{npq}$$

MISE EN SITUATION

Bien que, pour une binomiale $B(10; 0,5)$, il ne soit pas nécessaire de se servir de la loi normale pour calculer une probabilité, nous allons utiliser cette loi binomiale pour montrer comment effectuer une approximation de la loi binomiale avec la loi normale.

- ?
- Utiliser la fonction de probabilité pour calculer la probabilité d'obtenir 4 succès en 10 épreuves pour une variable aléatoire X qui suit une $B(10 ; 0,5)$.

$$P(X = 4) = \binom{10}{4} (0,5)^4 (0,5)^6 =$$

- ?
- Les conditions requises pour effectuer une approximation de cette probabilité avec une loi normale sont-elles satisfaites ? Si oui, quelle loi normale faut-il utiliser ?

- ?
- Pour une variable aléatoire X qui suit une normale $N(5 ; 1,6^2)$, quelle est la probabilité que $X = 4$?

$$P(X = 4) = \underline{\hspace{2cm}}$$

Pourquoi n'obtient-on pas 20,5 % ? Où est le problème ?

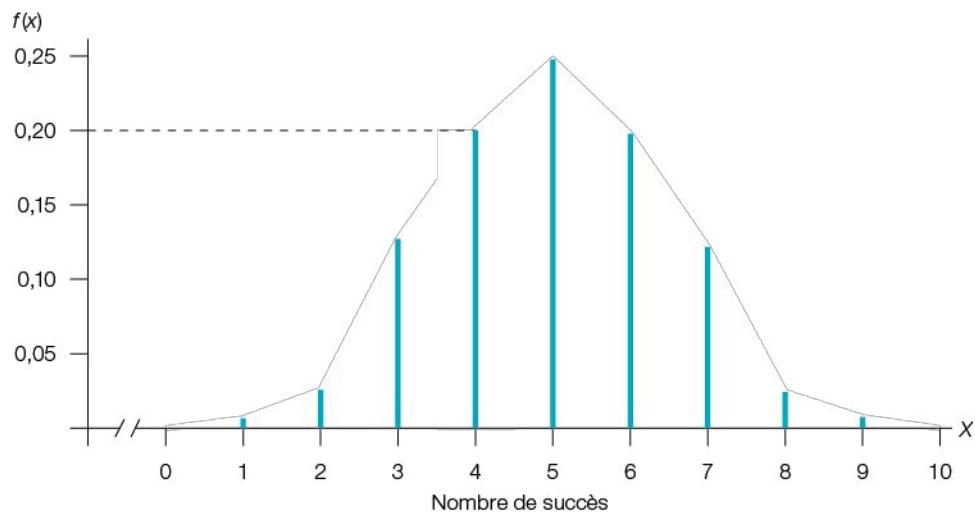
Correction de continuité

Voici comment on procède pour transformer une variable aléatoire discrète en une variable aléatoire continue pour effectuer une approximation d'une probabilité binomiale avec une loi normale. Dans la procédure qui suit, on écrit X_B pour indiquer que X suit une binomiale, et X_N pour indiquer que X suit une normale.

- ?
- Sur le graphique de la $B(10 ; 0,5)$ ci-dessous :

- Tracer la courbe normale $N(5 ; 1,6^2)$ sur le diagramme en bâtons de la $B(10 ; 0,5)$.
- Situer le bâtonnet dont la hauteur est égale à $P(X_B = 4) = 0,205$.

Distribution binomiale $B(10 ; 0,5)$



Le raisonnement suivant permet de calculer $P(X_B = 4)$ en utilisant la loi normale.

$$P(X_B = 4) = 0,205 = \text{hauteur du bâtonnet pour } X = 4$$

= $1 \times 0,205$: aire du rectangle de base 1 (de 3,5 à 4,5) et de hauteur 0,205

(Tracer ce rectangle sur le graphique.)

≈ aire sous la courbe normale entre 3,5 et 4,5

(Ombrer cette surface sur le graphique.)

≈ $P(3,5 < X_N < 4,5)$

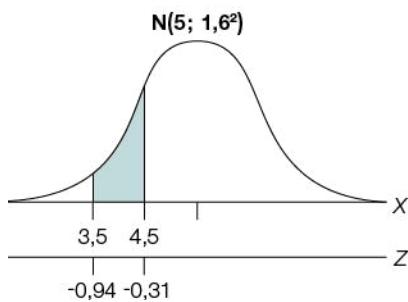
(en effectuant la correction de continuité)

$$\approx P\left(\frac{3,5 - 5}{1,6} < Z < \frac{4,5 - 5}{1,6}\right)$$

$$\approx P(-0,94 < Z < -0,31)$$

$$\approx 0,3783 - 0,1736$$

$$\approx 0,2047 \approx 20,5 \%$$



EXEMPLE

Pour la $B(10 ; 0,5)$ représentée à la page précédente, indiquer la correction de continuité qu'il faut faire pour obtenir une approximation des probabilités demandées en utilisant la loi normale.

a) $P(X_B > 7) =$

b) $P(6 \leq X_B \leq 8) \approx$

EXERCICE DE COMPRÉHENSION | 3.6

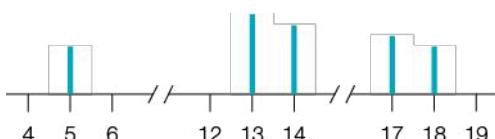
Indiquer la correction de continuité qu'il faut faire pour estimer les probabilités suivantes de la $B(50 ; 0,2)$ avec la loi normale.

Conseil

Une esquisse rapide de quelques bâtonnets de la binomiale aide à construire le bon raisonnement.

a) $P(X_B = 5) \approx$

b) $P(12 < X_B \leq 18) \approx$



EXEMPLE

Selon Statistique Canada, 56 % des internautes canadiens ont commandé des biens et des services en ligne en 2012. Compte tenu de cette statistique, quelles sont les chances de trouver plus de 100 personnes qui ont fait des achats en ligne dans un échantillon aléatoire de 200 internautes ?

Source: Statistique Canada. « Utilisation d'Internet et du commerce électronique par les particuliers, 2012 », *Le Quotidien*, octobre 2013.

Solution

EXERCICE DE COMPRÉHENSION | 3.7

Les jeunes sont-ils nombreux à jouer sur Internet ?

Une étude révèle que 57 % des internautes canadiens âgés de 16 à 24 ans jouent à des jeux en ligne. Si l'on prélève un échantillon aléatoire de 100 internautes dans ce groupe d'âge, quelle est la probabilité qu'au plus 55 d'entre eux jouent à des jeux sur Internet ?

Source: Statistique Canada. *Tableau 358-0153, CANSIM*, novembre 2013.

EXERCICES 3.6

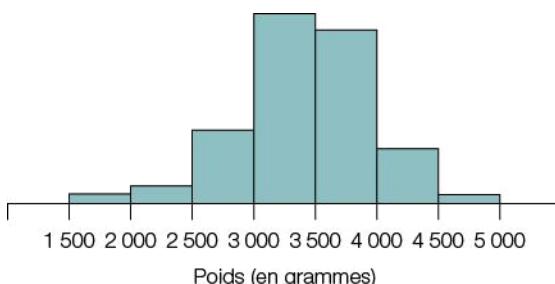
1. Voici la distribution du poids des garçons nés au Québec en 2011.

Répartition des nouveau-nés masculins selon le poids à la naissance, Québec, 2011

Poids à la naissance (en grammes)	Nombre de nouveau-nés
Moins de 2 000	852
[2 000; 2 500[1 531
[2 500; 3 000[6 290
[3 000; 3 500[16 274
[3 500; 4 000[14 871
[4 000; 4 500[4 718
4 500 et plus	777
Total	45 313

Source: Institut de la statistique du Québec. Août 2013.

Répartition des nouveau-nés masculins selon le poids à la naissance, Québec, 2011



a) Quelle loi normale peut être appliquée comme modèle mathématique à cette distribution ?

b) On classe dans la catégorie «Insuffisance de poids à la naissance» tous les nouveau-nés qui pèsent moins de 2 500 g. Déterminer le pourcentage de nouveau-nés masculins qui se trouvent dans cette catégorie :

- i) à partir des données du tableau de distribution ;
- ii) en estimant le pourcentage avec la loi normale.

c) Estimer, avec le modèle normal, le pourcentage de nouveau-nés qui pèsent entre 2 500 g et 4 000 g, et comparer ce résultat avec le pourcentage donné par le tableau de distribution.

2. a) Utiliser les indices suivants pour estimer l'âge de Nicole, infirmière dans un hôpital :

- L'âge moyen du personnel infirmier est de 42 ans et l'écart type, de 7 ans.
- La distribution de l'âge suit un modèle normal.
- Seulement 20 % du personnel infirmier est plus âgé que Nicole.

- b) Quel pourcentage du personnel infirmier a entre 39 ans et 56 ans ?

- c) Selon le modèle normal, l'âge de presque tout le personnel infirmier (99,7 %) serait compris entre quelles valeurs ?
3. On a établi que la distribution du quotient intellectuel (QI) suit un modèle normal dont la moyenne est 100 et l'écart type 15.
- Quel pourcentage de la population a un QI compris entre 92 et 108 ?
 - On dit qu'une personne présente une déficience intellectuelle si son QI est inférieur à 70. À combien pourrait-on estimer le pourcentage de personnes présentant une déficience intellectuelle dans la population ?
- Source:** Association des centres jeunesse du Québec. *Problèmes de santé mentale. Notions de pédopsychiatrie*.
- Claude prétend que 80 % de la population a un QI inférieur au sien. Quelle est la valeur de son QI ?
 - MENSA est une association dont les membres sont considérés comme supérieurement intelligents. Pour y être accepté, il faut un QI d'au moins 131. Sur 100 personnes dans la population, combien pourraient être acceptées dans cette association ?
4. On désire faire une approximation d'une binomiale par une normale. Indiquer la correction de continuité qu'il faut effectuer pour calculer les probabilités suivantes :
- $P(X_B > 12) \approx P(X_N > ?)$
 - $P(8 \leq X_B < 10) \approx P(? < X_N < ?)$
5. Une étude révèle que 39 % des travailleurs des secteurs des affaires, de la finance et de l'administration sont surqualifiés : ils occupent un emploi exigeant un niveau de scolarité inférieur à leur

qualification. Si l'on prélève un échantillon aléatoire de 400 travailleurs de ces secteurs d'activité, quelle est la probabilité que plus de 250 d'entre eux aient un emploi correspondant à leur qualification ?

Source: Institut de la statistique du Québec. *La surqualification au sein des grands groupes professionnels au Québec en 2012, 2013*.

6. On lance 100 fois une pièce de monnaie. Soit X : «nombre de faces obtenu en 100 lancers». Calculer les probabilités suivantes :
- $P(50 \leq X \leq 60)$
 - $P(49 < X < 61)$
 - $P(X = 54)$
 - $P(X > 62)$
7. La plupart des ordinateurs portables sont équipés d'une pile lithium-ion rechargeable. La durée de vie de ces piles suit un modèle normal dont la moyenne est de 60 mois et l'écart type, de 9,5 mois.
- Quelle est la probabilité que la pile d'un portable dure plus de 6 ans ?
 - Quelle est la probabilité que la pile d'un portable dure entre 46 et 74 mois ?
 - La garantie stipule que si la pile du portable dure moins de k mois, le fabricant s'engage à la remplacer. Quelle doit être la valeur de k si le fabricant ne veut pas remplacer plus de 0,5 % des piles ?
8. Un sondage est effectué auprès d'un échantillon aléatoire de 500 propriétaires d'une automobile afin de connaître le pourcentage de ceux-ci qui ont l'intention de changer de voiture d'ici 2 ans. Quelle est la probabilité qu'au moins 260 personnes de l'échantillon possèdent une automobile qui a plus de 5 ans, sachant que 55 % des automobiles immatriculées au Québec ont plus de 5 ans ?

Source: Société de l'assurance automobile du Québec. *Dossier statistique – Bilan 2012, accidents, parc automobile, permis de conduire*, juin 2013.

Les variables aléatoires

Variable aléatoire X

Fonction qui associe une valeur numérique à chaque résultat d'une expérience aléatoire.

Fonction de probabilité $f(x)$

Probabilité qu'une variable aléatoire X prenne une valeur x particulière.

$$f(x) = P(X = x) = P(A) \text{ où } A \text{ est l'événement antécédent de } x$$

Distribution de probabilité

Tableau qui donne la fonction de probabilité $f(x)$ associée à chaque valeur x de la variable aléatoire X .

$$\text{Espérance : } E(X) = \sum x f(x)$$

$$\text{Écart type : } \sigma = \sqrt{\sum [x - E(X)]^2 f(x)}$$

Les lois des probabilités

Binomiale $B(n; p)$	Poisson $Po(\lambda)$	Normale $N(\mu; \sigma^2)$
Contexte d'utilisation		
X est une variable aléatoire discrète <ul style="list-style-type: none"> • X: «nombre de succès en n épreuves» où $X = 0, 1, 2, \dots, n$ • n épreuves indépendantes • à chaque épreuve, il y a un succès de probabilité p et un échec de probabilité q 	X est une variable aléatoire discrète <ul style="list-style-type: none"> • X: «nombre de succès par région» où $X = 0, 1, 2, \dots$ • λ = moyenne de succès par région • région : temps, longueur, volume, surface, etc. 	X est une variable aléatoire continue (ou discrète s'il y a un grand nombre de valeurs différentes) <ul style="list-style-type: none"> • on indique dans la donnée du problème que X suit une loi normale • en théorie, $-\infty < X < \infty$
Espérance et écart type		
$E(X) = np$ $\sigma = \sqrt{npq}$	$E(X) = \lambda$ $\sigma = \sqrt{\lambda}$	$E(X) = \mu$, valeur connue ou à calculer σ , valeur connue ou à calculer

Démarche de résolution de problèmes

1. Identifier la variable aléatoire X et la loi de probabilité qui s'applique.
2. Exprimer la question en langage mathématique.
3. Appliquer la marche à suivre appropriée.
 - Si X suit une loi binomiale $B(n; p)$:
 - 1) La $B(n; p)$ est-elle dans la table binomiale ?

- 2) Peut-on faire une approximation de la $B(n; p)$ par :
- la loi de Poisson ? Critères : $n > 20$, $p \leq 0,10$ et $np \leq 5$
 $B(n; p) \approx Po(\lambda)$ où $\lambda = np$
 - la loi normale ? Critères : $np \geq 5$ et $nq \geq 5$
 $B(n; p) \approx N(\mu; \sigma^2)$ où $\mu = np$ et $\sigma = \sqrt{npq}$
- Attention !** Il faut faire une correction de continuité.

3) Utiliser la fonction de probabilité : $f(x) = \binom{n}{x} p^x q^{n-x}$

- Si X suit une loi de Poisson $Po(\lambda)$:
 - 1) Calculer λ pour la région étudiée, s'il y a lieu.
 - 2) Chercher la probabilité dans la table de Poisson $Po(\lambda)$.
- Si X suit une loi normale $N(\mu; \sigma^2)$:
 - 1) Représenter graphiquement la situation sur la courbe normale.
 - 2) Construire un raisonnement graphique, puis utiliser la table $N(0; 1)$.

EXERCICES RÉCAPITULATIFS

1. a) Pour calculer rapidement $P(X > x)$ pour les binomiales ci-dessous, dire s'il faut utiliser :
 - la table binomiale ;
 - la loi de Poisson (donner ses paramètres) ;
 - la loi normale (donner ses paramètres) ;
 - la fonction de probabilité.

i) $B(8; 0,22)$ iv) $B(250; 0,03)$
 ii) $B(15; 0,35)$ v) $B(25; 0,90)$
 iii) $B(30; 0,08)$

b) Calculer $P(X \geq 10)$ pour la loi binomiale $B(40; 0,3)$.
2. Vous avez droit à 3 essais pour obtenir face en lançant une pièce de monnaie.
 - Si vous obtenez face dès le 1^{er} lancer, vous gagnez 8 \$.
 - Si vous obtenez face après 2 lancers, vous gagnez 5 \$.
 - Si vous obtenez face après 3 lancers, vous gagnez 2 \$.
 - Si vous ne réussissez pas à obtenir face après 3 lancers, vous perdez 40 \$.

Accepteriez-vous de jouer à ce jeu? Justifier votre réponse par le calcul de l'espérance de gain.
3. Une entreprise fabrique des tiges métalliques pour un client qui les utilise comme rayons dans le montage de roues de bicyclette. Lorsque la machine qui coupe les tiges est bien réglée, la distribution de la longueur des tiges suit un modèle normal dont la moyenne est de 30 cm et dont l'écart type est égal à la précision de coupe de la machine, soit 0,1 cm.
 - a) Le contrat signé avec le client stipule que les tiges doivent avoir une longueur moyenne de 30 cm et qu'aucune tige ne doit s'éloigner de plus de 0,20 cm de cette moyenne, dans un sens ou dans l'autre; autrement dit, l'écart toléré est $E = 0,20$ cm. Si le fabricant respecte cette clause du contrat, quel pourcentage des tiges produites pourront être vendues au client?
 - b) Quel pourcentage des tiges produites ne pourront pas être vendues au client?
 - c) Le fabricant aimeraient bien diminuer le pourcentage de pertes. Pour ce faire, deux solutions s'offrent à lui: améliorer la précision de coupe afin de diminuer l'écart type ou demander au client s'il accepterait d'augmenter légèrement l'écart toléré. Il retient la seconde solution, car la première nécessiterait l'achat d'une nouvelle machine pour couper les tiges, ce qui est jugé trop

coûteux. Quelle devrait être la valeur de l'écart E toléré par le client pour que le fabricant puisse lui vendre 99 % de sa production ?

4. Les statistiques indiquent que lorsque l'équipe de hockey du cégep A affronte celle du cégep B, les probabilités qu'elle gagne, annule ou perde la partie sont respectivement de 60 %, de 10 % et de 30 %. Au cours de la prochaine saison, les deux équipes se renconteront 8 fois.
 - a) Quelle est la probabilité que l'équipe du cégep A gagne exactement 5 parties ?
 - b) Si l'équipe du cégep A gagne 7 parties sur 8 à la prochaine saison, peut-on en conclure que sa performance est vraiment exceptionnelle ? Justifier la réponse à l'aide de la valeur de la cote z .
 - c) Quelle est la probabilité que l'équipe du cégep A perde plus de 5 parties ?

5. Considérons la distribution suivante :



1. [15 ans; 20 ans] inclut les mères de moins de 15 ans;
[40 ans; 45 ans] inclut les mères de plus de 45 ans.

Source: Institut de la statistique du Québec. Août 2013.

- a) Quelle loi statistique peut être utilisée comme modèle mathématique de cette distribution ?
- b) Utiliser la loi normale pour estimer le pourcentage de nouveau-nés dont la mère a entre 35 et 40 ans. Donner l'écart entre cette estimation et le pourcentage réel.
- c) Selon le modèle normal, 10 % des nouveau-nés ont une mère ayant plus de quel âge ?

6. Y A-T-IL BEAUCOUP DE CENTENAIRES AU QUÉBEC?

En 2013, on dénombre 1 625 centenaires au Québec : 1 464 femmes et 161 hommes.

Source: Institut de la statistique du Québec. Décembre 2013.

- a) On pige sans remise 10 personnes parmi les 1 625 centenaires.
 - i) Pourquoi peut-on considérer que les tirages sont indépendants ?
 - ii) Quelle est la probabilité qu'il y ait exactement 8 femmes parmi les 10 personnes pigées ?
 - iii) Combien y a-t-il de façons d'obtenir exactement 8 femmes en 10 tirages ? Donner la probabilité de chacune de ces façons.
- b) Une compagnie d'assurance-vie évalue que la probabilité qu'une personne vive plus de 100 ans est de 0,000 03. Calculer la probabilité que, parmi 10 000 personnes assurées par cette compagnie, plus de 2 vivent plus de 100 ans.
7. Un conseiller municipal prétend que l'opinion des résidents de son arrondissement quant à l'établissement d'une maison d'accueil pour toxicomanes est partagée également entre ceux qui sont en faveur du projet et ceux qui sont contre. Un organisme d'aide aux toxicomanes commande un sondage sur le sujet. Dans un échantillon de 100 résidents, 60 sont en faveur du projet.
 - a) Si le conseiller municipal a raison, quelle est la probabilité d'obtenir au moins 60 répondants favorables au projet, dans un échantillon de 100 résidents ?
 - b) La probabilité obtenue devrait-elle faire douter de l'affirmation du conseiller municipal quant à la répartition des opinions des résidents de l'arrondissement ?
8. Avant de lancer un nouveau modèle de pneu sur le marché, un fabricant le soumet à des tests dans des conditions difficiles. Pendant les tests, sur une distance de 20 000 km, il y a eu en moyenne une crevaison. Si un acheteur utilise ce modèle de pneu dans les mêmes conditions que celles des tests, quelle est la probabilité, pour un pneu :
 - a) qu'il n'ait aucune crevaison pendant un voyage de 4 000 km ?
 - b) qu'il ait plus de 3 crevaisons sur un trajet de 40 000 km ?

PRÉPARATION À L'EXAMEN

Pour préparer votre examen, assurez-vous d'avoir les compétences suivantes :

	Si vous avez la compétence, cochez.
Les variables aléatoires	
• Construire la distribution de probabilité d'une variable aléatoire.	<input type="radio"/>
• Calculer l'espérance et l'écart type d'une variable aléatoire.	<input type="radio"/>
• Interpréter l'espérance et l'écart type d'une variable aléatoire.	<input type="radio"/>
• Calculer et interpréter l'espérance dans un contexte de jeu de hasard.	<input type="radio"/>
La loi binomiale	
• Reconnaître le contexte d'une expérience aléatoire binomiale.	<input type="radio"/>
• Calculer et interpréter l'espérance et l'écart type d'une variable aléatoire binomiale.	<input type="radio"/>
• Calculer une probabilité binomiale en utilisant :	
– la fonction de probabilité;	<input type="radio"/>
– la table de la loi binomiale;	<input type="radio"/>
– la loi de Poisson;	<input type="radio"/>
– la loi normale.	<input type="radio"/>
La loi de Poisson	
• Reconnaître le contexte d'une expérience de Poisson.	<input type="radio"/>
• Calculer et interpréter l'espérance et l'écart type d'une variable aléatoire de Poisson.	<input type="radio"/>
• Calculer une probabilité de Poisson en utilisant la table.	<input type="radio"/>
La loi normale	
• Connaître les caractéristiques propres à toute courbe normale.	<input type="radio"/>
• Résoudre des problèmes à l'aide du modèle normal.	<input type="radio"/>

Chapitre 4

La distribution d'échantillonnage et l'estimation



OBJECTIFS DU CHAPITRE

- Différencier les méthodes d'échantillonnage.
- Estimer une moyenne ou un pourcentage d'une population à partir de la moyenne ou du pourcentage d'un échantillon.

OBJECTIFS DU LABORATOIRE

Le laboratoire 4 vise, entre autres, à utiliser Excel pour construire et représenter un intervalle de confiance pour une moyenne.

Les lois de probabilité présentées au chapitre 3 sont définies en fonction de paramètres d'une population, soit la moyenne μ , l'écart type σ ou le pourcentage p . Or, ces paramètres sont souvent inconnus. Dans le présent chapitre, nous apprenons à estimer ces paramètres à partir des mesures (moyenne, écart type corrigé ou pourcentage) prises sur un échantillon prélevé dans la population. L'estimation de paramètres est le premier volet de l'inférence statistique. Le second volet, qui fait l'objet des chapitres 5 et 6, porte sur les tests d'hypothèse.

4.1 L'échantillonnage

Dans la présente section, nous examinons différentes méthodes employées pour prélever un échantillon dans une population.

4.1.1 Pourquoi faire des sondages ?

Il est important de savoir que, même si le recensement est la meilleure façon d'obtenir le portrait exact d'une population, certains facteurs le rendent souvent impossible à effectuer et obligent les chercheurs à procéder par sondage. Voici quelques-uns de ces facteurs :

- **La taille de la population.** Le coût de la collecte de renseignements auprès de toutes les unités de la population serait prohibitif par rapport à l'importance des résultats : est-il vraiment nécessaire de faire une enquête auprès de tous les Québécois afin de connaître leur opinion sur un projet de loi ?
- **Le temps.** Par sondage, en moins de trois jours, un gouvernement peut connaître le taux de satisfaction des électeurs à l'égard de son administration ; un recensement ne se fait pas aussi rapidement. On dit qu'un sondage, c'est la « météo » des politiciens : il leur permet de connaître rapidement le « temps » qu'il fait.
- **L'impossibilité de cerner la population.** Peut-on recenser toutes les personnes souffrant d'anorexie ?
- **L'aspect destructif de certains types de recensements.** Peut-on vérifier la résistance au choc des ampoules électriques ou la qualité des allumettes sans les détruire ?

Même quand il est possible de réaliser un recensement, on préfère souvent effectuer un sondage, car il a l'avantage de fournir un portrait assez juste d'une population en moins de temps et à un coût moindre qu'un recensement.

4.1.2 L'historique du sondage

La technique du sondage est mise au point par G. H. Gallup (1901-1984), un journaliste et statisticien américain né en Iowa et qui a présenté une thèse de doctorat sur les théories de l'échantillonnage. Gallup commence sa carrière professionnelle dès son entrée à l'université : à l'aide de sondages, il modifie le contenu du journal étudiant afin de le diffuser hors campus. En 1932, il mène son premier sondage préélectoral pour sa belle-mère, qui se présente pour le poste de secrétaire d'État en Iowa. Ensuite, il réalise des enquêtes de marketing pour ses propres affaires avant de fonder l'American Institute of Public Opinion. Aujourd'hui, l'institut Gallup est une entreprise multinationale qui compte des succursales et des associés dans plus de vingt pays ; son nom est devenu synonyme de « sondage ».

Au Québec, la méthode des sondages ne se développe qu'à la fin des années 1950. Le Groupe de recherches sociales dirige le premier sondage politique en 1959 pour la campagne électorale du Parti libéral du Québec. Entre 1960 et le début des années 1970, les principales maisons de sondage, parmi lesquelles on compte CROP, Sorécom et IQOP, sont fondées¹.

4.1.3 Comment choisir un échantillon?

Les méthodes d'échantillonnage, c'est-à-dire les méthodes employées pour constituer un échantillon, se divisent en deux groupes : les méthodes aléatoires ou probabilistes et les méthodes non aléatoires ou non probabilistes.

Méthodes d'échantillonnage aléatoire ou probabiliste

L'échantillonnage aléatoire repose sur un choix d'unités effectué au hasard dans la population. Ce n'est pas l'enquêteur qui choisit les unités ; c'est la méthode employée pour la sélection qui le fait. Une des caractéristiques de cette méthode est que chaque unité de la population a une probabilité mesurable et non nulle d'être choisie. Elle a comme avantage de permettre de généraliser les résultats de l'échantillon à l'ensemble de la population selon une théorie statistique reconnue. Son seul inconvénient est qu'il faut généralement posséder une liste de toutes les unités de la population avant de procéder au prélèvement de l'échantillon. Voici les quatre types d'échantillonnage aléatoire que l'on peut effectuer.

1. L'échantillonnage aléatoire simple

L'échantillonnage aléatoire simple est fondé sur le principe selon lequel toutes les unités de la population doivent avoir une chance égale de faire partie de l'échantillon. Pour respecter ce principe, on attribue un numéro à chaque unité de la population, puis on effectue un tirage au hasard des unités qui feront partie de l'échantillon au moyen d'un ordinateur ou d'une calculatrice. Le tirage peut se faire avec ou sans remise.

EXEMPLE

On veut choisir sans remise 5 individus dans une liste qui en contient 75. Après avoir attribué un numéro à chacun, on obtient, à l'aide de la touche Random (ou Rand) d'une calculatrice, les cinq numéros suivants :

8, 21, 43, 52, 63.

Méthodes pour générer des nombres aléatoires

Un ordinateur et une calculatrice sont des outils qui permettent de générer facilement des nombres aléatoires. Vérifions-le avec la situation présentée dans l'exemple précédent. Voici, selon l'outil utilisé, la marche à suivre pour générer aléatoirement 5 nombres compris entre 1 et 75 :

AU MOYEN DU LOGICIEL EXCEL

Pour obtenir un premier nombre aléatoire, on sélectionne une cellule de la feuille de calcul, puis on entre la formule =ALEA.ENTRE.BORNES(1 ; 75), où 1 et 75 sont le premier et le dernier numéro de la liste. Par la suite, on sélectionne la cellule contenant la formule, puis on clique sur la commande COPIER afin de copier la formule. On sélectionne ensuite une plage de quatre cellules, puis on clique sur la commande COLLER : on obtient quatre autres nombres aléatoires. Comme il s'agit d'une pige sans remise, on s'assure qu'il n'y a pas de doublons parmi les 5 nombres obtenus.

1. Source : André Tremblay. *Sondages. Histoire, pratique et analyse*, Boucherville, Gaëtan Morin, 1991.

AU MOYEN DE LA CALCULATRICE

Calculatrice scientifique (modèle Sharp EL-531W ou équivalent)

En mode normal, on appuie sur **2ND F**, sur **7** (**RANDOM** ou **RAND**, selon la calculatrice), puis on appuie deux fois sur **=**: on obtient des millièmes. À titre d'exemple, si 0,013 s'affiche, le nombre 13 est retenu (on multiplie par 1 000). Par contre, si 0,732 s'affiche, le nombre 732 n'est pas retenu, car il est supérieur à 75. On appuie de nouveau sur **=** pour obtenir d'autres nombres.

Calculatrice graphique (modèle TI-84 Plus)

On appuie sur **MATH**, on sélectionne le menu **PRB**, puis **5: RANDINT**. Pour obtenir le premier nombre aléatoire compris entre 1 et 75, on appuie sur **1**, sur **,**, sur **7**, sur **5**, puis sur **ENTER**. On appuie tout simplement sur **ENTER** pour obtenir d'autres nombres.

Avantages et désavantages de l'échantillonnage aléatoire simple

L'échantillonnage aléatoire simple assure le caractère représentatif de l'échantillon en utilisant une technique de sélection d'une grande simplicité. L'inconvénient de cette méthode réside dans le fait qu'elle nécessite une liste complète des unités statistiques de la population.

2. L'échantillonnage systématique

Pour sélectionner un échantillon en appliquant la méthode d'échantillonnage systématique, on doit prélever de façon systématique chaque k^{e} unité de la liste de la population. La valeur de k , que l'on nomme **pas de sondage**, dépend de la taille de la population et de celle de l'échantillon ; elle correspond approximativement à la valeur du rapport N/n .

EXEMPLE 1

On veut choisir 50 individus parmi une liste de 500, numérotés de 1 à 500. Comme $N=500$ et $n=50$, on prélèvera chaque 10^e ($500/50 = 10$) individu dans la liste ; le point de départ sera un nombre choisi au hasard entre 1 et 10. Par exemple, si le hasard désigne 4 comme point de départ, on retiendra les nombres 4, 14, 24, 34, 44, etc.

EXEMPLE 2

Lors du contrôle de la qualité en usine, l'échantillonnage systématique est très intéressant, car le prélèvement de l'échantillon peut être mécanisé. Par exemple, pour vérifier par échantillonnage le bon fonctionnement d'une machine qui emballle individuellement des petits gâteaux, on installe un bras mécanique qui expulse chaque 100^e gâteau qui tombe sur le tapis roulant à la sortie de la machine. Si la machine emballle 10 gâteaux à la minute, le contrôle de la qualité de l'emballage s'effectuera donc toutes les dix minutes, ce qui permet d'intervenir rapidement si un contrôle visuel indique que l'emballage présente une défectuosité.

Avantages et désavantages de l'échantillonnage systématique

La méthode d'échantillonnage systématique est plus agréable à employer que l'échantillonnage aléatoire simple dans le cas où la population et l'échantillon sont tous deux de grande taille, surtout si la sélection de l'échantillon se fait manuellement. Par contre, elle comporte un inconvénient, celui de la périodicité. Le problème peut se poser si la liste présente un caractère cyclique qui coïncide avec le « pas de sondage ». Il est alors probable que l'échantillon obtenu ne sera pas représentatif de la population. Par exemple, si le but de l'enquête est d'estimer le nombre de clients entrant dans un magasin au cours de certains mois, on peut prélever un échantillon de jours de ces mois et estimer le nombre de clients entrant dans le magasin aux jours choisis. Si les jours sont classés selon l'ordre habituel, un « pas de sondage » de 7, par exemple, donnera systématiquement le même jour de la semaine. Si l'on pense que la liste peut contenir un caractère cyclique, il est préférable d'effectuer un échantillonnage aléatoire simple.

3. L'échantillonnage stratifié

L'échantillonnage stratifié consiste à subdiviser la population en sous-groupes homogènes, ou strates, en fonction d'un ou de plusieurs critères : sexe, langue, province, ville de résidence, etc. On choisit ensuite un échantillon aléatoire dans chacune des strates, de manière qu'elle soit représentée dans l'échantillon proportionnellement à son importance dans la population.

EXEMPLE

Supposons que 60 % des étudiants d'un collège sont inscrits en techniques, et 40 % au secteur général ; pour former un échantillon de 120 étudiants en respectant la division en strates, on doit choisir au hasard $60\% \times 120 = 72$ étudiants en techniques et $40\% \times 120 = 48$ étudiants au secteur général.

Avantages et désavantages de l'échantillonnage stratifié

L'échantillonnage stratifié assure une bonne représentation des différentes strates de la population dans l'échantillon. Il permet aussi d'obtenir des estimations pour chaque strate. Toutefois, pour appliquer cette méthode, il faut avoir accès à la liste des unités de la population pour déterminer la répartition des strates.

4. L'échantillonnage par grappes

Il arrive souvent qu'une population soit répartie en grappes ou sous-ensembles plus ou moins homogènes : des électeurs d'une circonscription électorale sont répartis géographiquement en sections de vote d'environ 250 électeurs, des pompiers sont répartis dans des casernes sur le territoire d'une grande ville, des Amérindiens sont répartis dans des réserves, etc. L'échantillonnage par grappes consiste à tirer au hasard un certain nombre de grappes, puis à former l'échantillon avec tous les individus des grappes pigées.

EXEMPLE

Les étudiants de première année d'un cégep sont répartis dans les 30 groupes du premier cours de philosophie ; ces groupes sont numérotés de 1 à 30. On veut choisir un échantillon à l'aide de la méthode des grappes. On tire au hasard 4 nombres entre 1 et 30. Si, par exemple, on obtient les nombres 8, 13, 15 et 28, alors tous les étudiants de ces 4 groupes forment l'échantillon.

Avantages et désavantages de l'échantillonnage par grappes

Comparativement aux autres méthodes, l'échantillonnage par grappes a comme avantage qu'il n'est pas nécessaire d'avoir une liste de la population, mais seulement la liste des unités des grappes tirées au hasard. Cette méthode a par contre l'inconvénient de fournir des estimations habituellement moins précises que celles qu'on obtient avec un échantillonnage aléatoire simple, parce que des unités appartenant à une même grappe ont tendance à présenter des caractéristiques semblables. Il est toutefois possible de compenser cette perte de précision en augmentant la taille de l'échantillon.

Méthodes d'échantillonnage non aléatoire ou non probabiliste

L'échantillonnage non aléatoire repose sur un choix arbitraire des unités puisque c'est l'enquêteur qui choisit les unités et non le hasard. Pour cette raison, il serait hasardeux de généraliser à toute la population les résultats obtenus à partir de l'échantillon. On a néanmoins recours à des méthodes de ce type dans certains domaines, dont les études de marché ou les études de comportement des consommateurs.

1. L'échantillonnage à l'aveuglette ou accidentel

L'échantillonnage à l'aveuglette consiste à choisir les unités d'un échantillon de façon totalement arbitraire. Les résultats obtenus seront acceptables seulement si la population est relativement homogène, ce qui est rarement le cas. Autrement, certaines caractéristiques risquent d'être sous-représentées.

EXEMPLES

- Les interviews dans la rue, où les personnes interrogées sont sélectionnées au hasard des rencontres de l'intervieweur, fournissent rarement des résultats représentatifs de l'opinion de la population. En effet, seules les personnes qui se trouvent au même endroit que l'intervieweur ont des chances d'être interrogées, ce qui peut exclure plusieurs catégories de personnes. Par exemple, les étudiants qui sont en classe ou les travailleurs qui sont sur leur lieu de travail n'ont aucune chance de faire partie de l'échantillon si l'enquête se déroule un mardi avant-midi dans une rue commerciale.
- Un technicien prélève un échantillon d'eau dans un lac pour déterminer sa teneur en valeurs nutritives. Si l'on suppose que la composition de l'eau du lac est homogène, tout échantillon devrait donner des résultats assez semblables.

2. L'échantillonnage de volontaires

L'échantillonnage de volontaires consiste à choisir les individus d'un échantillon en faisant appel à des volontaires. On fait souvent appel à cette méthode en psychologie et en médecine quand la recherche peut s'avérer longue et exigeante, voire désagréable, pour les participants.

EXEMPLE

Des chercheurs publient une annonce dans les médias pour recruter des personnes qui accepteraient contre rémunération de tester un nouveau médicament.

3. L'échantillonnage par quotas

La méthode d'échantillonnage par quotas est largement employée dans les enquêtes d'opinion et les études de marché. Dans ce type d'échantillonnage, l'enquêteur choisit un échantillon aussi représentatif que possible des différentes strates de la population, selon le sexe, l'âge, la scolarité, etc. Cette méthode est peu coûteuse et assez rapide; de plus, elle ne requiert pas l'usage d'une liste de tous les individus de la population. Elle se distingue de l'échantillonnage stratifié par le fait que les individus ne sont pas choisis au hasard.

EXEMPLE

Dans une université, 70 % des étudiants sont inscrits au 1^{er} cycle, 20 % au 2^e cycle et 10 % au 3^e cycle. Pour constituer un échantillon de 200 étudiants de cette université, l'enquêteur choisit de façon arbitraire 140 étudiants au 1^{er} cycle, 40 au 2^e cycle et 20 au 3^e cycle.

NOTE

Les **sondages en ligne** dont les répondants proviennent d'un panel Web n'utilisent pas une méthode d'échantillonnage aléatoire. En effet, un **panel Web (ou Internet)** est composé d'internautes qui sont régulièrement invités à participer à des sondages d'opinion ou d'études de marché. Dans certains panels, les individus sont des volontaires, c'est-à-dire des gens qui se sont eux-mêmes inscrits pour remplir des sondages; dans d'autres, ils reçoivent une rétribution pour répondre au questionnaire. Ces derniers éléments et le fait que ce type d'échantillon laisse nécessairement de côté les personnes qui n'utilisent pas Internet (13 % des adultes québécois²) font en sorte que l'on ne peut pas projeter les résultats de ce type de sondage sur la population en général.

2. Source : CEFARIO. *Utilisation d'Internet au Québec en mai 2014*.

1. On désire choisir sans remise un échantillon de 6 personnes dans un groupe de 60. On numérote les individus du groupe de 1 à 60 et on procède à un tirage au hasard. Donner les numéros des individus de l'échantillon :
 - a) si l'on effectue un échantillonnage systématique dont le point de départ, tiré au hasard, est 3 ;
 - b) si l'on effectue un échantillonnage systématique dont le point de départ, tiré au hasard, est 8.
 2. Indiquer la méthode d'échantillonnage employée pour prélever les échantillons suivants :
 - a) Des biologistes font une étude sur une maladie qui attaque les arbres d'un parc. À l'aide d'une carte, ils ont divisé le territoire en 100 zones, puis ont choisi au hasard 10 zones en vue de procéder à l'analyse de chacun des arbres compris dans ces zones.
 - b) Une association réalise une enquête auprès d'un certain nombre de ses membres sélectionnés par tirage au sort dans la liste des membres.
 - c) Un journaliste de la télévision interroge des passants dans un centre commercial pour connaître leur opinion sur les résultats des dernières élections.
- d) Un chercheur demande la participation de jumeaux monozygotes pour une recherche médicale.
- e) Une usine produit 1 000 pièces par jour. Pour vérifier la qualité de celles-ci, on prélève chaque jour un échantillon de 50 pièces de la façon suivante : on retire une pièce de la production par 20 pièces produites en sélectionnant la première pièce au hasard entre la 1^{re} et la 20^e pièce produite.
- f) Dans le cadre d'une recherche auprès des membres de la coop étudiante d'un cégep, on désire constituer un échantillon de 30 membres en respectant la répartition des membres selon le sexe : 50 % de femmes et 50 % d'hommes. Pour ce faire, on sélectionne 15 femmes et 15 hommes au hasard des visites des membres à la coop.
- g) Dans le cadre de la recherche décrite en f), on sélectionne 15 femmes et 15 hommes au hasard dans la liste des membres de la coop.
- h) Parmi les échantillons décrits aux questions a) à g), lesquels sont aléatoires ?

4.2 La distribution d'échantillonnage d'une moyenne

Dans la présente section, nous nous intéresserons à la **loi du hasard**, qui décrit la relation entre la moyenne des données d'une variable de la population et la moyenne obtenue pour un échantillon aléatoire de ces données. Définissons d'abord certains termes et rappelons la notation des mesures présentées dans le chapitre 1.

Définition et notation

On donne le nom de **paramètres** aux mesures prises sur une population, et de **statistiques** à celles qui sont prises sur un échantillon. La notation suivante permet de distinguer ces différentes mesures.

Paramètres d'une population

- N : taille (nombre d'unités statistiques)
- μ : moyenne pour la variable étudiée
- σ^2 : variance pour la variable étudiée
- σ : écart type pour la variable étudiée

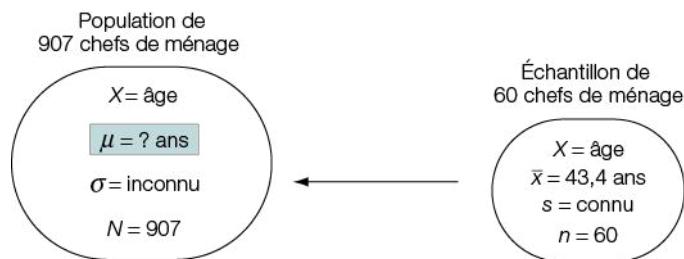
Statistiques d'un échantillon

- n : taille (nombre d'unités statistiques)
- \bar{X} : moyenne pour la variable étudiée
- s^2 : variance corrigée pour la variable étudiée
- s : écart type corrigé pour la variable étudiée

MISE EN SITUATION

On cherche à estimer la moyenne d'âge d'une population de 907 chefs de ménage résidant dans des habitations à loyer modique (HLM) à partir de la moyenne d'âge d'un échantillon aléatoire de 60 chefs de ménage.

Si la moyenne d'âge \bar{x} des 60 chefs de ménage de l'échantillon est de 43,4 ans, doit-on en conclure que la moyenne d'âge μ de la population est aussi de 43,4 ans, c'est-à-dire que $\mu = \bar{x}$?



De la même façon que la moyenne des notes de 6 étudiants d'une classe à un examen a très peu de chances d'être égale à la moyenne de l'ensemble des étudiants de la classe, il serait assez surprenant que la moyenne d'âge des 907 personnes de la population soit égale à celle des 60 personnes de l'échantillon. Il y a fort probablement un écart entre \bar{x} et μ .

Peut-on prédire les écarts possibles entre \bar{x} et μ avec une certaine certitude ? Par exemple, les chances que la moyenne \bar{x} soit à 0,5 an près (6 mois) de la moyenne μ de la population sont-elles grandes ? L'écart entre les deux moyennes peut-il être de 20 ans ?

Intuitivement, on peut penser qu'il y a plus de chances que l'écart entre \bar{x} et μ soit de 0,5 an que de 20 ans. Pour confirmer notre intuition, il faut connaître la loi de probabilité qui s'applique à cette expérience aléatoire dont la variable aléatoire est \bar{X} : « moyenne d'âge des 60 personnes pigées ».

Les caractéristiques de la distribution d'une moyenne d'échantillon

La réponse à la question suivante va permettre de trouver la loi de probabilité qui lie la moyenne μ de la population et la moyenne \bar{x} d'un échantillon :

« Si l'on connaît la moyenne μ de la population, peut-on prédire les valeurs que le hasard peut donner comme moyenne \bar{x} pour un échantillon de taille n tiré de cette population ? »

La réponse à cette question permettra par la suite de faire l'opération inverse : prédire la moyenne μ de la population à partir de la moyenne \bar{x} d'un échantillon aléatoire de taille n .

MISE EN SITUATION (suite)

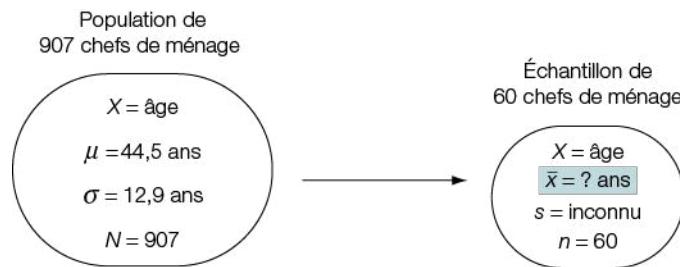
On sait que la moyenne d'âge μ des 907 chefs de ménage de la population est de 44,5 ans avec un écart type de 12,9 ans. Soit \bar{x} , la moyenne d'âge des 60 personnes d'un échantillon prélevé au hasard dans la population.

Posons les questions suivantes :

Q1. Quelles sont les chances que \bar{x} se situe à au plus 0,5 an de μ ?

Q2. Y a-t-il un modèle mathématique qui permet de prédire les valeurs possibles de \bar{x} ?

Représentons cette nouvelle situation où, à partir de la valeur connue de μ , on essaie de prédire la valeur de \bar{x} .



Valeurs possibles pour la moyenne \bar{x} d'un échantillon

Pour nous faire une idée des valeurs que le hasard peut donner comme moyenne d'échantillon, nous présentons ci-dessous les différentes moyennes \bar{x} obtenues par 145 étudiants à qui l'on avait demandé d'estimer l'âge moyen de 907 chefs de ménage en utilisant la moyenne d'âge de 60 chefs de ménage prélevés au hasard dans la population³.

\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}
40,0	42,3	43,0	43,5	43,8	44,3	44,8	45,1	45,6	47,0	
40,7	42,3	43,0	43,5	43,9	44,4	44,8	45,2	45,7	47,1	
40,8	42,3	43,1	43,5	44,0	44,4	44,8	45,2	45,8	47,2	
40,9	42,3	43,1	43,5	44,0	44,4	44,8	45,2	45,9	47,5	
41,4	S 42,5	43,2	43,5	44,0	44,4	44,9	45,2	45,9	47,6	
41,5	42,6	43,2	43,5	44,0	44,5	44,9	45,3	45,9	47,6	
41,6	42,6	43,3	43,6	44,1	44,5	44,9	45,3	46,3	47,7	
41,6	42,8	43,4	43,6	44,1	44,5	45,0	45,3	46,3	48,1	
41,7	42,8	43,4	43,6	44,1	13	44,6	45,0	45,4	46,3	48,8
41,7	42,8	43,4	43,6	44,2		44,6	45,1	45,4	46,4	48,9
41,8	42,9	43,4	43,7	44,2		44,6	45,1	45,5	46,4	
42,1	42,9	43,5	43,7	44,2		44,7	45,1	45,5	46,5	
42,2	42,9	43,5	43,7	44,2		44,8	45,1	45,6	J 46,5	
42,2	42,9	43,5	43,8	44,3		44,8	45,1	45,6	46,6	
42,3	43,0	43,5	43,8	44,3		44,8	45,1	45,6	46,7	

- ❓ Samuel et Jade sont deux des étudiants qui ont effectué ce travail. La moyenne \bar{x} qu'ils ont obtenue est désignée par les lettres **S** et **J** dans la liste. Pour ces deux échantillons, déterminer l'écart entre la moyenne d'âge des personnes de l'échantillon et la moyenne d'âge de 44,5 ans des personnes de la population.

Samuel: Écart entre \bar{x} et μ = _____

Jade: Écart entre \bar{x} et μ = _____

Dans les deux cas, on peut dire que la moyenne d'âge de l'échantillon se situe à 2 ans de la moyenne d'âge de la population. Dans l'estimation d'une moyenne, on accordera plus d'importance à la grandeur de l'écart (distance entre \bar{x} et μ) qu'au signe de cet écart (sens de l'écart par rapport à μ).

- ❓ Y a-t-il des étudiants qui ont obtenu une moyenne échantillonnale égale à la moyenne de 44,5 ans de la population ?

3. Résultats extraits d'un travail donné par l'auteure dans le cadre du cours Méthodes quantitatives.

On trouve _____ moyennes de 44,5 ans parmi les 145 moyennes de la liste.

Pourtant, nous avons indiqué au début de la présente section qu'une telle éventualité était peu probable. Y a-t-il là une contradiction ? En fait, nous observons ces égalités entre \bar{x} et μ parce que les moyennes sont calculées au dixième près. Si nous augmentions la précision à quatre décimales, ce que nous pouvons toujours faire avec une variable quantitative continue, aucune moyenne d'échantillon ne serait égale à la moyenne de la population.

Q1. Quelles sont les chances que \bar{x} se situe à au plus 0,5 an de μ ?

Plusieurs échantillons ont une moyenne qui se situe à au plus 0,5 an de μ (44,1 ans, 44,3 ans, 44,8 ans, etc.) ; en fait, tous ceux dans la liste dont la valeur est ombrée. Calculons le pourcentage des 145 échantillons qui présentent cette caractéristique :

$$P(44,5 - 0,5 \leq \bar{x} \leq 44,5 + 0,5) = P(44 \leq \bar{x} \leq 45) =$$

Comme environ 25 % des échantillons prélevés ont une moyenne qui se situe à au plus 0,5 an de μ , on peut estimer qu'un étudiant avait 25 % de chances de piger un tel échantillon.

Distribution des valeurs possibles pour la moyenne \bar{x} d'un échantillon

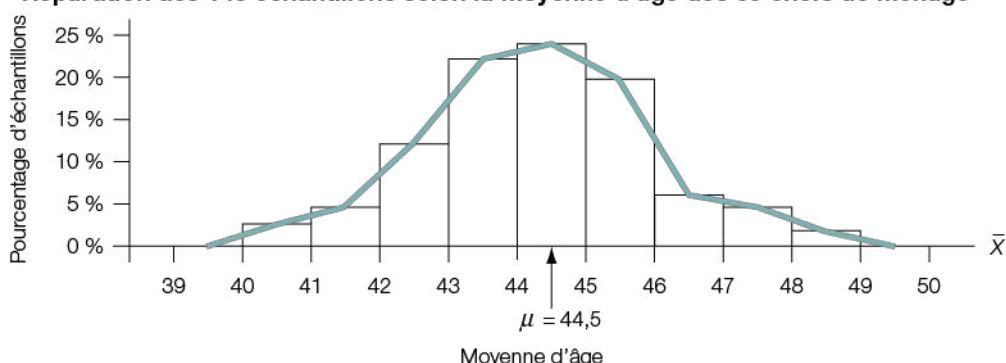
Q2. Y a-t-il un modèle mathématique qui permet de prédire les valeurs possibles de \bar{x} ?

Pour répondre à cette question, construisons le tableau et l'histogramme de la distribution des 145 moyennes \bar{x} .

Répartition des 145 échantillons selon la moyenne d'âge des 60 chefs de ménage

Moyenne d'âge	Nombre d'échantillons	Pourcentage
40 ans $\leq \bar{x} < 41$ ans	4	2,8 %
41 ans $\leq \bar{x} < 42$ ans	7	4,8 %
42 ans $\leq \bar{x} < 43$ ans	18	12,4 %
43 ans $\leq \bar{x} < 44$ ans	33	22,8 %
44 ans $\leq \bar{x} < 45$ ans	35	24,1 %
45 ans $\leq \bar{x} < 46$ ans	29	20,0 %
46 ans $\leq \bar{x} < 47$ ans	9	6,2 %
47 ans $\leq \bar{x} < 48$ ans	7	4,8 %
48 ans $\leq \bar{x} < 49$ ans	3	2,1 %
Total	145	100,0 %

Répartition des 145 échantillons selon la moyenne d'âge des 60 chefs de ménage



NOTE

On désigne l'axe horizontal de l'histogramme par la lettre \bar{X} majuscule, car tous les nombres qui s'y trouvent sont des moyennes \bar{x} d'échantillons.

Les 145 moyennes échantillonnelles semblent se distribuer selon un modèle normal autour de μ . En fait, si au lieu de prélever 145 échantillons, nous avions prélevé tous les échantillons possibles (il y en a $\binom{907}{60} = 4,7 \times 10^{94}$), la distribution de probabilité de la variable aléatoire \bar{X} , que l'on appelle **distribution d'échantillonnage de \bar{X}** suivrait un modèle normal dont la moyenne serait effectivement de 44,5 ans, soit la valeur de la moyenne de la population. Comme la moyenne de la courbe normale correspond à la moyenne des moyennes de tous les échantillons possibles, on la note $\mu_{\bar{x}}$ (lire mu de x barre) :

$$\mu_{\bar{x}} = \mu = 44,5 \text{ ans}$$

Pour utiliser cette loi normale, il faut connaître son écart type. Comme ce dernier correspond à l'écart type des moyennes de tous les échantillons possibles, on le note $\sigma_{\bar{x}}$ (lire sigma de x barre). Un théorème mathématique, qui porte le nom de **théorème central limite**, a pu établir qu'il existe un lien entre l'écart type $\sigma_{\bar{x}}$ et l'écart type σ de la population. Ce lien se traduit par l'égalité suivante :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\text{écart type de la population}}{\sqrt{\text{taille de l'échantillon}}}$$

Toutefois, pour un échantillon sans remise, si la taille N de la population est considérée comme petite par rapport à la taille n de l'échantillon (on considère que c'est le cas quand $N < 20n$)⁴, on doit alors multiplier σ/\sqrt{n} par le **facteur de correction** $\sqrt{(N-n)/(N-1)}$.

 Calculer l'écart type $\sigma_{\bar{x}}$ de la distribution d'échantillonnage de \bar{X} .

Vérifions d'abord s'il faut utiliser le facteur de correction :

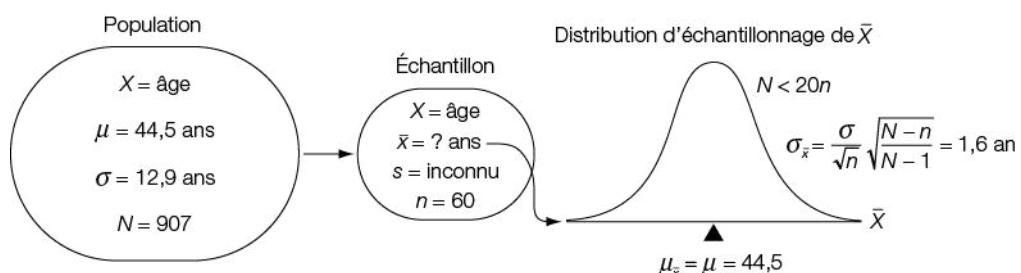
La population est-elle petite par rapport à l'échantillon : $N < 20n$? _____

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} =$$

(Pour les 145 échantillons de la mise en situation, la moyenne est de 44,3 ans et l'écart type de 1,7 an, ce qui est assez près des résultats ci-dessus.)

Conclusion

De ce qui précède, on conclut que la distribution d'échantillonnage de \bar{X} suit une loi normale de moyenne $\mu_{\bar{x}} = 44,5$ ans et d'écart type $\sigma_{\bar{x}} = 1,6$ an.



4. Dans certains ouvrages, on donne le nom de «taux de sondage» au rapport (n/N) et l'on considère que la population est petite si ce rapport est supérieur à 5 %, ce qui est équivalent à $N < 20n$. En privilégiant cette dernière expression, on élimine la confusion possible entre «un taux de sondage supérieur à 5 %» et «un risque d'erreur inférieur à 5 %» présenté au chapitre 5.

Le modèle normal confirme notre intuition voulant qu'il existe une forte probabilité que la moyenne d'âge \bar{x} des 60 personnes pigées soit près de la moyenne d'âge μ de la population plutôt que très éloignée de celle-ci.

? Peut-on prédire la plus petite et la plus grande valeur que le hasard peut donner pour \bar{x} ?

Sachant que dans une distribution normale presque toutes les valeurs (99,7 %) se trouvent à ± 3 écarts types de la moyenne, en négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues, on obtient ces valeurs ainsi :

- plus petite moyenne échantillonnale : $\bar{x}_{\min} = \underline{\hspace{10cm}}$
- plus grande moyenne échantillonnale : $\bar{x}_{\max} = \underline{\hspace{10cm}}$

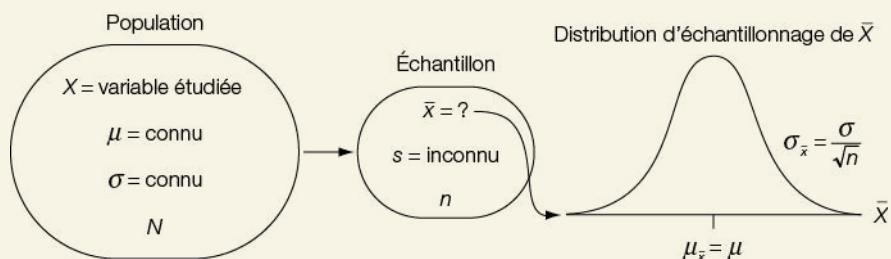
(La liste des 145 moyennes \bar{x} de la page 193 ne compte effectivement aucune valeur inférieure à 39,7 ans ou supérieure à 49,3 ans.)

Le théorème suivant permet de répondre à la question posée en début de section :

« Si l'on connaît la moyenne μ de la population, peut-on prédire les valeurs que le hasard peut donner comme moyenne \bar{x} pour un échantillon de taille n tiré de cette population ? »

Théorème central limite

Si, dans une population de taille N , de moyenne μ et d'écart type σ , on prélève sans remise un échantillon aléatoire de taille n , la distribution d'échantillonnage de \bar{X} présente alors les caractéristiques suivantes :



- Sa moyenne $\mu_{\bar{x}}$ est égale à la moyenne μ de la population : $\mu_{\bar{x}} = \mu$.
- Son écart type est :

$$\sigma_{\bar{x}} = \begin{cases} \frac{\sigma}{\sqrt{n}} & \text{si la population est grande } (N \geq 20n); \\ \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} & \text{si la population est petite } (N < 20n). \end{cases}$$

On donne le nom de facteur de correction à l'expression $\sqrt{(N-n)/(N-1)}$.

- Sa forme est normale si l'on a l'une ou l'autre des conditions suivantes :
 - la taille de l'échantillon est supérieure ou égale à 30 ($n \geq 30$) ;
 - la population suit un modèle normal.

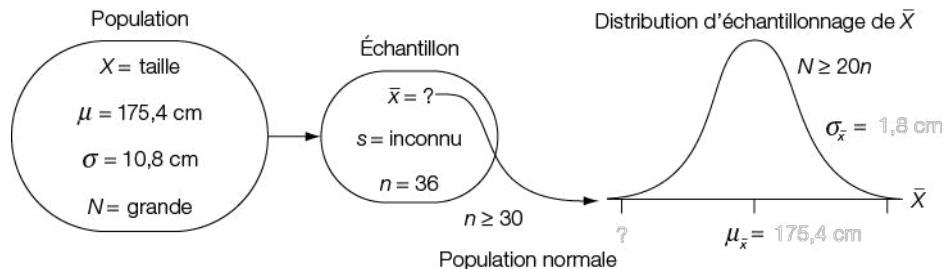
NOTE

- En pratique, il est rare qu'un échantillon soit prélevé avec remise ; quand c'est le cas, l'écart type $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, quelle que soit la taille de la population.
- Lorsque la taille de la population n'est pas spécifiée dans un problème, c'est que l'on considère que cette population est de grande taille.

EXEMPLE 1

La taille moyenne des hommes québécois de 20 ans et plus suit une loi normale dont la moyenne est de 175,4 cm et dont l'écart type est de 10,8 cm. On prélève au hasard 36 hommes parmi les Québécois de 20 ans et plus. On s'intéresse à la taille moyenne des 36 hommes pigés.

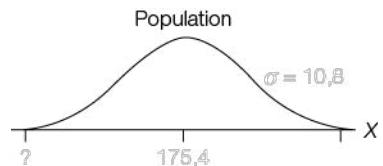
Source: Institut de la statistique du Québec. *Portraits et trajectoires*, novembre 2007.



a) En négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues avec un modèle normal :

- indiquer la plus petite taille que l'on puisse trouver dans la population des hommes de 20 ans et plus.

Solution



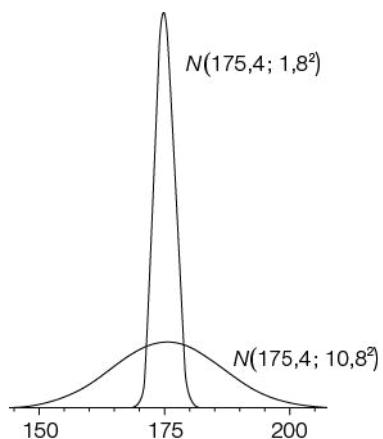
- indiquer la plus petite moyenne que le hasard puisse donner pour la taille des 36 hommes pigés.

Solution

En comparant les réponses en i) et en ii), on note que $\bar{x}_{\min} \neq x_{\min} = 143 \text{ cm}$. Ce résultat n'est pas surprenant, car pour que la taille moyenne d'un échantillon de 36 hommes soit de 143 cm, il faudrait que le hasard nous donne presque uniquement des hommes de très petite taille (autour de 143 cm), ce qui est pratiquement impossible.

La variation des moyennes \bar{x} sera toujours plus petite (σ/\sqrt{n}) que la variation des valeurs x de la population (σ). Pour illustrer cette idée, on peut faire une analogie avec les résultats à un examen: une note sur 100 points peut varier entre 0 et 100, mais la moyenne pour un groupe de 30 étudiants ne variera pas autant; par exemple, elle pourrait se situer entre 55 et 85.

Le graphique ci-contre est une représentation à l'échelle des courbes normales de la distribution de la taille X de la population et de la distribution d'échantillonnage de \bar{X} .



- b) Quelles sont les chances que la moyenne échantillonnale se situe à au plus 1 cm de la moyenne μ de la population ?

Solution

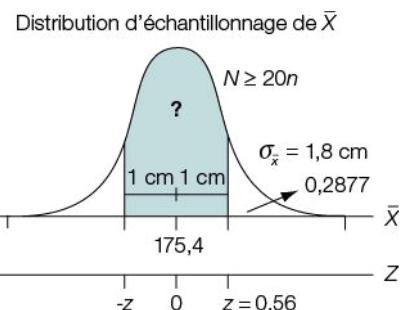
Pour un écart de 1 cm entre \bar{x} et μ , on a :

$$z = \frac{1 \text{ cm}}{1,8 \text{ cm}} = 0,56$$

Les chances que la grandeur de l'écart entre \bar{x} et μ soit d'au plus 1 cm se calculent ainsi :

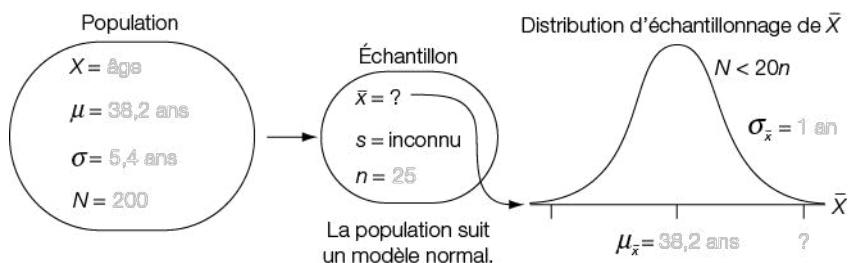
$$\begin{aligned} P(|\bar{x} - \mu| \leq 1 \text{ cm}) &= P(-0,56 < Z < 0,56) = 1 - 2 \times 0,2877 \\ &= 0,4246 \end{aligned}$$

Il y a donc 42,5 % de chances que la moyenne \bar{x} se situe à au plus 1 cm de μ .



EXEMPLE 2

La moyenne d'âge des 200 travailleurs d'une entreprise est de 38,2 ans avec un écart type de 5,4 ans. La distribution de l'âge des travailleurs suit un modèle normal. On pique un échantillon de 25 travailleurs et l'on s'intéresse à leur moyenne d'âge.



- a) Indiquer la plus grande moyenne échantillonnale possible, en négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues.

Solution

- b) Trouver une valeur E telle qu'il y ait 95 % de chances que la moyenne échantillonnale \bar{x} se situe à au plus E ans de la moyenne μ de la population.

Solution

On cherche E tel que $P(|\bar{x} - \mu| \leq E \text{ ans}) = 95\%$.

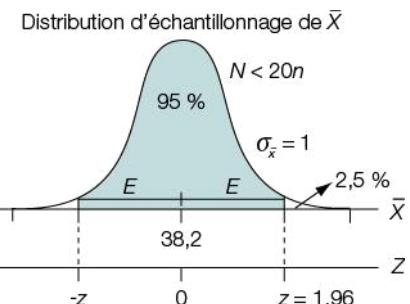
On a $P(Z > z) = 2,5\% = 0,025$; donc $z = 1,96$.

Pour un écart égal à E ans entre \bar{x} et μ on a :

$$z = \frac{E}{\sigma_{\bar{x}}} \Rightarrow E = z \sigma_{\bar{x}}$$

D'où $E = 1,96 \times 1 \approx 2$ ans.

Pour 95 % des échantillons possibles, l'écart entre \bar{x} et μ sera d'au plus 2 ans.

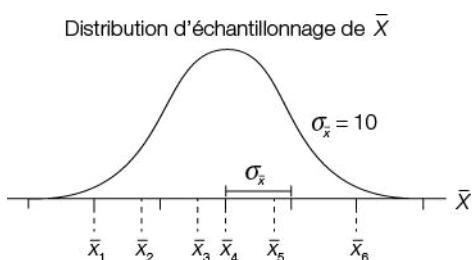


- c) Vrai ou faux ? Pour calculer l'écart maximal entre \bar{x} et μ associé à une certaine probabilité, il n'est pas nécessaire de connaître les valeurs de \bar{x} et de μ .

EXERCICES DE COMPRÉHENSION | 4.1

1. On prélève six échantillons de même taille dans une population. La moyenne de chacun est indiquée sur la distribution d'échantillonnage de \bar{X} ci-dessous.

a) Quelle moyenne \bar{x} a une valeur qui semble égale à la moyenne μ de la population ?

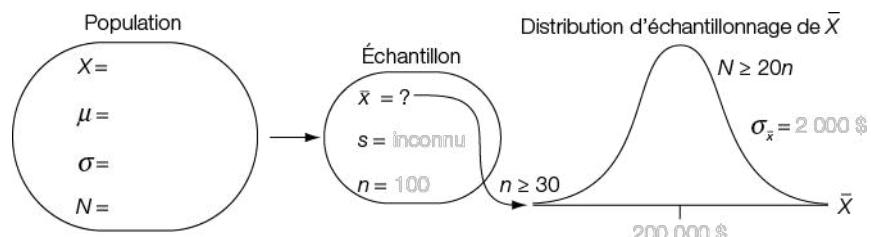


b) Quelles moyennes \bar{x} sont situées à au plus 15 unités de la moyenne μ de la population ?

c) Quelles moyennes \bar{x} semblent situées à une même distance de la moyenne μ de la population ?

2. Supposons que le revenu moyen des 3 000 médecins d'une région soit de 200 000 \$, avec un écart type de 20 000 \$. On s'intéresse à la moyenne de revenu des 100 médecins d'un échantillon aléatoire prélevé parmi ces 3 000 médecins.

a) Inscrire dans les 2 ovales le nom de la variable étudiée, les paramètres de la population et les statistiques de l'échantillon qui reflètent la situation décrite ci-dessus.



- b) La distribution d'échantillonnage de \bar{X} suit un modèle normal, car on a une des conditions, soit _____.

La moyenne $\mu_{\bar{x}}$ de cette distribution est _____. Incrire cette valeur sur la courbe normale.

La notation employée pour symboliser l'écart type de cette distribution est _____.

Faut-il utiliser le facteur de correction dans le calcul de cet écart type ? _____

Calculer l'écart type $\sigma_{\bar{x}}$, puis inscrire sa valeur sur la courbe normale.

- c) Encercler la ou les valeurs qu'on a peu de chances d'obtenir comme moyenne de revenu pour les 100 médecins de l'échantillon prélevé.

$\bar{x} = 202\,500 \$$ $\bar{x} = 193\,800 \$$ $\bar{x} = 200\,500 \$$ $\bar{x} = 198\,000 \$$ $\bar{x} = 206\,500 \$$

- d) Quelles sont les chances qu'il y ait un écart d'au plus 2 000 \$ entre le revenu moyen des 100 médecins de l'échantillon et celui des médecins de la population ? Représenter graphiquement la situation.

- e) Compléter l'énoncé.

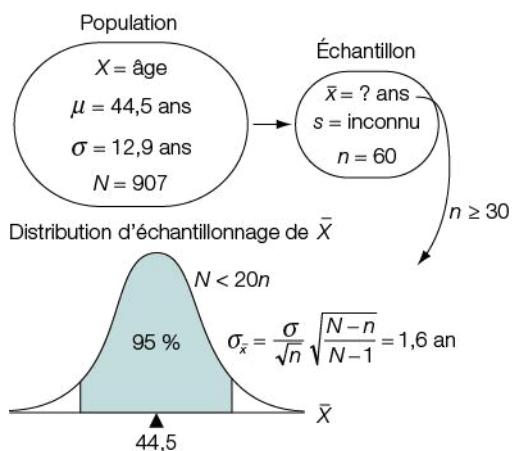
Il y a 95 % de chances que l'écart entre le revenu moyen \bar{x} des 100 médecins de l'échantillon et le revenu moyen μ des médecins de la population soit d'au plus _____ \$.

Représenter graphiquement la situation.

EXERCICES 4.2

1. Considérons de nouveau la mise en situation de la section 4.2 (*voir la page 192*), où 145 étudiants avaient prélevé un échantillon de taille 60 dans une population constituée de 907 chefs de ménage dont la moyenne d'âge était de 44,5 ans, avec un écart type de 12,9 ans.

- a) Nous avons établi que, si l'on prélevait tous les échantillons possibles, la distribution des moyennes échantillonnelles suivrait la loi normale $N(44,5 ; 1,6^2)$. La zone bleue sous la courbe normale contient 95 % des échantillons possibles. Pour les échantillons situés dans cette zone, quelle est la valeur maximale de l'écart entre \bar{x} et μ ?



- b) Utiliser la liste des moyennes échantillonnelles obtenues par les 145 étudiants (*voir la page 193*) pour répondre aux questions suivantes.
- Combien d'étudiants ont obtenu une moyenne \bar{x} qui ne se situe pas dans la zone bleue du graphique?
 - À quel pourcentage des 145 échantillons étudiés le nombre trouvé en i) correspond-il?
 - On sait que, dans une distribution normale, 68,3 % des données sont situées à au plus un écart type de la moyenne de la courbe normale. Donc, 68,3 % de tous les échantillons possibles devraient avoir une moyenne \bar{x} se situant à au plus 1,6 an de la moyenne μ de la population. Parmi les 145 étudiants qui ont prélevé un échantillon, quel pourcentage ont obtenu une moyenne d'échantillon située à au plus 1,6 an de la moyenne μ de la population?

2. Une entreprise fabrique des câbles d'acier. On désire vérifier si le diamètre (X) des câbles de la production est bien conforme aux normes : une distribution

normale avec un diamètre moyen de 0,90 cm et un écart type de 0,06 cm. Pour ce faire, on prélève un échantillon de 36 câbles dans la production. Le diamètre moyen des 36 câbles est de 0,88 cm, avec un écart type corrigé de 0,075 cm.

- a) Donner les valeurs de μ , σ , \bar{x} , s , $\mu_{\bar{x}}$ et $\sigma_{\bar{x}}$.
- b) En négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues :
- déterminer les valeurs entre lesquelles le diamètre des câbles de la production peut se situer.
 - déterminer les valeurs entre lesquelles le diamètre moyen d'un échantillon de 36 câbles devrait se situer. Est-ce que la moyenne échantillonnelle obtenue se situe entre ces deux valeurs?
3. Luc a prélevé au hasard 36 étudiants parmi les 3 000 étudiants d'un cégep où la moyenne d'heures de cours par semaine est de 23 heures, avec un écart type de 3 heures. Il a ensuite calculé la moyenne et l'écart type corrigé du nombre d'heures de cours hebdomadières des étudiants de son échantillon, ce qui lui a donné 22 heures et 2,5 heures respectivement.
- i) Indiquer la valeur des moyennes μ , \bar{x} et $\mu_{\bar{x}}$.
 - Donner la valeur des écarts types σ , s et $\sigma_{\bar{x}}$.
 - Tracer la courbe de la distribution d'échantillonnage de \bar{X} et y situer la moyenne de 23 heures et celle de 22 heures, en tenant compte de l'écart type de cette courbe normale.
- b) En négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues, indiquer la plus petite moyenne que Luc pouvait obtenir pour son échantillon.
- Dans 90 % des cas, pour un échantillon de 36 étudiants pigés parmi les 3 000 étudiants du cégep, quel devrait être l'écart entre \bar{x} et μ ? Représenter la situation sur la courbe tracée en a).
 - i) Calculer l'écart entre la moyenne que Luc a obtenue pour son échantillon et la moyenne de la population. Est-ce que l'échantillon de Luc fait partie des 90 % de cas considérés en c)?
 - Fournir deux exemples de moyennes \bar{x} qui feraient partie des cas considérés en c).
- e) En prélevant un échantillon de 36 étudiants, quelles étaient les chances que la moyenne d'heures de cours des étudiants pigés se situe à au plus 0,5 heure de la moyenne d'heures de cours des étudiants du cégep?

4. En 2013, dans le secteur privé, le salaire moyen d'un technicien en administration sans expérience est de 47 206 \$ avec un écart type de 2 300 \$. On prélève au hasard un échantillon de 100 techniciens dans cette population. Quelles sont les chances que le salaire moyen des 100 techniciens de l'échantillon :

Source: Institut de la statistique du Québec. *Enquête sur la rémunération globale au Québec, Collecte 2013*, janvier 2014.

- se situe à au plus 300 \$ du salaire moyen des techniciens de la population ?
- se situe entre 47 000 \$ et 47 500 \$?

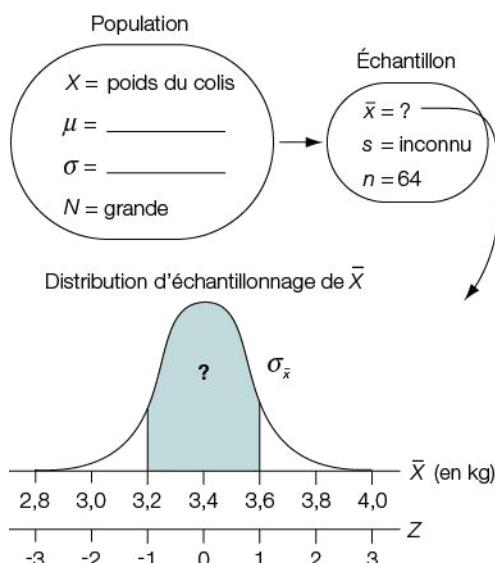
5. Une étude indique que les 1 400 artistes embauchés pour jouer dans des pièces de théâtre au cours de la saison 2009-2010 ont eu un revenu moyen de 8 457 \$ avec un écart type de 2 925 \$. On prélève au hasard 300 acteurs parmi ces 1 400 artistes et l'on s'intéresse à la moyenne de leurs revenus.

Source: Conseil québécois du théâtre. *Profil statistique de la saison théâtrale 2009-2010*, octobre 2012.

Compléter les énoncés.

- La distribution d'échantillonnage de \bar{X} suit une loi normale, car _____. Sa moyenne est de ____ \$ et son écart type est de ____ \$.
- Il y a ____ % de chances que l'écart entre \bar{x} et μ soit d'au plus 250 \$.
- Il y a 80 % de chances que l'écart entre \bar{x} et μ soit d'au plus ____.

6. Un échantillon de 64 colis est prélevé parmi tous les colis reçus à la messagerie Colisbus. Utiliser l'information présentée dans le graphique pour répondre aux questions suivantes :



a) Quelle est la variable étudiée ? Quelle est l'unité de mesure de cette variable ?

b) D'après la représentation graphique de la distribution d'échantillonnage de \bar{X} , quelles sont les chances que le hasard produise une moyenne d'échantillon \bar{x} comprise entre 3,2 kg et 3,6 kg ?

c) Quelle est la valeur de l'écart type de la distribution d'échantillonnage de \bar{X} ?

d) Quel est le poids moyen μ de l'ensemble des colis reçus à la messagerie Colisbus ?

e) Quel est l'écart type σ du poids de l'ensemble des colis reçus à la messagerie Colisbus ?

f) Indiquer, parmi les intervalles suivants de la courbe normale, celui qui a le plus de chances de contenir la moyenne \bar{x} de l'échantillon prélevé :

- [3,6 kg ; 3,8 kg]
- [2,8 kg ; 3,0 kg]
- [3,2 kg ; 3,4 kg]

7. La simulation suivante permet de vérifier que la distribution d'échantillonnage de \bar{X} , pour un échantillon prélevé dans une petite population, a bien une moyenne égale à la moyenne de la population et un écart type égal à σ/\sqrt{n} fois le facteur de correction, comme le stipule le théorème central limite.

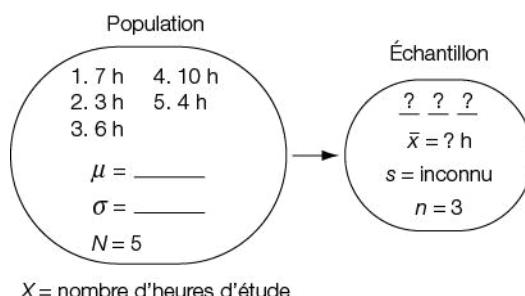
Simulation

Pour une petite population de 5 étudiants, numérotés de 1 à 5, on s'intéresse à la variable aléatoire X : «nombre d'heures d'étude par semaine».

Voici le nombre d'heures d'étude par semaine de chaque étudiant de la population :

1. 7 h
2. 3 h
3. 6 h
4. 10 h
5. 4 h

- a) Calculer la moyenne μ et l'écart type σ du nombre d'heures d'étude pour la population et inscrire les valeurs obtenues dans l'ovale représentant la population.



- b) Pour se faire une idée des valeurs que peut prendre la moyenne du nombre d'heures d'étude d'un échantillon de 3 étudiants pigés au hasard et sans remise dans la population, on prélève tous les échantillons possibles de cette taille et on calcule la moyenne \bar{x} . Le tableau suivant fournit la liste de ces échantillons. Compléter ce tableau.

Échantillons possibles	Nombre d'heures d'étude des étudiants pigés	Moyenne d'heures d'étude dans l'échantillon pigé
{1, 2, 3}	7 h, 3 h, 6 h	$\bar{x} = 5,33 \text{ h}$
{1, 2, 4}	7 h, 3 h, 10 h	$\bar{x} = 6,67 \text{ h}$
{1, 2, 5}	7 h, 3 h, 4 h	$\bar{x} = 4,67 \text{ h}$
{1, 3, 4}	7 h, 6 h, 10 h	$\bar{x} = 7,67 \text{ h}$
{1, 3, 5}	7 h, 6 h, 4 h	$\bar{x} = 5,67 \text{ h}$
{1, 4, 5}	7 h, 10 h, 4 h	$\bar{x} = 7,00 \text{ h}$
{2, 3, 4}	3 h, 6 h, 10 h	$\bar{x} = 6,33 \text{ h}$
{2, 3, 5}	3 h, 6 h, 4 h	$\bar{x} = \underline{\hspace{2cm}}$
{2, 4, 5}	3 h, 10 h, 4 h	$\bar{x} = \underline{\hspace{2cm}}$
{3, 4, 5}	6 h, 10 h, 4 h	$\bar{x} = \underline{\hspace{2cm}}$

- c) Quel nom donne-t-on à la distribution des résultats de la colonne de droite ?
- d) En utilisant le mode statistique de la calculatrice, calculer la moyenne $\mu_{\bar{x}}$ et l'écart type $\sigma_{\bar{x}}$ de la distribution d'échantillonnage de \bar{X} .
- e) Selon le théorème central limite, la moyenne de la distribution d'échantillonnage de \bar{X} est égale à la moyenne de la population ($\mu_{\bar{x}} = \mu$). Cette égalité est-elle vérifiée dans ce cas-ci ?
- f) Selon le théorème central limite, pour une population considérée comme petite par rapport à la taille de l'échantillon ($N < 20n$), on obtient :

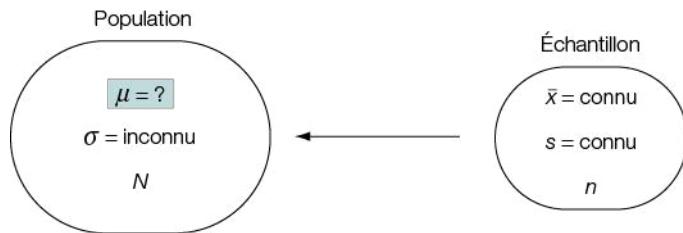
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Cette égalité est-elle vérifiée dans ce cas-ci ?

4.3 L'estimation de la moyenne d'une population

Dans la section précédente, où nous connaissons la moyenne μ et l'écart type σ de la population, nous avons appris à prédire les valeurs que le hasard peut générer comme moyenne \bar{x} pour un échantillon de taille n . Nous allons maintenant examiner la situation inverse : la moyenne μ et l'écart type σ de la population seront inconnus, alors que la moyenne \bar{x} et l'écart type corrigé s de l'échantillon seront connus. Nous tenterons de prédire, avec une certaine probabilité, la moyenne μ de la population à partir de la valeur connue de la moyenne \bar{x} d'un échantillon. Plus précisément, nous chercherons à répondre à la question suivante :

«Comment peut-on estimer la moyenne μ de la population à l'aide de la moyenne \bar{x} d'un échantillon?»

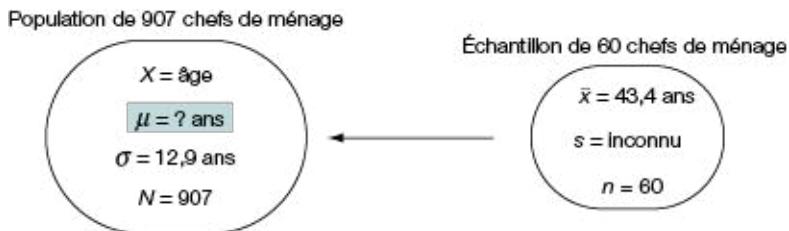


4.3.1 L'échantillon de grande taille ($n \geq 30$)

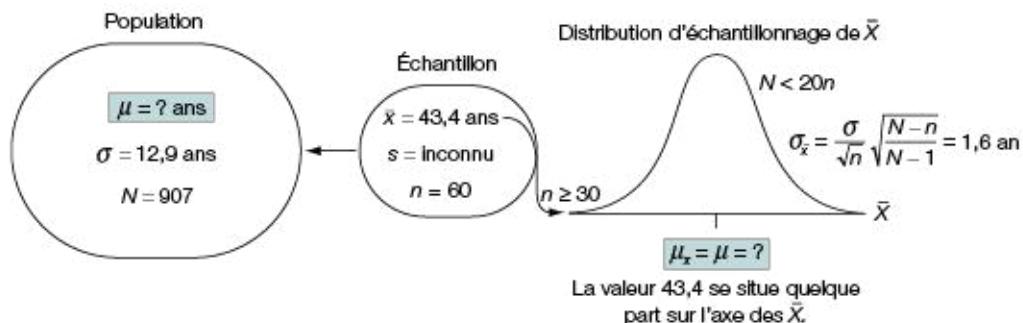
Estimation de la moyenne μ par intervalle de confiance

La mise en situation suivante sert à illustrer la marche à suivre pour estimer la moyenne μ d'une population à l'aide de la moyenne \bar{x} d'un échantillon dont la taille est d'au moins 30 unités. Le cas des échantillons de petite taille, inférieure à 30 unités, est abordé dans la prochaine section.

Au début de la section 4.2, notre objectif était d'estimer la moyenne d'âge μ d'une population composée de 907 chefs de ménage à l'aide de la moyenne d'âge de 43,4 ans obtenue pour un échantillon de 60 personnes pigées dans la population. Nous supposons, pour le moment, que l'écart type σ de l'âge de la population est connu. Le graphique qui suit représente la situation.



- Comme $n \geq 30$, nous savons que la distribution de toutes les valeurs possibles de \bar{X} suit une loi normale dont la moyenne $\mu_{\bar{x}}$ est égale à μ , soit la moyenne de la population, que nous supposons ici inconnue, et dont l'écart type $\sigma_{\bar{x}}$ est de 1,6 an. Par conséquent, la moyenne \bar{x} de 43,4 ans est une des valeurs de cette distribution.



- Comme μ , le centre de la cloche, est ici inconnu, il est impossible de savoir où exactement, sur l'axe \bar{X} , se situe la moyenne \bar{x} de 43,4 ans. On sait toutefois qu'il est peu probable que cette moyenne se trouve exactement au centre de la courbe normale, c'est-à-dire que $\mu = \bar{x}$. Par contre, il y a de fortes chances que \bar{x} se situe assez près du centre de la courbe.
- Nous savons que la zone bleue du graphique représenté ci-dessous contient 95 % de tous les échantillons possibles. L'échantillon prélevé a donc 95 % de chances d'être compris dans cette zone et, si c'est le cas, l'écart entre \bar{x} et μ est d'au plus E ans.

Rappelons comment déterminer la valeur E :

$$E = z\sigma_{\bar{x}} = 1,96 \times 1,6 = 3,1 \text{ ans}$$

- Étant donné qu'il y a 95 % de chances que l'écart entre \bar{x} et μ soit d'au plus 3,1 ans, prétendre que la moyenne μ de la population est égale à \bar{x} conduirait à commettre une erreur d'estimation pour μ pouvant atteindre 3,1 ans. En tenant compte de ce fait, on dira que la moyenne \bar{x} de 43,4 ans permet d'estimer μ à 3,1 ans près, autrement dit :

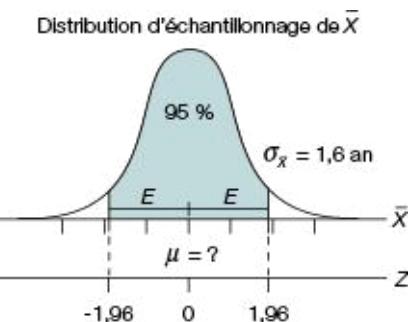
$$\mu = 43,4 \text{ ans} \pm 3,1 \text{ ans} \quad \Rightarrow \quad \mu = \bar{x} \pm E$$

La valeur de μ serait dans ce cas comprise dans l'intervalle suivant :

$$\mu \in [40,3 \text{ ans} ; 46,5 \text{ ans}] \quad \Rightarrow \quad \mu \in [\bar{x} - E ; \bar{x} + E]$$

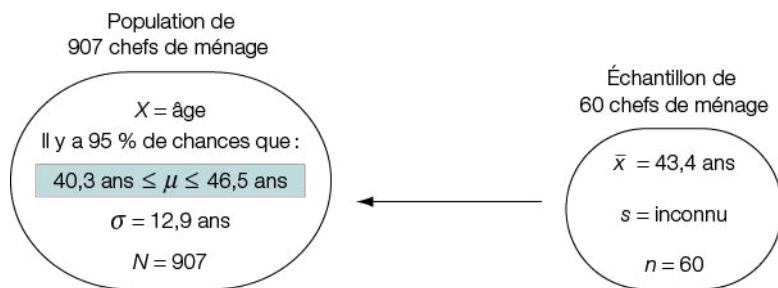
On utilise aussi la notation suivante :

$$40,3 \text{ ans} \leq \mu \leq 46,5 \text{ ans} \quad \Rightarrow \quad \bar{x} - E \leq \mu \leq \bar{x} + E$$



Interprétation de l'intervalle construit

Il y a 95 % de chances que l'intervalle [40,3 ans ; 46,5 ans] contienne la moyenne d'âge μ des 907 chefs de ménage de la population : l'âge moyen de la population se situerait dans ce cas quelque part entre 40,3 et 46,5 ans ; ce qui donnerait un écart d'au plus 3,1 ans entre la moyenne de l'échantillon et la moyenne de la population.



On donne le nom d'**intervalle de confiance** à l'intervalle [40,3 ans ; 46,5 ans]. Les valeurs 40,3 ans et 46,5 ans sont les **bornes de l'intervalle**.

Puisqu'il y a 95 % de chances que cet intervalle contienne μ , on dit que le **niveau de confiance** de l'intervalle est de 95 %.

L'écart d'au plus 3,1 ans entre \bar{x} et μ est appelé **marge d'erreur** ou **précision de l'estimation**. On désigne la marge d'erreur par la lettre E .

Nous venons de voir qu'il est impossible de déterminer la valeur exacte de la moyenne μ d'une population à l'aide de la moyenne \bar{x} d'un échantillon. En revanche, on peut construire un intervalle de rayon E , centré en \bar{x} , qui a de bonnes chances de contenir la moyenne μ de la population. C'est ce qu'on appelle «effectuer une **estimation de μ par intervalle de confiance**».

Risque d'erreur

S'il y a 95 % de chances que l'intervalle [40,3 ans ; 46,5 ans] contienne la moyenne μ de la population, il y a donc 5 % de risques que cet intervalle ne contienne pas μ et, par conséquent, que l'écart entre \bar{x} et μ soit supérieur à 3,1. On dit alors que le risque d'erreur de l'estimation est de 5 %. On emploie la notation α (lettre «*a*» de l'alphabet grec, qui se lit alpha) pour représenter un risque d'erreur.

Voici la démarche générale pour construire un intervalle de confiance pour μ :

Estimation de la moyenne d'une population par intervalle de confiance

Démarche pour construire un intervalle de confiance pour μ :

- Vérifier les conditions d'application.
- Déterminer l'écart type $\sigma_{\bar{x}}$ de la distribution d'échantillonnage de \bar{X} .
- Calculer la marge d'erreur associée au niveau de confiance considéré : $E = z \sigma_{\bar{x}}$.
- Calculer les bornes de l'intervalle de confiance et l'interpréter : $[\bar{x} - E; \bar{x} + E]$.

EXEMPLE 1

Trois étudiants, Nicole, Hélène et Claude, prélèvent chacun un échantillon de taille 60 dans la population des 907 chefs de ménage afin d'estimer la moyenne d'âge μ de cette population avec un niveau de confiance de 95 %.

a) Compléter le tableau suivant.

Nom	Moyenne \bar{x}	Marge d'erreur $E = z\sigma_x$	Intervalle de confiance $[\bar{x} - E; \bar{x} + E]$
Nicole	$\bar{x} = 43,4$ ans	$E = 3,1$ ans	[40,3 ans ; 46,5 ans]
Hélène	$\bar{x} = 46,2$ ans	$E =$	
Claude	$\bar{x} = 41,0$ ans	$E =$	

Interprétation

Chacun des trois étudiants estime qu'il y a 95 % de chances que la moyenne d'âge μ de la population se situe à au plus 3,1 ans de la moyenne d'âge des 60 personnes de son échantillon. Comme le hasard leur a donné des échantillons différents, ils ont donc obtenu des intervalles de confiance différents, chacun affirmant qu'il y a 95 % de chances que l'intervalle de confiance qu'il a construit contienne μ .

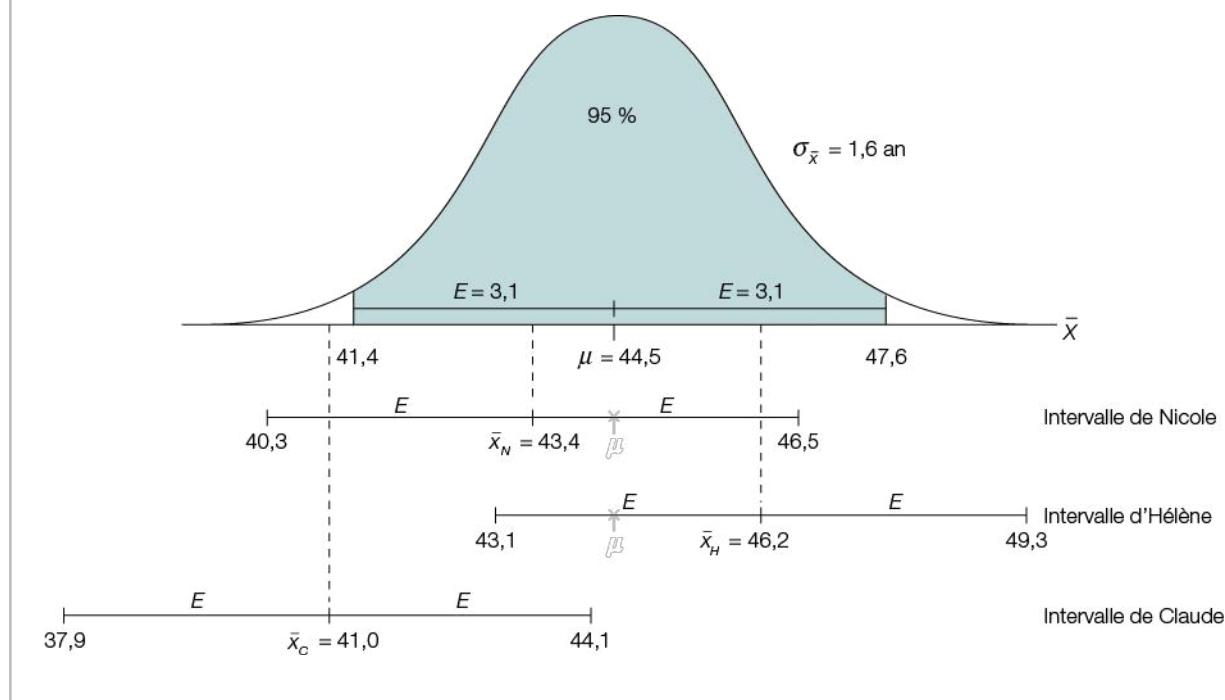
b) Exceptionnellement, dans ce cas-ci, on connaît l'âge moyen μ de la population (*voir la page 193*). Sachant que $\mu = 44,5$ ans, y a-t-il, parmi les intervalles construits, un intervalle qui ne contient pas la moyenne μ de la population ?

Si oui, qui a construit cet intervalle ? _____

Quels étaient les risques que cela se produise ? _____

c) Le fait que la moyenne μ soit connue permet de représenter graphiquement les intervalles de confiance sur la distribution d'échantillonnage de \bar{X} . Pour chacun des intervalles de confiance construits, marquer d'une croix la position de la moyenne μ de la population.

Représentation graphique d'un intervalle de confiance



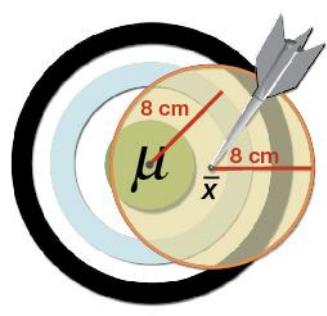
NOTE

Il importe de saisir que, dans le graphique de la page précédente, le centre de l'intervalle de confiance est la moyenne \bar{x} de l'échantillon, alors que le centre de la courbe normale de la distribution d'échantillonnage de \bar{X} est la moyenne μ de la population.

Analogie avec le jeu de fléchettes

Utilisons un jeu de fléchettes pour mieux comprendre le concept d'intervalle de confiance. Pour ce faire, associons le trou fait par la fléchette sur la cible à la moyenne \bar{x} d'un échantillon et le centre de la cible à la moyenne μ de la population.

- L'examen d'une cible sur laquelle une fléchette a été lancée un très grand nombre de fois montrera que les trous \bar{x} laissés par la fléchette sont très nombreux près du centre μ de la cible, et de moins en moins nombreux à mesure que l'on s'éloigne du centre; cette situation est analogue à celle qu'on observe pour la distribution des moyennes d'échantillons \bar{x} autour de la moyenne μ de la population comme l'indique la courbe normale de la page précédente.
- Supposons que l'on observe que 95 % des trous laissés par une fléchette se situent à au plus 8 cm du centre de la cible (8 cm est la marge d'erreur associée à ces lancers). Graphiquement, la surface du cercle de 8 cm de rayon, centré en μ , contient 95 % de tous les trous \bar{x} laissés par la fléchette.
- Imaginons maintenant que l'on bande les yeux d'une personne et qu'on lui demande de choisir un trou \bar{x} au hasard sur la cible. De l'énoncé du paragraphe précédent, on peut déduire que si l'on trace un cercle de 8 cm de rayon centré sur le trou \bar{x} choisi, la surface du cercle a 95 % de chances de contenir le centre μ de la cible. À titre d'exemple, si le trou \bar{x} illustré sur la cible est choisi, la surface du cercle jaune centré en \bar{x} contient le centre μ de la cible, car la distance entre \bar{x} et μ est inférieure à 8 cm. Par contre, si le trou \bar{x} choisi au hasard se situe dans la zone noire ou dans la zone blanche adjacente, la surface du cercle de 8 cm de rayon centré en \bar{x} ne contiendra pas le centre μ de la cible, car la distance entre \bar{x} et μ est supérieure à 8 cm; les risques que cela se produise sont de 5 %. L'idée d'intervalle de confiance est appliquée ici à une surface, la surface d'un cercle de 8 cm de rayon centré en \bar{x} .



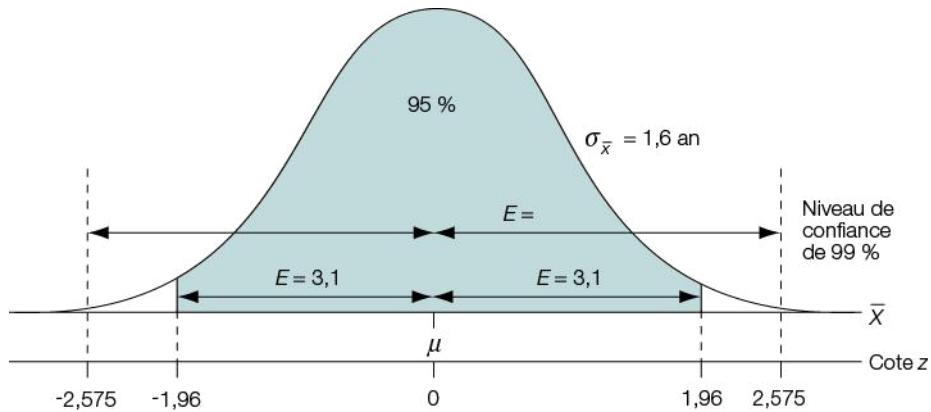
EXEMPLE 2

Que se passera-t-il si le niveau de confiance d'une estimation passe de 95 % à 99 %?

Représentation graphique de la variation d'un niveau de confiance

On peut facilement voir sur la distribution d'échantillonnage de \bar{X} représentée à la page suivante que l'augmentation du niveau de confiance de 95 % à 99 % a pour effet d'élever la marge d'erreur de l'estimation. Par conséquent, l'intervalle de confiance sera plus grand.

Indiquer sur le graphique la valeur de la marge d'erreur E pour un niveau de confiance de 99 %.



Effet de la variation du niveau de confiance sur la marge d'erreur

L'augmentation du niveau de confiance d'une estimation a l'avantage d'accroître les chances de voir l'intervalle de confiance contenir la moyenne μ de la population, mais a l'inconvénient d'augmenter la marge d'erreur de cette estimation, ce qui entraîne une estimation moins précise de μ .

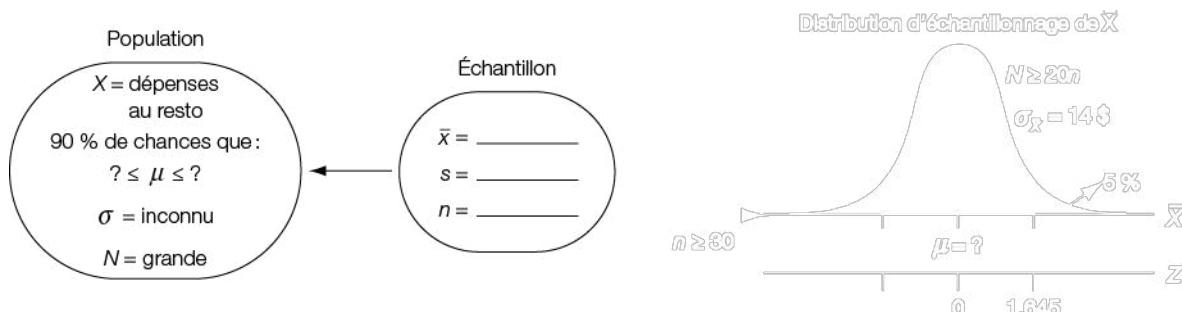
EXEMPLE 3

Dans le cadre de l'*Enquête sur les dépenses des ménages 2011*, Statistique Canada a établi que les 1 574 ménages québécois de l'échantillon dépensaient en moyenne 1 807 \$ par année au restaurant avec un écart type corrigé de 556 \$. Construire un intervalle de confiance au niveau de confiance de 90 % permettant d'estimer le montant annuel moyen des dépenses au restaurant pour l'ensemble des ménages du Québec.

Sources: Statistique Canada. *Tableau 203-0021, CANSIM*.

Statistique Canada. *Guide de l'utilisateur, Enquête sur les dépenses des ménages 2011*, février 2013.

Solution



Interprétation de l'intervalle de confiance

Il y a _____ de chances que le montant moyen des dépenses au restaurant des ménages québécois se situe entre _____ et _____ par année. Il y a donc un risque de 10 % que ce montant moyen ne se situe pas entre ces deux valeurs.

Estimation de l'écart type σ d'une population lorsque celui-ci est inconnu

Lorsque l'écart type σ de la population est inconnu, on se sert de l'écart type corrigé s de l'échantillon comme estimateur de σ pour calculer $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

Il ne faut pas oublier de multiplier par le facteur de correction $(\sqrt{(N-n)/(N-1)})$ si $N < 20n$.

Estimation ponctuelle de la moyenne μ

Dans une étude statistique, on fait souvent une estimation ponctuelle de μ , surtout si l'on présente un court résumé des résultats obtenus. Une telle estimation consiste à prétendre que la moyenne μ de la population est égale à la moyenne \bar{x} de l'échantillon. Par exemple, les nouvelles télévisées diffuseraient ainsi la conclusion de l'étude réalisée à l'exemple précédent :

« Selon une enquête de Statistique Canada, les ménages québécois ont dépensé en moyenne 1 807 \$ au restaurant en 2011. »

Le défaut de ce type d'estimation est que l'égalité est rarement vraie puisqu'il y a généralement un écart entre \bar{x} et μ . Toutefois, si la marge d'erreur entre \bar{x} et μ est petite par rapport à la valeur de la moyenne (comme c'est le cas ici avec moins de 23 \$), l'estimation est acceptable.

Pour que le public puisse juger de la qualité d'un sondage, le document *Droits et responsabilités de la presse* du Conseil de presse du Québec mentionne que les médias doivent donner les éléments méthodologiques d'un sondage lors de la publication.

Le résultat du sondage de l'exemple précédent serait publié de la façon suivante dans un journal (à remarquer que l'on utilise l'expression 18 fois sur 20 pour indiquer le niveau de confiance de 90 %) :

Les ménages québécois ont dépensé en moyenne 1 807 \$ au restaurant en 2011. [...]

Méthodologie du sondage

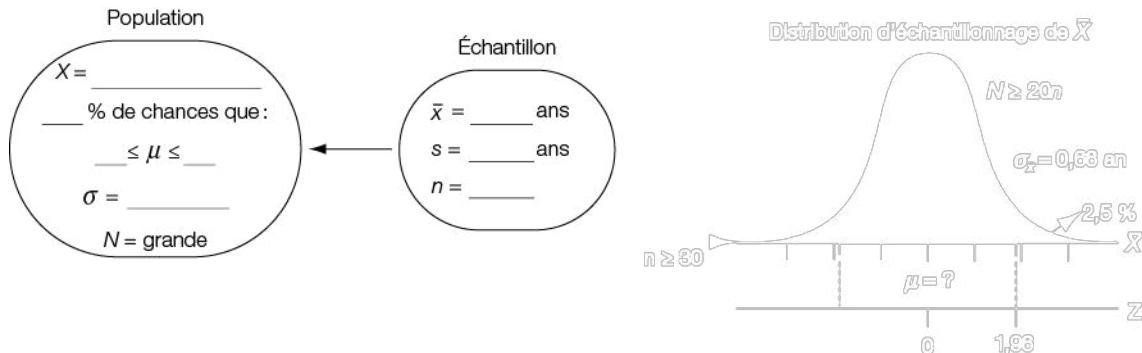
Ce sondage a été réalisé par Statistique Canada auprès d'un échantillon aléatoire de 1 574 ménages québécois. Avec un échantillon de cette taille, la marge d'erreur de l'estimation est de 23 \$, 18 fois sur 20.

EXERCICE DE COMPRÉHENSION | 4.2

Un sondage est effectué auprès d'un échantillon aléatoire de 100 Québécoises qui se sont mariées pour la première fois en 2012. La moyenne d'âge de celles-ci est de 31,3 ans avec un écart type corrigé de 6,6 ans.

Source: Institut de la statistique du Québec. *Le bilan démographique du Québec, édition 2013*, décembre 2013.

- a) Construire un intervalle de confiance, au niveau de confiance de 95 %, pour estimer l'âge moyen de l'ensemble des Québécoises qui se sont mariées pour la première fois en 2012.



Interprétation de l'intervalle de confiance

Il y a _____ % de chances que l'âge moyen de l'ensemble des Québécoises qui se sont mariées pour la première fois en 2012 soit compris entre _____ ans et _____ ans.

- b) Un article paru dans le journal résumait ce sondage ainsi :

Un sondage indique que la moyenne d'âge des Québécoises qui se sont mariées pour la première fois en 2012 est de 31,3 ans.

Méthodologie

Ce sondage a été mené auprès d'un échantillon de _____ Québécoises qui en étaient à leur premier mariage en 2012. Avec un échantillon de cette taille, la marge d'erreur de l'estimation est de _____ an, _____ fois sur 20.

- i) Compléter la méthodologie du sondage.

ii) L'estimation donnée dans l'article est-elle ponctuelle ou par intervalle ? _____

Selon vous, ce type d'estimation est-il acceptable dans ce cas-ci ? _____

Sur quoi vous basez-vous pour porter ce jugement ?

► c) Vrai ou faux ?

- i) Il y a 95 % de chances que l'écart entre \bar{x} et μ soit de 1,3 an. _____
- ii) Si l'on augmente le niveau de confiance à 99 %, la marge d'erreur sera plus petite. _____

4.3.2 L'échantillon de petite taille ($n < 30$)

Jusqu'à présent, nous avons utilisé la cote z dans le calcul de la marge d'erreur de l'estimation de μ , car l'une ou l'autre des conditions permettant d'affirmer que la distribution d'échantillonnage de \bar{X} suit une loi normale était respectée, soit :

- un échantillon de grande taille ($n \geq 30$) avec l'écart type σ de la population connu ou estimé par s ;
- un échantillon de petite taille ($n < 30$) tiré d'une population normale dont l'écart type σ est connu.

Dans cette section, nous apprendrons à calculer la marge d'erreur pour un échantillon de petite taille ($n < 30$) tiré d'une population normale où l'écart type σ de la population est inconnu.

Dans ce cas, la valeur $(\bar{x} - \mu_{\bar{x}})/\sigma_{\bar{x}}$ ne suit pas une loi normale centrée réduite, mais une loi de Student. Les caractéristiques de cette loi sont énoncées ci-dessous.

La loi de Student

La courbe de la distribution de Student ressemble à celle de la loi normale ; elle a la forme d'une cloche symétrique par rapport à la moyenne, mais elle est généralement un peu plus aplatie que la courbe normale. On désigne la variable d'une loi de Student par la lettre T et ses valeurs par t . Il existe plusieurs distributions de Student ; c'est la taille de l'échantillon qui indique laquelle choisir. Il est à souligner que, plus la taille de l'échantillon augmente, plus la distribution de Student s'approche de la loi normale $N(0; 1)$.

La table de Student

Pour trouver une valeur t_α dans la table de Student (voir la page 348), il faut connaître :

- l'aire α sous la courbe pour $T > t_\alpha$: $P(T > t_\alpha) = \alpha$;
- le nombre de **degrés de liberté** que l'on détermine ainsi : $dl = n - 1$.

L'idée de degrés de liberté peut s'expliquer ainsi :

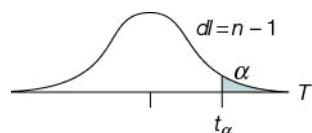
Supposons que la moyenne d'un échantillon de 3 données est 20. En tenant compte de cette contrainte, on peut facilement déterminer une des 3 données pourvu que l'on connaisse la valeur des 2 autres. Par exemple, si l'échantillon comprend les données 8 et 22, on trouve la 3^e donnée ainsi :

$$\frac{8 + 22 + x_3}{3} = 20$$

$$8 + 22 + x_3 = 60$$

$$x_3 = 30$$

On considère alors que, dans un échantillon de taille 3, il y a 1 donnée liée et 2 données libres, soit 2 degrés de liberté.



La marge d'erreur avec $n < 30$, population normale et σ inconnu

Sous ces conditions, la valeur $(\bar{x} - \mu_{\bar{x}})/\sigma_{\bar{x}}$ suit une loi de Student et la marge d'erreur de l'estimation de la moyenne μ se calcule ainsi :

$$E = t\sigma_{\bar{x}} \quad \text{où} \quad t = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \text{ suit une loi de Student avec } n - 1 \text{ degrés de liberté ;}$$

$$\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}, \text{ car } \sigma \text{ est inconnu.} \quad (\text{Si } N < 20n, \text{ on multiplie par le facteur de correction.})$$

NOTE

Le cas où $n < 30$ et où la population ne suit pas une loi normale ne sera pas traité dans ce cours.



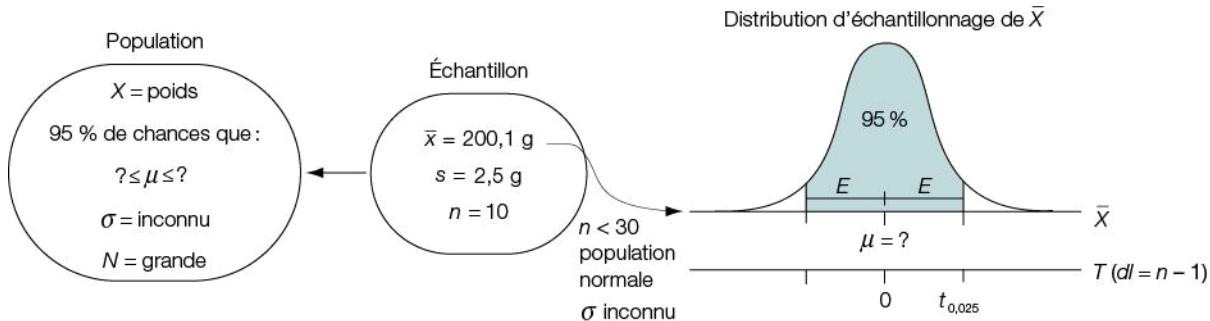
William Gosset (1876-1937)

Diplômé de chimie et de mathématiques, William Sealy Gosset est embauché comme chimiste par la brasserie Guinness en 1899 pour uniformiser le goût de la bière. Il étudie alors la statistique et met au point une méthode de contrôle de la qualité utilisant une distribution basée sur de petits échantillons. On lui attribue la loi de Student, nommée d'après le pseudonyme sous lequel il a publié sa découverte en 1908.

EXEMPLE

Dans une usine, une machine est réglée de telle sorte que le poids du produit qu'elle verse dans un contenant est distribué selon une loi normale. Un échantillon aléatoire de 10 contenants prélevé dans la production d'une journée donne un poids moyen $\bar{x} = 200,1$ g et un écart type corrigé $s = 2,5$ g. Estimer, à partir de cet échantillon, le poids moyen par contenant pour l'ensemble de la production de la journée. Le niveau de confiance est fixé à 95 %.

Solution



- Avec $n < 30$, population normale et σ inconnu, on a :

$$E = t\sigma_{\bar{x}}, \text{ où } t \text{ suit une loi de Student avec } dl = \underline{\hspace{2cm}}$$

- Valeur de $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} =$$

- Marge d'erreur :

$$E = t\sigma_{\bar{x}} = 2,262 \times 0,8 = 1,8 \text{ g}$$

- Intervalle de confiance :

$$\mu = \bar{x} \pm E = 200,1 \pm 1,8$$

$$\mu \in [198,3 \text{ g}; 201,9 \text{ g}]$$

Il y a 95 % de chances que le poids moyen par contenant de la production se situe entre 198,3 g et 201,9 g.

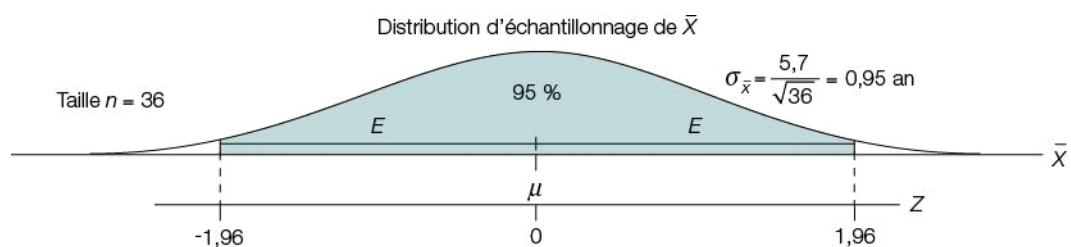
4.3.3 Le choix de la taille de l'échantillon

Quel est l'effet de la variation de la taille de l'échantillon sur la marge d'erreur ?

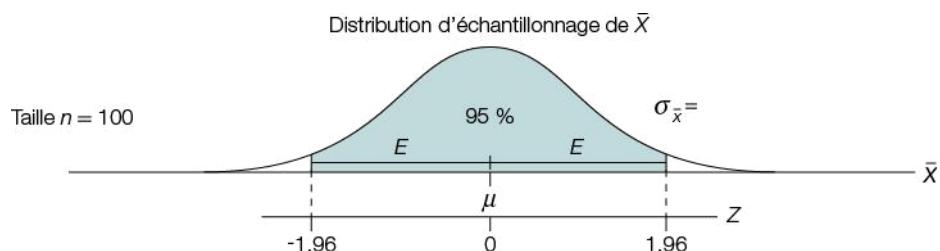
EXEMPLE

On prélève trois échantillons de tailles différentes (36, 100 et 500) parmi tous les étudiants d'une université afin d'estimer l'âge moyen des étudiants avec un niveau de confiance de 95 %. Des études antérieures ont démontré que l'écart type σ de l'âge des étudiants de l'université pouvait être estimé à 5,7 ans. Voici les courbes normales de la distribution d'échantillonnage de \bar{X} pour chaque taille d'échantillon. Calculer la marge d'erreur de l'estimation de μ pour chaque cas et comparer les résultats.

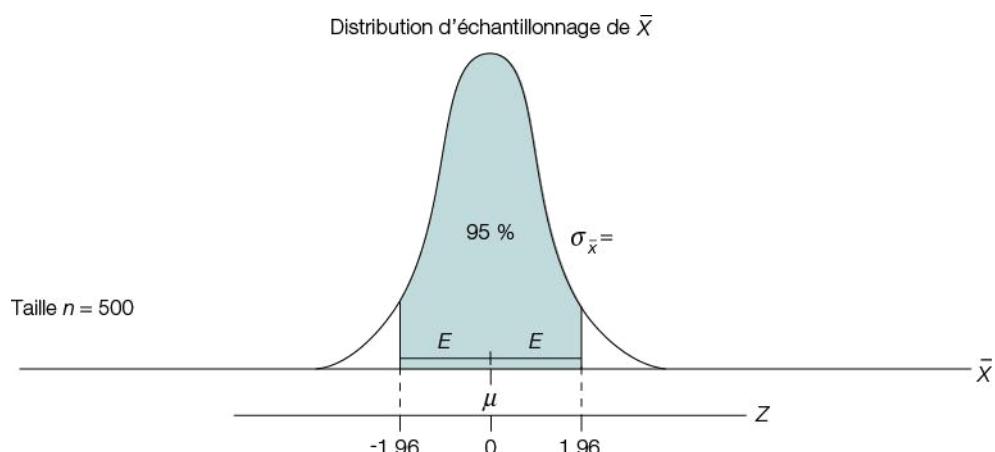
- Pour $n = 36$, on a $E =$ _____



- Pour $n = 100$, on a $E =$ _____



- Pour $n = 500$, on a $E =$ _____



Effet de la variation de la taille de l'échantillon sur la marge d'erreur

Pour un même niveau de confiance, plus on augmente la taille de l'échantillon, plus l'écart type $\sigma_{\bar{x}}$ diminue, ce qui a pour conséquence de diminuer la marge d'erreur et, par le fait même, de donner une estimation plus précise de la moyenne de la population.

On peut fixer d'avance la marge d'erreur que l'on ne veut pas excéder et choisir la taille de l'échantillon en conséquence. Comme nous venons de le voir, plus l'échantillon est grand, plus la marge d'erreur est petite, mais plus les coûts du sondage sont élevés.

EXEMPLE

Quelle taille minimale d'échantillon faudrait-il prendre pour estimer la moyenne d'âge des étudiants d'une université avec une marge d'erreur d'au plus 1,5 an et un niveau de confiance de 95 %, si des études antérieures ont donné un écart type σ de 5,7 ans pour la population ?

Solution

On cherche n tel que la marge d'erreur soit $E = 1,5$ an pour un niveau de confiance à 95 %.

Écrivons la formule donnant la marge d'erreur, puis insérons-y les valeurs connues :

$$E = z\sigma_{\bar{x}} \Leftrightarrow E = z \frac{\sigma}{\sqrt{n}}$$

$$1,5 = 1,96 \times \frac{5,7}{\sqrt{n}}$$

En isolant n , on obtient : $\sqrt{n} = \frac{1,96 \times 5,7}{1,5}$

En éllevant au carré, on a : $n = \left(\frac{1,96 \times 5,7}{1,5} \right)^2 = 55,5$

Il faudrait prélever un échantillon de 56 étudiants, au minimum, pour obtenir une marge d'erreur d'au plus 1,5 an dans l'estimation de l'âge de tous les étudiants de l'université.

NOTE

Quand la valeur de σ est inconnue, on fait une enquête préliminaire avec un échantillon d'au moins 30 unités, et on utilise l'écart type corrigé s de cet échantillon comme estimateur de σ .

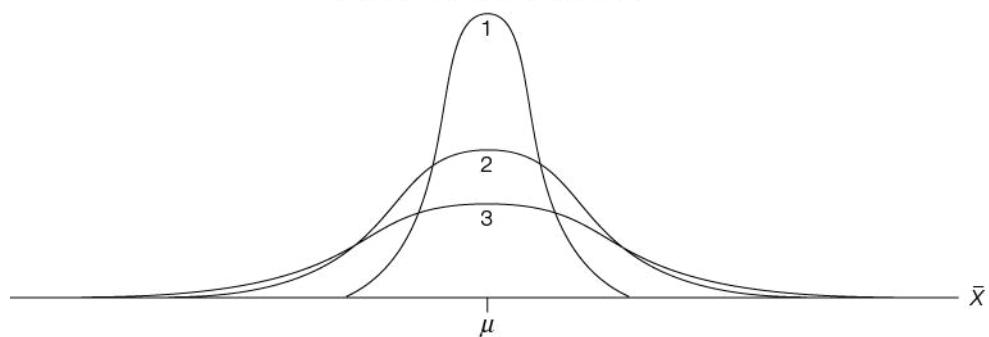
EXERCICE DE COMPRÉHENSION | 4.3

Une machine, précise à 80 ml près, remplit des contenants de jus selon une distribution normale. On désire estimer le volume moyen de jus contenu dans les 1 600 contenants remplis durant la dernière heure.

- a) Quelle taille doit avoir l'échantillon prélevé si l'on veut effectuer une estimation de μ , au niveau de confiance de 99 %, avec une marge d'erreur d'au plus 30 ml en prenant la précision de la machine comme écart type σ de la population ?

- b) On utilise 3 échantillons de taille différente pour estimer le volume moyen de jus dans la production de la dernière heure : un échantillon de 25 contenants, un de 75 contenants et un de 150 contenants.
- Avec quelle taille d'échantillon l'estimation sera-t-elle la plus précise ? _____
 - Les trois courbes qui suivent représentent la distribution des valeurs possibles de \bar{X} pour chacun de ces échantillons. Indiquer la courbe normale qui représente la distribution d'échantillonnage de \bar{X} :
 - pour l'échantillon de taille 150, c'est la courbe _____ ;
 - pour l'échantillon de taille 75, c'est la courbe _____ .

Distribution d'échantillonnage de \bar{X}



Résumé de la section 4.3

Marche à suivre pour construire un intervalle de confiance pour μ

- Vérifier les conditions d'application : $n \geq 30$ ou population normale
- Déterminer l'écart type $\sigma_{\bar{x}}$ de la distribution d'échantillonnage de \bar{X} :

- σ connu : $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- σ inconnu : $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$

NOTE

Quand la population est petite ($N < 20n$), on multiplie les rapports ci-dessus par le facteur de correction $\sqrt{(N - n)/(N - 1)}$.

► 3. Calculer la marge d'erreur associée au niveau de confiance:

Taille de l'échantillon	Distribution de la population	Écart type σ de la population	Marge d'erreur
$n \geq 30$	Quelconque	Connu Inconnu	$E = z \sigma_x$
$n < 30$	Normale	Connu	$E = z \sigma_x$
$n < 30$	Normale	Inconnu	$E = t \sigma_x$ t suit une Student avec $dl = n - 1$

4. Calculer les bornes de l'intervalle et interpréter cet intervalle en tenant compte du contexte:

$$\mu \in [\bar{x} - E; \bar{x} + E]$$

Estimation ponctuelle de la moyenne de la population

$$\mu = \bar{x}$$

Prédétermination de la taille d'un échantillon associée à une marge d'erreur donnée

On isole n dans l'équation de la marge d'erreur après avoir remplacé E , z et σ (ou son estimation) par leurs valeurs:

$$E = z \frac{\sigma}{\sqrt{n}}$$

EXERCICES 4.3

1. a) On sait que le poids des contenants remplis par une machine obéit à une loi normale dont l'écart type σ est de 0,7 g. Pour un échantillon aléatoire de 100 contenants prélevés dans la production de cette machine, on obtient un poids moyen de 49,7 g. Construire et interpréter l'intervalle de confiance, au niveau de confiance de 95 %, permettant d'estimer le poids moyen de tous les contenants remplis par la machine.
 b) Donner et interpréter le risque d'erreur.
 c) Donner et interpréter la marge d'erreur.
2. On construit un intervalle de confiance pour estimer μ à l'aide d'un échantillon de taille 50.
 - a) Donner la cote z des niveaux de confiance suivants:
 - i) 80 %
 - ii) 93 %
 - iii) 97 %

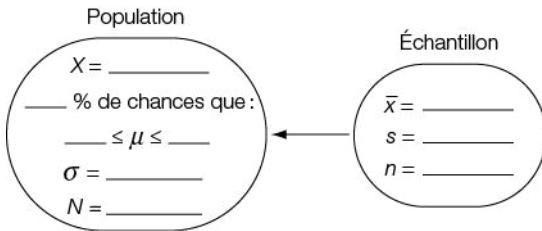
- b) Lequel de ces trois niveaux de confiance donnera:

- i) la plus petite marge d'erreur?
- ii) le plus grand risque d'erreur?

3. a) Afin de pallier un problème de surcharge du réseau dû aux appels interurbains le jour de la fête des Mères, Québec-tel désire estimer la durée moyenne de ces appels. Un échantillon aléatoire de 36 appels interurbains a donné les temps ci-dessous (en minutes). Construire un intervalle de confiance, au niveau de confiance de 95 %, pour estimer la durée moyenne des appels interurbains ce jour-là.

3,5	3,5	5,2	8,6
2,3	6,3	3,4	3,1
4,2	6,0	10,2	6,7
2,2	2,1	4,9	2,6
12,4	3,1	3,8	7,2
4,5	7,4	4,9	2,9
4,0	3,8	9,3	4,6
3,3	20,7	5,3	3,3
5,2	4,3	4,7	2,8

- b) Le graphique suivant présente les résultats obtenus en a). Le compléter.



- c) L'article suivant présente le sondage. Le compléter.

Selon une étude effectuée par sondage par Québec-tel, la durée moyenne des appels interurbains le jour de la fête des Mères est de _____ minutes. [...]

Méthodologie

Ce sondage a été mené à partir d'un échantillon de _____ appels interurbains effectués le jour de la fête des Mères. Avec un échantillon de cette taille, la marge d'erreur de l'estimation est de _____ minute, _____ fois sur 20.

4. On désire estimer par intervalle de confiance, au niveau de confiance de 95 %, le volume moyen de bière par bouteille d'une production de plusieurs milliers de bouteilles. La machine qui remplit celles-ci est précise à 3 ml près. On prélève au hasard 125 bouteilles et on obtient un volume moyen de 337 ml par bouteille. Construire l'intervalle de confiance demandé en utilisant la précision de la machine comme écart type σ de la production.

5. Vrai ou faux ?

- a) On peut interpréter le niveau de confiance de 95 % pour l'intervalle construit au numéro 4 ainsi :
- i) Il y a 95 % de chances que le volume moyen de bière par bouteille de l'échantillon se situe entre 336,5 ml et 337,5 ml.
 - ii) Il y a 95 % de chances d'être dans l'intervalle de confiance construit.
 - iii) Il y a 95 % de chances que le volume moyen de bière par bouteille de la production se situe entre 336,5 ml et 337,5 ml.
 - iv) Il y a 95 % de chances que le volume moyen de bière par bouteille de la production soit compris dans l'intervalle construit.
- b) On peut interpréter le risque d'erreur de 5 % pour l'intervalle construit au numéro 4 ainsi :
- i) Il y a 5 % de risques qu'on fasse des erreurs en calculant l'intervalle de confiance.

- ii) Il y a 5 % de risques que le volume moyen de bière par bouteille de l'échantillon ne se situe pas dans l'intervalle construit.

- iii) Il y a 5 % de risques que le volume moyen de bière par bouteille de la production ne se situe pas dans l'intervalle construit.

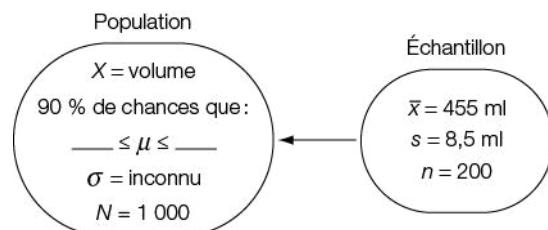
- iv) Il y a 5 % de risques que le volume moyen de bière par bouteille soit inférieur à 336,5 ml ou supérieur à 337,5 ml.

6. Soit une population normale d'écart type inconnu. On désire construire un intervalle de confiance pour μ en utilisant la moyenne \bar{x} d'un échantillon de petite taille ($n < 30$).

- a) Énoncer la formule de la marge d'erreur.
 b) Déterminer la valeur de t dans le cas où :
- i) le niveau de confiance est 99 % et $n = 15$.
 - ii) le niveau de confiance est 95 % et $n = 28$.
- c) Laquelle des deux situations décrites en b) donne la plus petite marge d'erreur ?

7. Un psychologue fait exécuter une tâche à 22 enfants de 10 ans choisis au hasard et obtient un temps moyen d'exécution de 48 minutes avec un écart type corrigé de 6 minutes. Construire un intervalle de confiance, au niveau de confiance de 99 %, pour estimer le temps moyen que l'ensemble des enfants de cet âge prendrait pour exécuter cette tâche. On suppose que la distribution du temps obéit à une loi normale.

8. a) Utiliser l'information contenue dans le graphique suivant pour estimer μ par intervalle de confiance.



- b) Compléter l'énoncé : Il y a _____ de chances que l'écart entre la moyenne de l'échantillon et la moyenne de la population soit supérieur à _____.

- c) Quel type d'estimation utilise-t-on si l'on pose :

- i) $\mu = 455 \text{ ml}$?
- ii) $454,1 \text{ ml} \leq \mu \leq 455,9 \text{ ml}$?

- d) En négligeant les moyennes \bar{x} ayant moins de 0,3 % de chances d'être obtenues, quelle est la plus grande marge d'erreur possible entre μ et \bar{x} ? Pourquoi n'utilise-t-on pas cette marge d'erreur pour estimer μ ?

9. Une équipe de chercheurs suit le développement de jeunes enfants depuis leur naissance afin d'établir une courbe de croissance indiquant la distribution de leur taille et de leur poids selon l'âge. Voici le tableau de distribution du poids des 500 filles de l'échantillon, à l'âge de trois ans.

Répartition de 500 filles de 3 ans selon le poids

Poids (en kg)	Nombre de filles
[11; 12[45
[12; 13[80
[13; 14[140
[14; 15[125
[15; 16[70
[16; 17[40
Total	500

- a) Estimer, par un intervalle de confiance au niveau de confiance de 95 %, le poids moyen des filles de 3 ans.
- b) Dans ce cas-ci, l'estimation ponctuelle serait-elle acceptable? Justifier la réponse.
10. a) Si l'on augmente le niveau de confiance de 90 % à 99 %, la marge d'erreur dans l'estimation de μ sera-t-elle plus grande ou plus petite?
- b) Si l'on augmente la taille d'un échantillon, tout en gardant le même niveau de confiance, la marge d'erreur dans l'estimation de μ sera-t-elle plus grande ou plus petite?

11.

Les jeunes s'informent de plus en plus sur Internet

Une étude indique qu'en 2013 les Québécois de moins de 35 ans ont passé en moyenne 34 minutes par jour sur Internet à s'informer de l'actualité; c'est 21 minutes de plus qu'il y a 4 ans. De fait, Internet compte pour 47 % du temps quotidien qu'ils consacrent à s'informer. [...]

Méthodologie

Cette étude a été effectuée par le Centre d'études sur les médias auprès d'un échantillon aléatoire de 150 Québécois de moins de 35 ans. Avec un échantillon de cette taille, la marge d'erreur est de 1,3 minute, 19 fois sur 20.

Source: Centre d'études sur les médias. *Comment les Québécois s'informent-ils?*, novembre 2013.

L'étude vise à déterminer par sondage le temps moyen que les Québécois de moins de 35 ans passent chaque jour sur Internet à s'informer de l'actualité.

- a) Donner une estimation ponctuelle du temps moyen cherché et dire si celle-ci est acceptable. Justifier la réponse.

- b) En utilisant l'information donnée dans la méthodologie, construire et interpréter l'intervalle de confiance permettant d'estimer la moyenne de temps cherchée.

12. On veut estimer la consommation d'essence d'un modèle de voiture prétendument économique. Pour un échantillon de 16 voitures, on a obtenu la consommation ci-dessous, en litres par 100 kilomètres. Construire un intervalle de confiance au niveau de confiance de 99 % pour estimer la consommation moyenne de cette voiture. La distribution de la consommation d'essence obéit à une loi normale.

5,62	5,54	5,80	5,92
5,75	5,64	5,46	5,58
5,39	5,55	5,60	5,61
5,64	5,68	5,56	5,60

13. Calculer la taille minimale de l'échantillon à prélever pour estimer le poids moyen des sacs de sucre remplis par une machine, avec une marge d'erreur d'au plus 0,03 kg, en utilisant un intervalle de confiance au niveau de 99 %. On considère que la distribution du poids des sacs obéit à une loi normale dont l'écart type est de 0,1 kg.

14. Calculer la taille minimale de l'échantillon à prélever pour estimer à 500 \$ près le revenu familial moyen des familles d'un quartier, avec un niveau de confiance de 95 %, si l'on estime l'écart type des revenus à 3 500 \$.

15. Quelle taille d'échantillon devrait-on utiliser pour effectuer l'enquête du numéro 3 si l'on veut une marge d'erreur d'au plus 0,4 min ?

4.4 L'estimation d'un pourcentage d'une population

Nous venons de voir comment, à partir de la moyenne d'un échantillon, on peut estimer la moyenne d'une population pour des variables quantitatives comme l'âge, le salaire, le poids, etc. Nous allons maintenant travailler principalement avec des variables qualitatives telles que le sexe, les préférences, les opinions, etc. Nous chercherons à estimer, pour une des catégories d'une variable qualitative, le pourcentage d'unités de la population dans cette catégorie à partir du pourcentage observé dans un échantillon. Les sondages par lesquels on cherche à estimer un pourcentage pour une population sont ceux que l'on rencontre le plus souvent dans les journaux : tous les sondages portant sur les intentions de vote, l'opinion des consommateurs, l'appréciation des clients sont de ce type.

Nous aborderons ce sujet en suivant une démarche analogue à celle que nous avons suivie pour l'estimation d'une moyenne. Dans un premier temps, nous chercherons à prédire les valeurs que le hasard peut générer comme pourcentage échantillonnaux lorsque le pourcentage d'une population est connu. Une fois que nous aurons découvert la loi probabiliste qui s'applique, nous l'utiliserons pour estimer le pourcentage d'une population à l'aide de celui d'un échantillon.

Notation

Le pourcentage d'unités statistiques ayant une même caractéristique se note :

- p s'il s'agit d'unités statistiques d'une population ;
- \hat{p} (lire « p chapeau») s'il s'agit d'unités statistiques d'un échantillon.

NOTE

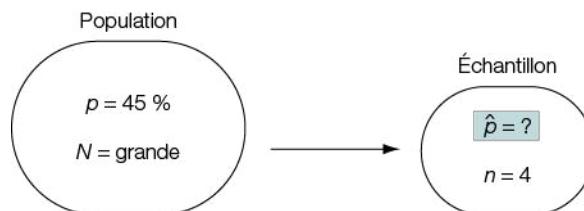
Dans certains ouvrages, on utilise la lettre grecque π , au lieu de p , pour noter le pourcentage d'unités statistiques d'une population.

4.4.1 La distribution d'échantillonnage d'un pourcentage

La mise en situation suivante va nous faire découvrir la loi de probabilité qui lie le pourcentage p de la population et le pourcentage \hat{p} d'un échantillon.

MISE EN SITUATION

Dans une grande population comptant 45 % de célibataires, on prélève sans remise un échantillon aléatoire de 4 personnes et on s'intéresse au pourcentage \hat{p} de célibataires dans l'échantillon.



Posons les questions suivantes :

- Q1.** Quelles sont les valeurs possibles pour \hat{p} ?
- Q2.** Quelles sont les chances que l'écart entre \hat{p} et p soit d'au plus 5 %?

Pour répondre à la question Q1, définissons les deux variables aléatoires suivantes :

- X : « le nombre de célibataires parmi les 4 personnes de l'échantillon ».

Les valeurs possibles pour X sont : 0, 1, 2, 3 et 4 ;

- \hat{P} : « le pourcentage de célibataires parmi les 4 personnes de l'échantillon ».

On a : $\hat{P} = \frac{X}{4}$. Les valeurs possibles pour \hat{P} sont : $\frac{0}{4}, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$ et $\frac{4}{4}$.

En pourcentage, la réponse à la question Q1 est : 0 %, 25 %, 50 %, 75 % et 100 %.

Pour répondre à la question Q2, présentons la distribution de probabilité de X et \hat{P} dans un même tableau.

Notons d'abord que la variable aléatoire X suit une distribution binomiale $B(4; 0,45)$.

- On a 4 épreuves indépendantes (tirage sans remise dans une grande population).
- À chaque épreuve, on obtient un succès ou un échec :
 - succès : « la personne est célibataire » avec $p = 0,45$;
 - échec : « la personne n'est pas célibataire » avec $q = 0,55$.
- La variable aléatoire X correspond au nombre de succès en 4 épreuves.

Distribution de probabilité de X et de \hat{P}

x	\hat{p}	$P(X = x)$ ou $P(\hat{P} = \hat{p})$
0	0 %	$\binom{4}{0}(0,45)^0(0,55)^4 = 0,0915 \approx 9,2\%$
1	25 %	$\binom{4}{1}(0,45)(0,55)^3 = 0,2995 \approx 30\%$
2	50 %	$\binom{4}{2}(0,45)^2(0,55)^2 = 0,3675 \approx 36,8\%$
3	75 %	$\binom{4}{3}(0,45)^3(0,55) = 0,2005 \approx 20,0\%$
4	100 %	$\binom{4}{4}(0,45)^4(0,55)^0 = 0,0410 \approx 4,1\%$

Nous pouvons maintenant répondre à la seconde question :

Q2. Quelles sont les chances que l'écart entre \hat{p} et p soit d'au plus 5 % ?

Comme $p = 45\%$, l'écart entre \hat{p} et p est d'au plus 5 % si le pourcentage de célibataires dans l'échantillon est compris entre 40 % et 50 %. Le tableau de distribution de probabilité indique que les chances d'obtenir un tel résultat sont de 37 %.

Distribution d'échantillonnage d'un pourcentage

On donne le nom de **distribution d'échantillonnage de \hat{P}** à la distribution de probabilité de la variable aléatoire \hat{P} . Cette distribution a les caractéristiques suivantes :

- L'espérance et l'écart type de la distribution d'échantillonnage de \hat{P}

L'espérance de la variable aléatoire \hat{P} est notée $\mu_{\hat{p}}$ et son écart type est noté $\sigma_{\hat{p}}$.

❓ Sachant que X suit une $B(4; 0,45)$, en moyenne, combien de célibataires peut-on espérer trouver dans un échantillon de 4 personnes? En déduire le pourcentage moyen de célibataires que l'on peut espérer.

$$E(X) =$$

$$E(\hat{P}) = \mu_{\hat{P}} =$$

En généralisant le raisonnement à une binomiale $B(n; p)$, on a :

$$E(\hat{P}) = \mu_{\hat{P}} = \frac{E(X)}{n} = \frac{np}{n} = p \text{ d'où l'égalité } \mu_{\hat{P}} = p$$

❓ Calculer l'écart type du nombre de célibataires que l'on peut espérer trouver dans un échantillon de 4 personnes et en déduire l'écart type du pourcentage de célibataires que l'on peut espérer.

$$\sigma(X) =$$

$$\sigma(\hat{P}) = \sigma_{\hat{P}} =$$

En généralisant le raisonnement à une binomiale $B(n; p)$, on a :

$$\sigma_{\hat{P}} = \frac{\sigma(X)}{n} = \frac{\sqrt{npq}}{n} = \sqrt{\frac{npq}{n^2}} = \sqrt{\frac{pq}{n}} \text{ d'où l'égalité } \sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$$

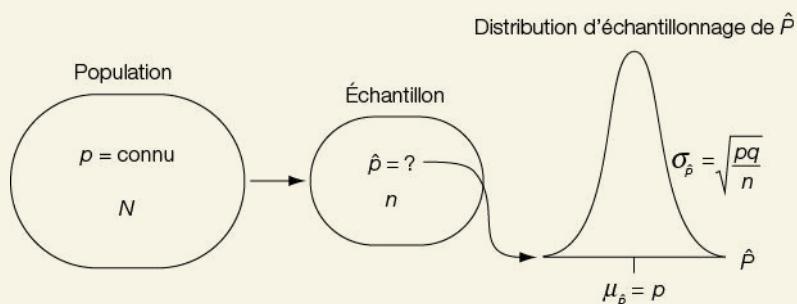
En vous servant du mode statistique de la calculatrice, vous pouvez vérifier l'exactitude de ces mesures en entrant les données de la distribution de probabilité de la page précédente.

- La forme de la distribution d'échantillonnage de \hat{P}

On sait que l'on peut utiliser une loi normale pour faire une approximation d'une loi binomiale $B(n; p)$ quand $np \geq 5$ et $nq \geq 5$. Si l'on ajoute $n \geq 30$ à ces conditions, on peut aussi utiliser une loi normale pour faire une approximation de la distribution d'échantillonnage de \hat{P} .

Théorème central limite pour un pourcentage

Si l'on prélève un échantillon aléatoire de taille n dans une population ayant un pourcentage p d'unités statistiques possédant une même caractéristique, alors la distribution d'échantillonnage de \hat{P} (pourcentage d'unités de l'échantillon ayant cette caractéristique) possède les propriétés suivantes :



- Sa moyenne $\mu_{\hat{p}}$ est égale au pourcentage p de la population : $\mu_{\hat{p}} = p$.
- Son écart type est :

$$\sigma_{\hat{p}} = \begin{cases} \sqrt{\frac{pq}{n}} & \text{si la population est grande } (N \geq 20n); \\ \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}} & \text{si la population est petite } (N < 20n). \end{cases}$$

On donne le nom de **facteur de correction** à l'expression $\sqrt{(N-n)/(N-1)}$.

- Sa forme est celle d'une courbe normale si les conditions suivantes sont réunies :
 - 1) $n \geq 30$
 - 2) $np \geq 5$
 - 3) $nq \geq 5$

NOTE

Si les conditions exigées pour appliquer le modèle normal ne sont pas respectées, on utilise la loi binomiale après avoir converti le pourcentage de succès désiré en un nombre de succès.

EXEMPLE

Dans un cégep, 22 % des 3 500 étudiants sont inscrits en sciences de la nature. On prélève un échantillon de 400 étudiants et l'on s'intéresse au pourcentage d'étudiants en sciences de la nature dans l'échantillon.

- a) Soit les variables aléatoires suivantes :

X : «le nombre d'étudiants en sciences de la nature parmi les 400 étudiants de l'échantillon».

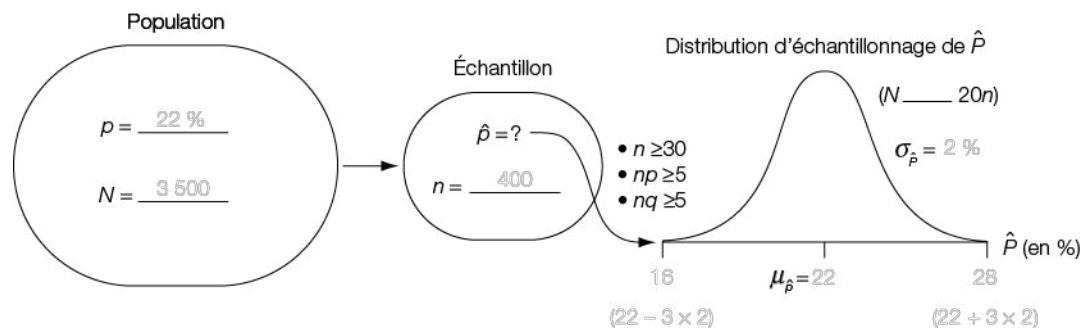
\hat{P} : «le pourcentage d'étudiants inscrits en sciences de la nature parmi les 400 étudiants de l'échantillon».

On a $\hat{P} = \frac{X}{400}$, où X suit une distribution binomiale.

Vérifier si l'on respecte les trois conditions permettant d'affirmer que la distribution d'échantillonnage de \hat{P} suit un modèle normal.

Solution

- b) Compléter le graphique de manière à refléter la situation décrite.

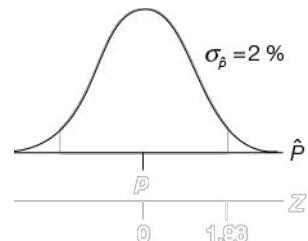


Solution

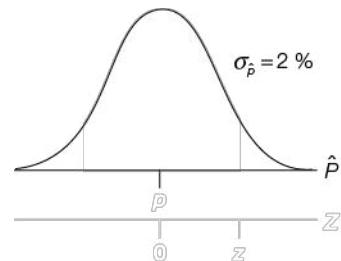
- c) Encercler le pourcentage qui a très peu de chances d'être obtenu comme pourcentage échantillonnaux d'étudiants inscrits en sciences de la nature.

25,7 % 18,2 % 14,3 % 26,6 %

- d) Pour 95 % des échantillons possibles, quelle est la valeur maximale de l'écart entre le pourcentage \hat{p} de l'échantillon et le pourcentage p de la population ? Représenter la situation sur la courbe normale.

Solution

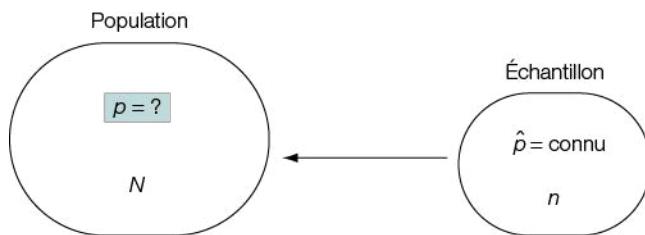
- e) Quelle est la probabilité qu'il y ait un écart d'au plus 2,5 % entre \hat{p} et p ? Représenter la situation sur la courbe normale.

Solution

4.4.2 L'estimation d'un pourcentage par intervalle de confiance

Maintenant que nous connaissons la distribution d'échantillonnage de \hat{P} , nous sommes en mesure de répondre à la question suivante :

« Si l'on connaît le pourcentage \hat{p} d'un échantillon, peut-on estimer le pourcentage p de la population ? »



Comme nous venons de le voir, on sait qu'il y aura un écart E , associé à une probabilité, entre p et \hat{p} . Par conséquent, affirmer que le pourcentage p de la population est égal au pourcentage \hat{p} de l'échantillon constituerait une erreur d'estimation pouvant atteindre la valeur de l'écart E . En tenant compte de ce fait, on estime la valeur de p par intervalle de confiance en posant $p = \hat{p} \pm E$, donc $p \in [\hat{p} - E; \hat{p} + E]$.

On procède de la même façon pour construire un intervalle de confiance pour un pourcentage p que pour construire un intervalle de confiance pour une moyenne μ .

Estimation d'un pourcentage d'une population par intervalle de confiance

Voici la marche à suivre pour construire un intervalle de confiance pour p :

- Vérifier les conditions d'application.
- Déterminer l'écart type $\sigma_{\hat{p}}$ de la distribution d'échantillonnage de \hat{P} .
- Calculer la marge d'erreur associée au niveau de confiance souhaité : $E = z\sigma_{\hat{p}}$.
- Calculer les bornes de l'intervalle de confiance et l'interpréter : $p \in [\hat{p} - E; \hat{p} + E]$.

Il reste un dernier problème à résoudre avant de pouvoir construire un intervalle de confiance pour p . Comment calculer la valeur de l'écart type $\sigma_{\hat{p}}$ quand on ne connaît pas le pourcentage p de la population ?

Pour contourner cette difficulté, il est d'usage de remplacer p par \hat{p} dans la formule de $\sigma_{\hat{p}}$:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \approx \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

NOTE

Avec la même logique, on dira que la distribution d'échantillonnage de \hat{P} suit une loi normale si :

$$n \geq 30, n\hat{p} \geq 5 \text{ et } n\hat{q} \geq 5$$

EXEMPLE

Le problème suivant est inspiré des résultats d'un sondage publié dans *Le Journal de Québec* du 11 mars 2012.

Les deux solitudes s'éloignent

Il y a vraiment deux Canada en un. Le sondage Léger Marketing publié aujourd’hui montre à quel point les Québécois sont distincts des autres Canadiens.

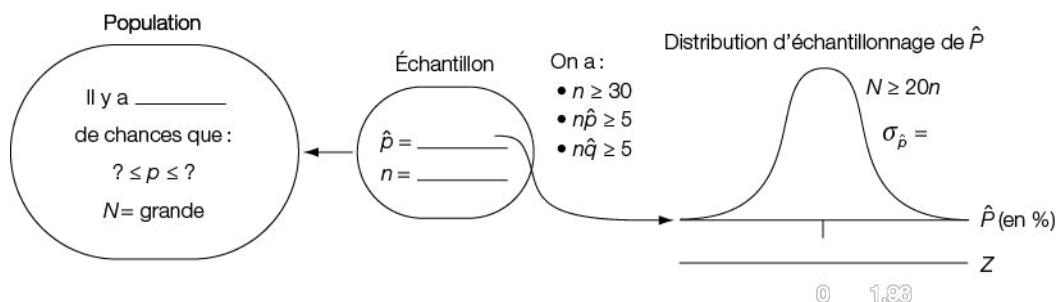
- D'une part, les Québécois sont proportionnellement plus nombreux que les Canadiens à être d'avis que les choses vont mal au Canada (71 % contre 43 %) et à être favorables au droit à l'avortement (85 % contre 66 %).
- D'autre part, ils sont, toujours en proportion, moins nombreux que les Canadiens à se dire favorables : à l'extraction du pétrole des sables bitumineux (36 % contre 63 %); à la mise en valeur de la monarchie (9 % contre 36 %); au financement accru de l'armée canadienne (19 % contre 37 %).

Méthodologie

Ce sondage a été réalisé du 28 février au 5 mars 2012 par Léger Marketing. Les résultats reposent sur 2 509 entrevues téléphoniques : 1 001 au Québec et 1 508 dans le reste du Canada. La marge d'erreur est d'au plus 3,1 % pour l'échantillon québécois et d'au plus 2,5 % pour l'échantillon hors Québec, et cela, 19 fois sur 20.

- a) Estimer par intervalle de confiance au niveau de 95 % le pourcentage des Québécois qui sont d'avis que les choses vont mal au Canada, sachant que 71 % des 1 001 Québécois interrogés dans le sondage sont de cet avis.

Solution



On a :

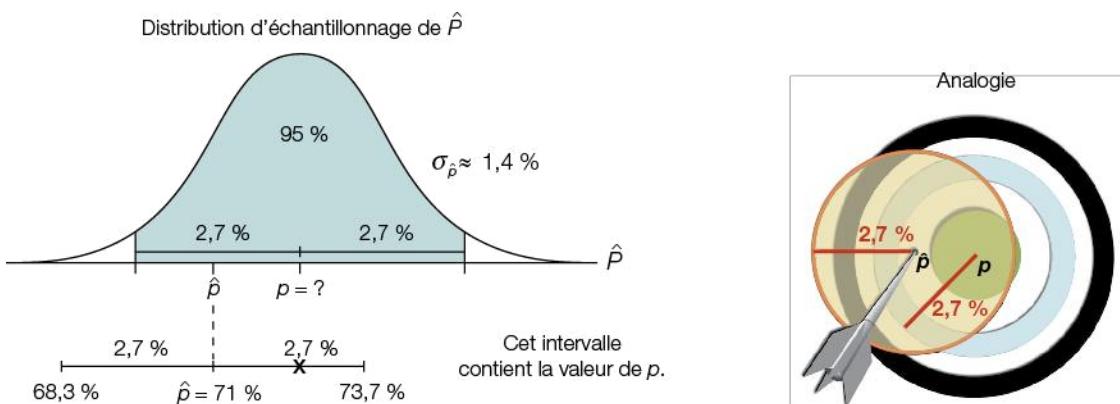
- $n = 1\,001 \geq 30$
- $n\hat{p} = 1\,001 \times 0,71 \approx 710 \geq 5$
- $n\hat{q} = 1\,001 \times 0,29 \approx 290 \geq 5$

Conclusion du sondage

Il y a 95 % de chances que le pourcentage réel de Québécois qui sont d'avis que les choses vont mal au Canada se situe entre _____ et _____.

Représentation graphique de l'intervalle de confiance

En supposant que le pourcentage de l'échantillon soit effectivement dans la zone qui contient 95 % des pourcentages d'échantillons possibles, le graphique suivant pourrait représenter l'intervalle de confiance construit (la position de \hat{p} a été choisie arbitrairement dans la zone de 95 %). Il faut bien observer que le centre de l'intervalle de confiance est \hat{p} et non p . La position de p sur l'intervalle de confiance est indiquée par une croix : l'écart entre \hat{p} et p est à l'intérieur de la marge d'erreur de 2,7 %. L'analogie avec le jeu de fléchettes est toujours valable : dans ce cas-ci, le centre de la cible est p et le trou fait par la fléchette est \hat{p} .



- b) En sachant que, dans l'échantillon de 1 508 Canadiens hors Québec, 63 % sont favorables à l'extraction du pétrole des sables bitumineux, construire un intervalle de confiance, au niveau de confiance de 95 %, qui permet d'estimer le pourcentage réel de Canadiens hors Québec qui partagent cette opinion.

Solution

- On a: $n = 1\ 508 \geq 30$; $n\hat{p} = 1\ 508 \times 0,63 \geq 5$; $n\hat{q} = 1\ 508 \times 0,37 \geq 5$
- $\sigma_{\hat{p}} = \sqrt{\frac{63 \times 37}{1508}} = 1,2 \%$
- Marge d'erreur:
 $E = z\sigma_{\hat{p}} = 1,96 \times 1,2 \% \approx 2,4 \%$
- Intervalle de confiance:
 $p = 63 \% \pm 2,4 \%$
 $p \in [60,6 \% ; 65,4 \%]$

Conclusion du sondage

On peut estimer qu'il y a 95 % de chances que le pourcentage réel de Canadiens hors Québec qui sont favorables à l'extraction du pétrole des sables bitumineux se situe entre 60,6 % et 65,4 %.

NOTE

Généralement, un sondage d'opinion contient plusieurs questions, et donc autant de pourcentages échantillonnaux et de marges d'erreur. À la publication du sondage, il est impensable de donner toutes les marges d'erreur associées à chaque question dans la méthodologie. En pratique, on donnera la plus grande marge d'erreur que l'on peut obtenir pour le niveau de confiance et la taille de l'échantillon considéré, soit celle associée à un pourcentage de 50 %. En effet, pour une même valeur n et z , plus le produit $\hat{p}\hat{q}$ est grand, plus la marge d'erreur E est grande : c'est pour $\hat{p} = 50\%$ que ce produit atteint sa valeur maximale, soit $50\% \times 50\% = 25\%$.

À titre de comparaison, pour $\hat{p} = 10\%$, on obtient $10\% \times 90\% = 9\%$; pour $\hat{p} = 40\%$, on a $40\% \times 60\% = 24\%$; pour $\hat{p} = 80\%$, on a $80\% \times 20\% = 16\%$, etc. Tous ces produits sont inférieurs à 25 %.

EXERCICE DE COMPRÉHENSION | 4.4

Pour affronter une concurrence de plus en plus grande dans la restauration, la chaîne de restaurants Coq-délices a procédé à un vaste plan de relance de son entreprise. Pour connaître l'opinion de sa clientèle sur les changements apportés, elle a mené un sondage auprès de 800 clients. Ces clients ont notamment été invités à répondre à la question suivante :

« Dans l'ensemble, quel est votre niveau de satisfaction quant au rapport qualité-prix de notre nouveau menu ? »

Répartition des répondants selon le niveau de satisfaction quant au rapport qualité-prix

Niveau de satisfaction	Très satisfait	Satisfait	Insatisfait	Pas d'opinion	Total
Nombre de répondants	320	256	192	32	800

- a) En se basant sur les résultats de l'échantillon, et en acceptant de courir le risque de se tromper 2 fois sur 20, estimer le véritable pourcentage de clients satisfaits ou très satisfaits du rapport qualité-prix du nouveau menu.

Solution

On a $\hat{p} =$ _____ et un niveau de confiance de _____.

Les conditions d'application du modèle normal sont vérifiées : $n \geq 30$, $n\hat{p} \geq 5$ et $n\hat{q} \geq 5$.

- ▶ b) Voici la façon dont les résultats du sondage ont été publiés dans la chronique «La bonne bouffe» d'un quotidien et à la radio. Compléter l'énoncé.

Dans le quotidien :

La relance de Coq-délices, un succès !

Les efforts faits par Coq-délices depuis un an pour conserver sa part du marché de la restauration semblent avoir porté leurs fruits.

Un récent sondage confirme en effet que sa clientèle accueille favorablement les changements apportés. Ce sondage révèle entre autres que les clients semblent grandement apprécier le rapport qualité-prix du nouveau menu puisque _____ % d'entre eux s'en sont montrés satisfaits ou très satisfaits. [...]

Méthodologie

Ce sondage a été mené auprès d'un échantillon de _____ clients. Avec un échantillon de cette taille, la marge d'erreur de l'estimation est de _____ %, _____ fois sur 20.

À la radio :

«Le nouveau menu des restaurants Coq-délices est apparemment bien accueilli par sa clientèle. C'est du moins ce qui semble se dégager d'un récent sondage mené par cette entreprise, où _____ % des clients ont dit apprécier le rapport qualité-prix du nouveau menu.»

- c) Quel média ne donne qu'une estimation ponctuelle du résultat du sondage ? _____

4.4.3 Le choix de la taille de l'échantillon

Tout comme nous l'avons fait pour une moyenne, nous pouvons fixer d'avance la marge d'erreur maximale désirée et choisir la taille de l'échantillon en conséquence. Plus l'échantillon est grand, plus la marge d'erreur est petite mais, en revanche, plus les coûts du sondage sont élevés.

EXEMPLE

Quelle devrait être la taille de l'échantillon à prélever si l'on désire estimer le pourcentage des électeurs qui appuient le parti A avec une marge d'erreur inférieure à 2 %, au niveau de confiance de 95 % ?

Solution

On cherche n tel que la marge d'erreur maximale est $E = 2\%$, pour un niveau de confiance de 95 %. Écrivons la formule exprimant la marge d'erreur :

$$E = z\sigma_{\hat{p}} = z\sqrt{\frac{pq}{n}} \approx z\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Dans cette équation, on connaît les valeurs de E et de z , mais pas celle de \hat{p} puisque l'étude n'est pas commencée. Si l'on ignore la valeur approximative de \hat{p} , on lui attribue la valeur 50 %, soit celle qui donne au produit $\hat{p}\hat{q}$ sa valeur maximale.

Remplaçons E par 2 et \hat{p} et \hat{q} par 50 (en %) dans l'équation :

$$2 = 1,96\sqrt{\frac{50 \times 50}{n}}$$

Afin d'isoler l'inconnue n , il faut éléver au carré chaque terme de l'équation :

$$2^2 = 1,96^2 \left(\frac{50 \times 50}{n} \right)$$
$$n = \frac{1,96^2 \times 50 \times 50}{2^2} = 2\,401$$

Si l'on veut que la marge d'erreur soit inférieure à 2 %, il faut prélever un échantillon comptant au moins 2 401 électeurs.

NOTE

La taille minimale exigée pour une marge d'erreur inférieure à 2 % pourrait être plus petite que 2 401 s'il s'avérait que le pourcentage, estimé ici à 50 %, soit assez différent de cette valeur. On peut supposer, par exemple, qu'un sondage préliminaire effectué auprès de 60 électeurs indique que le pourcentage d'électeurs qui appuient le parti A se situe plutôt autour de 10 %. En calculant à nouveau la taille d'échantillon nécessaire pour une marge d'erreur maximale de 2 %, on obtient :

$$2 = 1,96 \sqrt{\frac{10 \times 90}{n}}, \text{ ce qui donne } n = \frac{(1,96^2) \times 900}{2^2} \approx 865$$

Un minimum de 865 électeurs serait alors suffisant pour obtenir une marge d'erreur inférieure à 2 % dans le sondage.

EXERCICE DE COMPRÉHENSION | 4.5

Afin d'inciter ses citoyens à économiser l'eau potable, une municipalité songe à instaurer une tarification de l'eau en fonction du volume consommé par résidence. Pour savoir si ce projet recevra un bon accueil dans la population, elle demande à une firme d'effectuer un sondage visant à estimer le pourcentage de citoyens qui appuieraient un tel projet.

- a) Quelle taille devrait avoir l'échantillon si l'on veut que la marge d'erreur de l'estimation n'excède pas 3 %, avec un niveau de confiance de 95 % ?

- b) Quelle taille devrait avoir l'échantillon si, *a priori*, on estime à environ 20 % le pourcentage de personnes favorables au projet ?

4.4.4 La répartition des indécis

Au cours d'un sondage d'opinion, il arrive souvent que des personnes se disent indécises ou refusent de répondre à certaines questions : c'est le cas notamment lorsqu'on demande aux répondants leurs intentions de vote à une élection ou à un référendum. Si leur nombre est élevé, il peut alors être difficile de prédire l'issue du scrutin.

Supposons, à titre d'exemple, qu'un sondage effectué deux semaines avant un référendum donne les résultats présentés dans le tableau suivant.

Intentions de vote	Nombre de répondants	Pourcentage
Option OUI	420	42 %
Option NON	380	38 %
Indécis ou refus de répondre	200	20 %
Total	1 000	100 %

Ces statistiques indiquent que, parmi les personnes sondées, l'option OUI est en avance. Elles révèlent également qu'il y a un pourcentage élevé de répondants indécis ou qui refusent de répondre. Le jour du vote, ces électeurs devront se prononcer et leur choix sera déterminant pour l'issue du scrutin. C'est pourquoi les spécialistes des sondages essaient souvent de prédire comment ils se répartiront entre les différentes options le jour du vote. Ce n'est pas une tâche facile : il faut beaucoup de flair et d'expérience pour prévoir le comportement de cette catégorie de répondants.

On avance souvent l'hypothèse que, le jour du scrutin, ces personnes voteront globalement dans les mêmes proportions que celles qui se sont prononcées lors du sondage. En langage mathématique, cela signifie que, si les 1 000 personnes de l'échantillon avaient répondu par oui ou par non, elles l'auraient fait de la même façon que les 800 personnes qui ont donné une de ces deux réponses. Sous cette hypothèse, on obtient les pourcentages suivants :

$$\text{Pourcentage pour le OUI parmi les répondants} = \frac{420}{800} \times 100 = 52,5 \%$$

$$\text{Pourcentage pour le NON parmi les répondants} = \frac{380}{800} \times 100 = 47,5 \%$$

Intentions de vote	Nombre de répondants	Pourcentage	Pourcentage après répartition des répondants indécis ou ayant refusé de répondre
Option OUI	420	42 %	52,5 %
Option NON	380	38 %	47,5 %
Indécis ou refus de répondre	200	20 %	—
Total	1 000	100 %	100,0 %

Cette répartition proportionnelle des répondants indécis ou refusant de répondre confirme l'avantage du camp du OUI.

On peut émettre d'autres hypothèses sur la répartition de cette catégorie de répondants : par exemple, le jour du scrutin, ils voteront de la même façon que les personnes qui se sont prononcées et qui ont les mêmes caractéristiques socioéconomiques qu'eux (sexe, âge, langue maternelle, etc.). Ce type d'hypothèse exige que les sondeurs fassent une analyse très fine du profil de ces personnes. Le calcul du pourcentage de répondants indécis ou ayant refusé de répondre qu'il faut attribuer à chacune des options devient alors très complexe. Ce n'est que le jour du scrutin que l'on pourra vérifier la justesse de l'hypothèse de répartition.

Anecdote

Avant le référendum du 30 octobre 1995 sur la souveraineté du Québec, les camps du OUI (les souverainistes) et du NON (les fédéralistes) courtisaient avec ferveur les indécis. Plusieurs sondages indiquaient que leurs votes décideraient de l'issue du référendum. Les spécialistes étaient nombreux à spéculer sur la répartition du vote de cette catégorie de répondants ; l'un d'entre eux émit l'hypothèse que la répartition serait la suivante : 2/3 pour le NON et 1/3 pour le OUI. Le résultat du référendum lui a donné raison ! Grâce à cette répartition, le camp du NON l'a finalement emporté par moins de 1 % des votes (54 288 votes), les résultats ayant été NON : 50,6 % et OUI : 49,4 %.

Si l'on applique ce type de répartition à notre exemple, on obtient les résultats suivants :

Intentions de vote	Nombre de répondants	Pourcentage après répartition des répondants indécis ou ayant refusé de répondre
Option OUI	$420 + (200 \times 33,3 \%) = 487$	48,7 %
Option NON	$380 + (200 \times 66,7 \%) = 513$	51,3 %
Indécis ou refus de répondre	—	—
Total	1 000	100,0 %

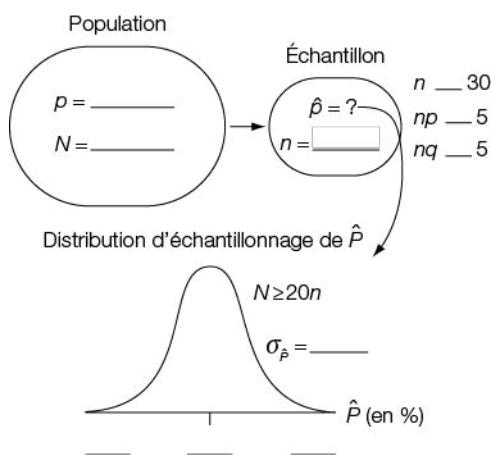
L'option gagnante n'est plus la même ! On comprend pourquoi les bons sondeurs se montrent prudents lorsqu'ils spéculent sur la répartition des indécis et de ceux qui ont refusé de répondre. Il y va de leur réputation.

EXERCICES 4.4

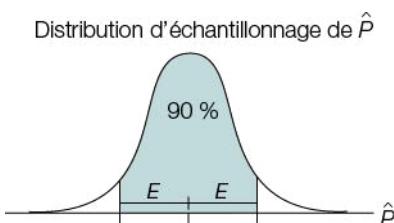
1. En 2011, 38 % des couples québécois vivent en union libre. On prélève un échantillon de 250 couples parmi tous les couples du Québec et l'on s'intéresse au pourcentage de couples en union libre dans l'échantillon.

Source: Statistique Canada. *Recensement 2011*.

- a) En négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues, quels sont le plus petit et le plus grand pourcentage échantillonnel que le hasard peut générer? Compléter l'information du graphique de manière à illustrer la situation.



- b) i) Pour 80 % de tous les échantillons possibles, l'écart maximal entre le pourcentage \hat{p} de couples en union libre dans l'échantillon et le pourcentage de 38 % de couples en union libre dans la population sera de quelle valeur?
 ii) On trouve 42,6 % de couples en union libre dans l'échantillon. L'échantillon prélevé fait-il partie des 80 % d'échantillons considérés en i)?
 iii) Le pourcentage de 42,6 % trouvé dans l'échantillon se situe-t-il dans la zone bleue du graphique ci-dessous? Si oui, indiquer la position approximative de \hat{p} .



2. Une machine produit 2 % de pièces défectueuses. On prélève un échantillon aléatoire de 100 pièces dans la production de cette machine et on s'intéresse au pourcentage de pièces défectueuses que le hasard peut générer dans l'échantillon.

- a) Dans ce cas-ci, peut-on affirmer que la distribution d'échantillonnage de \hat{P} suit une loi normale? Justifier la réponse.

- b) Utiliser la loi binomiale pour déterminer les chances que le pourcentage de pièces défectueuses dans l'échantillon soit de 3 % ou plus.

3. Dans une ville, 52 % de la population est de sexe féminin. On prélève un échantillon de 1 000 personnes dans cette population.

- a) Quelles sont les chances que le pourcentage de femmes dans l'échantillon se situe à au plus 2 % du pourcentage de femmes dans la population?

- b) Pour 95 % des échantillons possibles, l'écart entre le pourcentage de l'échantillon et celui de la population est inférieur à une certaine valeur. Laquelle?

4. Les 4 536 employés d'une usine de textile se répartissent ainsi: 3 280 femmes et 1 256 hommes. On projette d'effectuer un sondage auprès des employés, au niveau de confiance de 95 %.

- a) Le tableau suivant indique, pour différentes tailles d'échantillon, l'écart maximal que le hasard peut produire entre le pourcentage de femmes dans l'échantillon et le pourcentage à l'usine. Compléter le tableau.

Marge d'erreur selon la taille de l'échantillon

Taille de l'échantillon	Écart type $\sigma_{\hat{p}}$	Marge d'erreur E
$n = 100$		
$n = 150$		
$n = 200$		

- b) Quel est l'effet de l'augmentation de la taille de l'échantillon sur la marge d'erreur?

5. Utiliser l'information présentée dans l'article suivant pour répondre aux questions.

L'engouement pour les médias sociaux

Un sondage effectué en 2013 révèle que 82 % des internautes québécois utilisent les médias sociaux, alors qu'ils étaient 73 % à le faire en 2011. Cependant, les jeunes internautes demeurent plus actifs sur les médias sociaux que leurs aînés: 92 % chez les 18 à 25 ans contre moins de 62 % chez les 65 ans et plus. [...]

Méthodologie

Ce sondage a été effectué en 2013 auprès d'un échantillon aléatoire de 764 internautes âgés de 18 ans et plus. Avec un échantillon de cette taille, la marge d'erreur est d'au plus _____, 19 fois sur 20. Rappelons que la marge d'erreur tend à augmenter lorsque les résultats portent sur des sous-groupes.

Source: CEFARIO. NETendances 2013: Les adultes québécois toujours très actifs sur les médias sociaux, vol. 4, n° 1.

- a) Compléter la méthodologie du sondage en calculant la marge d'erreur.
- b) Construire et interpréter l'intervalle de confiance permettant d'estimer le pourcentage d'internautes québécois qui utilisent les médias sociaux.
- c) Si l'on augmente le niveau de confiance à 99 %:
 - i) la marge d'erreur sera-t-elle plus grande ou plus petite ?
 - ii) le risque d'erreur sera-t-il plus grand ou plus petit ?
 - iii) l'intervalle de confiance sera-t-il plus grand ou plus petit ?
- d) Calculer la marge d'erreur du pourcentage d'internautes de 65 ans et plus qui sont actifs sur les médias sociaux, sachant que 100 répondants sont dans cette tranche d'âge. Une estimation ponctuelle est-elle acceptable pour ce sous-groupe d'internautes ?

6. AVERTISSEMENT: CE PROBLÈME PEUT CONTENIR DES ARACHIDES.

Une étude québécoise a mesuré la prévalence des allergies alimentaires chez les enfants de 5 à 9 ans. Parmi les 1 693 enfants de l'échantillon, 25 sont allergiques aux arachides.

Source: Association québécoise des allergies alimentaires et Centre universitaire McGill.

a) Construire et interpréter l'intervalle de confiance au niveau de confiance de 90 % permettant d'estimer le pourcentage d'enfants québécois de 5 à 9 ans allergiques aux arachides.

b) Quels sont les risques que le pourcentage d'enfants allergiques aux arachides obtenu dans l'échantillon ne soit pas compris dans l'intervalle de confiance construit en a) ?

7.

Le divertissement en ligne

Le Web est devenu une source de divertissement de plus en plus populaire. Voici quelques statistiques à ce sujet: 48 % des adultes québécois regardent des vidéos sur Internet, 29 % regardent de la Webtélé; 33 % écoutent de la musique et 31 % jouent en ligne. [...]

Méthodologie

Les résultats du sondage reposent sur 1 000 entrevues téléphoniques effectuées en août 2013. Les répondants ont été choisis au hasard parmi les Québécois de 18 ans et plus. Avec un échantillon de cette taille, la marge d'erreur est d'au plus _____, 19 fois sur 20.

Source: CEFARIO. NETendances 2013: Divertissement en ligne: webtélé et téléviseur branché s'imposent, vol. 4, n° 4.

- a) En considérant la note de la page 227, compléter la méthodologie du sondage en calculant la marge d'erreur.
 - b) Calculer la marge d'erreur associée au pourcentage d'adultes québécois qui jouent en ligne et dire si elle est compatible avec celle qui est indiquée dans la méthodologie.
8. En 2011, un sondage est effectué auprès des titulaires d'un baccalauréat en administration de la promotion 2009. Parmi les répondants entrés sur le marché du travail à la fin de leurs études, 97 % occupent un emploi à temps plein au moment de l'enquête. La marge d'erreur de cette estimation est de 2 %, 19 fois sur 20.

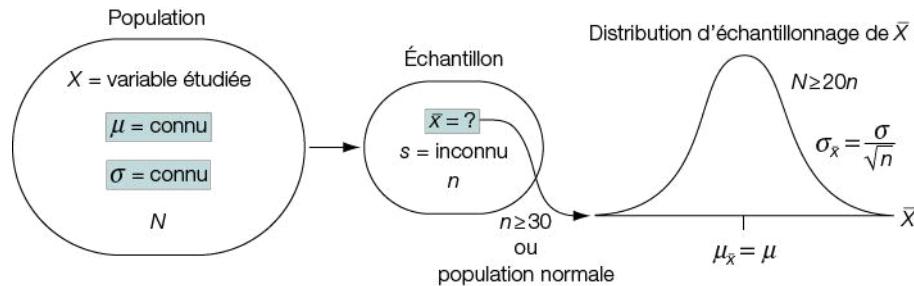
Source: Ministère de l'Enseignement supérieur. *La relance à l'université – 2011. La situation d'emploi de personnes diplômées. Enquêtes de 2007/2009/2011, 2012.*

- a) Le sondage visait à déterminer le pourcentage de titulaires d'un bac en administration qui occupent un emploi à temps plein 2 ans après l'obtention de leur diplôme.
 - i) Donner une estimation ponctuelle de ce pourcentage.

- ii) Donner une estimation par intervalle de confiance en utilisant la marge d'erreur donnée.
- b) Quel est le risque d'erreur (en %) qui s'applique à l'intervalle de confiance construit?
- c) Interpréter la phrase suivante : «La marge d'erreur de cette estimation est de 2 %, 19 fois sur 20.»
- d) i) Déterminer la taille de l'échantillon en utilisant l'estimation ponctuelle de p .
ii) Supposons que le sondeur ait envisagé au départ de poser plusieurs questions aux personnes sondées. Dans ce cas, quelle taille d'échantillon aurait-il dû prélever pour s'assurer que les estimations faites avec les différents pourcentages échantillonnaux auront une marge d'erreur d'au plus 2 %?
- e) Pour réduire la marge d'erreur à 1,5 %, aurait-il fallu:
i) augmenter ou diminuer la taille de l'échantillon?
ii) augmenter ou diminuer le niveau de confiance?
9. M. Tremblay est candidat aux prochaines élections. On veut estimer par sondage le pourcentage de votes qu'il recueillera.
- a) Quelle taille d'échantillon faut-il prendre pour que la marge d'erreur de l'estimation soit d'au plus 5 %, 19 fois sur 20?
- b) Parmi les personnes sondées, 160 sont en faveur de M. Tremblay. En utilisant un niveau de confiance de 95 %, estimer par intervalle de confiance le pourcentage d'électeurs favorables à M. Tremblay.
10. Un échantillon aléatoire de 625 électeurs est prélevé afin de déterminer le pourcentage d'électeurs favorables à un projet de loi. Sur les 625 personnes interrogées, 350 se déclarent en faveur du projet de loi.
- a) Estimer le pourcentage véritable des électeurs favorables au projet de loi à l'aide d'un intervalle de confiance au niveau de confiance de 95 %.
- b) Pour un même niveau de confiance, quelle taille d'échantillon faudrait-il prendre pour réduire la marge d'erreur du sondage à 2 %? Comme le sondage effectué auprès de 625 personnes donne un pourcentage échantillonnaux de 56 %, utiliser cette valeur pour déterminer la nouvelle taille d'échantillon.

RÉSUMÉ DU CHAPITRE 4

Distribution d'échantillonnage de \bar{X}



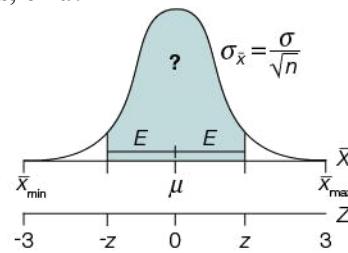
- En négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues, on a:

Plus petite moyenne échantillonnale : $\bar{x}_{\min} = \mu - 3\sigma_{\bar{x}}$.

Plus grande moyenne échantillonnale : $\bar{x}_{\max} = \mu + 3\sigma_{\bar{x}}$.

- Probabilité associée à un écart maximal E entre \bar{x} et μ :
 - représenter la situation graphiquement;
 - appliquer la définition de la cote z :

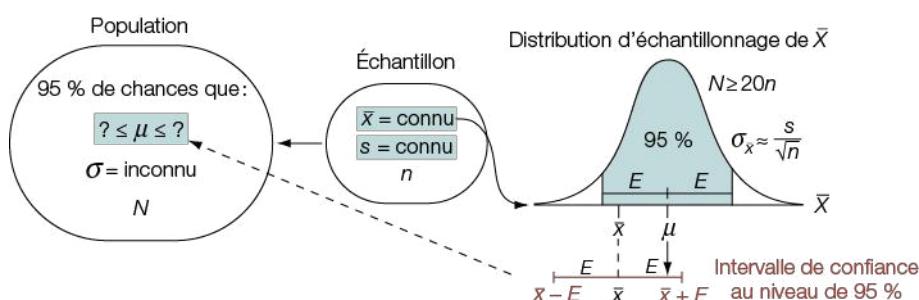
$$z = \frac{E}{\sigma_{\bar{x}}} \quad \text{ou} \quad E = z\sigma_{\bar{x}}$$



Attention !

Quand $N < 20n$ (petite population), on utilise le facteur de correction $\sqrt{(N-n)/(N-1)}$ dans le calcul de $\sigma_{\bar{x}}$

Estimation de la moyenne d'une population par intervalle de confiance



Démarche pour construire un intervalle de confiance pour μ :

- Vérifier les conditions d'application.
- Déterminer l'écart type $\sigma_{\bar{x}}$.
- Calculer la marge d'erreur associée au niveau de confiance souhaité : $E = z\sigma_{\bar{x}}$ ou $E = t\sigma_{\bar{x}}$.
- Calculer les bornes de l'intervalle de confiance et l'interpréter : $\mu \in [\bar{x} - E ; \bar{x} + E]$.

Attention !

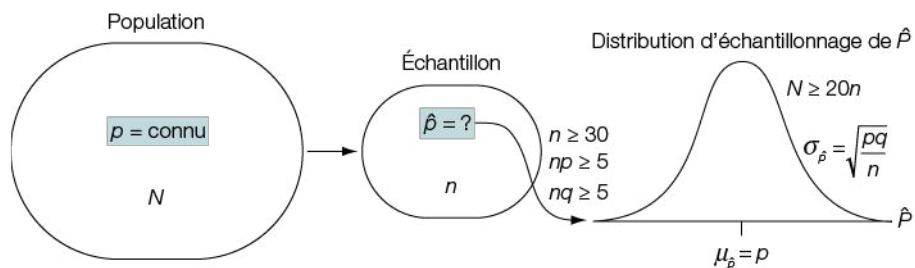
Si $n < 30$, σ inconnu et population normale, alors $E = t\sigma_{\bar{x}}$, où t suit une loi de Student avec $dl = n - 1$.

Estimation ponctuelle de la moyenne d'une population

On pose $\mu = \bar{x}$.

L'estimation est acceptable si la marge d'erreur est petite par rapport à la valeur de \bar{x} .

Distribution d'échantillonnage de \hat{P}

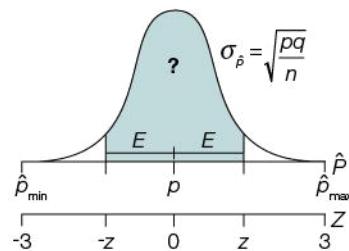


- En négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues, on a :

Plus petit pourcentage échantillonnaux : $\hat{p}_{\min} = p - 3\sigma_{\hat{P}}$.
Plus grand pourcentage échantillonnaux : $\hat{p}_{\max} = p + 3\sigma_{\hat{P}}$.

- Probabilité associée à un écart maximal E entre \hat{p} et p :
 - représenter la situation graphiquement ;
 - appliquer la définition de la cote z :

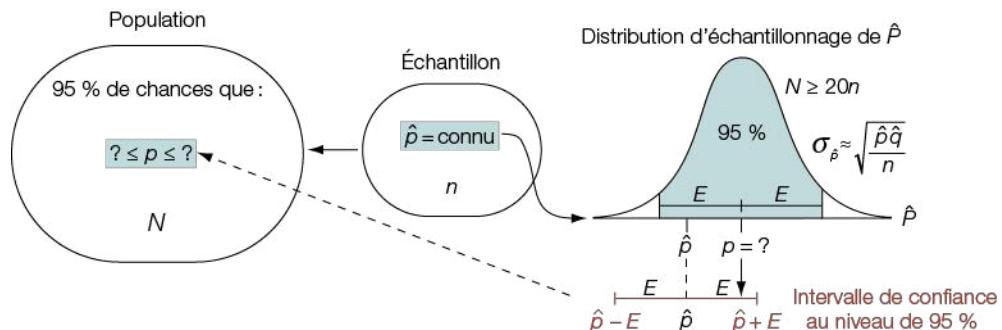
$$z = \frac{E}{\sigma_{\hat{P}}} \quad \text{ou} \quad E = z\sigma_{\hat{P}}$$



Attention !

Quand $N < 20n$ (petite population), on utilise le facteur de correction $\sqrt{(N-n)/(N-1)}$ dans le calcul de $\sigma_{\bar{x}}$.

Estimation d'un pourcentage d'une population par intervalle de confiance



Démarche pour construire un intervalle de confiance pour p :

- Vérifier les conditions d'application.
- Déterminer l'écart type $\sigma_{\hat{P}}$.
- Calculer la marge d'erreur associée au niveau de confiance souhaité : $E = z\sigma_{\hat{P}}$.
- Calculer les bornes de l'intervalle de confiance et l'interpréter : $p \in [\hat{p} - E; \hat{p} + E]$.

Estimation ponctuelle du pourcentage d'une population

On pose $p = \hat{p}$.

L'estimation est acceptable si la marge d'erreur est petite par rapport à la valeur de \hat{p} .

EXERCICES RÉCAPITULATIFS

1. Au cours de la saison théâtrale 2009-2010, 1 400 acteurs ont été embauchés pour jouer dans les 725 pièces produites. Le tableau ci-dessous donne la distribution de l'âge des acteurs.

**Répartition des acteurs selon l'âge,
saison théâtrale 2009-2010**

Âge	Nombre	Pourcentage
Moins de 25 ans	111	7,9 %
[25 ans; 35 ans[542	38,7 %
[35 ans; 45 ans[351	25,1 %
[45 ans; 55 ans[217	15,5 %
55 ans et plus	179	12,8 %
Total	1 400	100,0 %

Source: Conseil québécois du théâtre. *Profil statistique de la saison théâtrale 2009-2010*, octobre 2012.

- a) Quel groupe d'âge a été le plus avantage dans l'attribution des rôles ?
- b) Si l'on pige 60 artistes parmi les 1 400 acteurs embauchés pour la saison 2009-2010, serait-il surprenant d'obtenir une moyenne d'âge de 33 ans pour les 60 acteurs pigés ?
- c) Compléter l'énoncé. Il y a 85 % de chances que l'écart entre la moyenne d'âge des 60 acteurs pigés et la moyenne d'âge des 1 400 acteurs de la population étudiée soit d'au plus _____ ans.
- d) On pige 200 acteurs parmi les 1 400 acteurs de la population étudiée et l'on s'intéresse au pourcentage d'acteurs qui ont moins de 35 ans. Dans 90 % des cas, l'écart entre le pourcentage trouvé dans l'échantillon et le pourcentage d'acteurs de la population dans cette tranche d'âge sera inférieur à une certaine valeur. Laquelle ?
2. Afin de mieux connaître sa clientèle, un centre de jardinage demande à une firme d'effectuer un sondage auprès d'un échantillon aléatoire de 300 clients. Voici la distribution des réponses à quatre des questions posées. Il est à souligner que seulement 28 clients ont accepté de répondre à la question 2.

Q1. Quel est votre sexe ?

Sexe	Nombre de répondants
Féminin	170
Masculin	130
Total	300

Q2. Quel âge avez-vous ?

Âge	Nombre de répondants
Moins de 35 ans	5
[35 ans; 50 ans[11
[50 ans; 65 ans[6
65 ans et plus	6
Total	28

Q3. Quel est le montant de vos achats aujourd'hui ?

Montant (en \$)	Nombre de répondants
Moins de 25	95
[25; 50[83
[50; 75[68
[75; 100[30
100 et plus	24
Total	300

Q4. De quelle façon avez-vous payé vos achats ?

Mode de paiement	En argent	Carte de crédit	Carte de débit	Total
Nombre de répondants	90	150	60	300

- a) Donner une estimation ponctuelle du pourcentage de clients de sexe féminin. Quelle est la marge d'erreur de cette estimation, au niveau de confiance de 95 % ?
- b) En utilisant un niveau de confiance de 95 %, estimer entre quelles valeurs se situe la moyenne d'âge des clients de cette entreprise.
- c) Estimer par intervalle de confiance, au niveau de 95 %, le montant moyen des achats des clients et interpréter cet intervalle.
- d) Compléter l'énoncé. Il y a 95 % de chances que le pourcentage de clients du centre de jardinage qui utilisent la carte de débit se situe entre _____ % et _____ %.
- e) Écrire un court texte de style journalistique résumant les résultats du sondage : utilisation de l'estimation ponctuelle pour présenter les résultats, suivie de la méthodologie du sondage. La méthodologie doit contenir la taille de l'échantillon, la marge d'erreur de l'estimation et le niveau de confiance. (Dans ce cas-ci, indiquer la marge d'erreur de chaque variable estimée.)

PRÉPARATION À L'EXAMEN

Pour préparer votre examen, assurez-vous d'avoir les compétences suivantes :

	Si vous avez la compétence, cochez.
Échantillonnage	
• Différencier l'échantillonnage aléatoire de l'échantillonnage non aléatoire. <input type="radio"/>	
• Savoir effectuer un échantillonnage systématique. <input type="radio"/>	
• Reconnaître la méthode d'échantillonnage employée dans une situation donnée : aléatoire simple, systématique, stratifié, par grappes, à l'aveuglette, de volontaires, par quotas. <input type="radio"/>	
Distribution d'échantillonnage de \bar{X}	
• Trouver les caractéristiques de la distribution d'échantillonnage de \bar{X} : – forme de la distribution et conditions d'application ; <input type="radio"/>	
– moyenne de la distribution ; <input type="radio"/>	
– écart type de la distribution, que la population soit grande ou petite. <input type="radio"/>	
• Donner la plus petite et la plus grande moyenne d'échantillon que le hasard peut donner, en négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues. <input type="radio"/>	
• Calculer la probabilité associée à un écart maximal entre \bar{X} et μ , et vice versa. <input type="radio"/>	
Estimation d'une moyenne μ	
• Estimer μ de façon ponctuelle. <input type="radio"/>	
• Calculer la marge d'erreur E pour différents niveaux de confiance. <input type="radio"/>	
• Estimer μ par intervalle de confiance, selon différents niveaux de confiance, que σ soit connu ou non. <input type="radio"/>	
• Interpréter un niveau de confiance et un risque d'erreur. <input type="radio"/>	
• Prédire l'effet de la variation du niveau de confiance sur la marge d'erreur E . <input type="radio"/>	
• Rédiger et interpréter un texte donnant la méthodologie d'un sondage sur μ . <input type="radio"/>	
Taille de l'échantillon	
• Prédire l'effet de la variation de la taille de l'échantillon sur la marge d'erreur E . <input type="radio"/>	
• Évaluer la taille de l'échantillon nécessaire pour obtenir une marge d'erreur précisée. <input type="radio"/>	
Représentation graphique	
Être capable de représenter graphiquement :	
– la distribution d'échantillonnage de \bar{X} avec sa moyenne et son écart type ; <input type="radio"/>	
– la marge d'erreur E sur la distribution d'échantillonnage de \bar{X} ; <input type="radio"/>	
– un niveau de confiance et un risque d'erreur sur la distribution d'échantillonnage de \bar{X} . <input type="radio"/>	

Distribution d'échantillonnage de \hat{P}

- Trouver les caractéristiques de la distribution d'échantillonnage de \hat{P} :
 - forme de la distribution et conditions d'application;
 - moyenne de la distribution;
 - écart type de la distribution, que la population soit grande ou petite.
- Trouver le plus petit et le plus grand pourcentage d'échantillon que le hasard peut donner, en négligeant les valeurs ayant moins de 0,3 % de chances d'être obtenues.
- Calculer la probabilité associée à un écart maximal entre \hat{p} et p , et vice versa.

Estimation d'un pourcentage p

- Calculer la marge d'erreur E pour différents niveaux de confiance.
- Estimer p de façon ponctuelle.
- Estimer p par intervalle de confiance, selon différents niveaux de confiance.
- Interpréter un niveau de confiance et un risque d'erreur.
- Prédire l'effet de la variation du niveau de confiance sur la marge d'erreur E .
- Rédiger et interpréter un texte décrivant la méthodologie d'un sondage sur p .

Taille de l'échantillon

- Prédire l'effet de la variation de la taille de l'échantillon sur la marge d'erreur E .
- Évaluer la taille de l'échantillon nécessaire pour obtenir une marge d'erreur précisée:
 - lorsqu'on n'a aucune idée de la valeur de \hat{p} ;
 - lorsqu'on a une idée de la valeur approximative de \hat{p} .

Représentation graphique

Être capable de représenter graphiquement:

- la distribution d'échantillonnage de \hat{P} avec sa moyenne et son écart type;
- la marge d'erreur E sur la distribution d'échantillonnage de \hat{P} ;
- un niveau de confiance et un risque d'erreur sur la distribution d'échantillonnage de \hat{P} .



Chapitre 5

Les tests paramétriques



OBJECTIFS DU CHAPITRE

- Valider une hypothèse sur une moyenne ou un pourcentage.
- Valider une hypothèse sur l'égalité de deux moyennes ou de deux pourcentages.

OBJECTIF DU LABORATOIRE

Le laboratoire 4 vise, entre autres, à apprendre à utiliser les données échantillonnelles du laboratoire pour tester un pourcentage.



Dans l'estimation d'un paramètre, premier volet de l'inférence statistique, nous avons utilisé les statistiques d'un échantillon (\bar{x} , s , \hat{p}) pour estimer les paramètres d'une population (μ , σ , p). Dans le second volet de l'inférence, les statistiques de l'échantillon servent à valider une hypothèse quant à la modification possible des paramètres d'une population.

5.1 Le test d'hypothèse sur une moyenne

À partir de la mise en situation suivante, nous apprendrons à valider statistiquement une hypothèse formulée sur une modification possible de la moyenne d'une population en utilisant la moyenne d'un échantillon de grande taille ($n \geq 30$) tiré de cette population.

MISE EN SITUATION

Dans les années 1990, l'âge moyen des acheteurs d'une première voiture neuve était de 30 ans, avec un écart type de 6 ans. Un chercheur estime que cette moyenne ne correspond plus à la réalité d'aujourd'hui : il croit que les consommateurs se procurent une voiture neuve plus tôt qu'autrefois, délaissant ainsi le marché des voitures d'occasion. Pour vérifier son hypothèse voulant que la moyenne d'âge des acheteurs d'une première voiture neuve soit inférieure à 30 ans, notre chercheur décide de mener une étude statistique sur le sujet. Puisqu'il s'agit ici de tester une hypothèse, on donne le nom de **test d'hypothèse** à ce type d'étude.

Pour réaliser son étude, le chercheur suivra la démarche ci-dessous :

- Son hypothèse de travail sera d'accepter, jusqu'à preuve du contraire, que l'âge moyen de la population des acheteurs d'une première voiture neuve est encore de 30 ans aujourd'hui : soit $\mu = 30$ ans.

Son objectif consistera à démontrer que cette affirmation est fausse (on dira de rejeter l'hypothèse de travail $\mu = 30$ ans) et donc de voir ainsi son hypothèse de recherche acceptée : soit $\mu < 30$ ans.

- Il prendra la décision de rejeter ou non l'hypothèse voulant que la moyenne d'âge de la population soit de 30 ans, en se basant sur la valeur de la moyenne \bar{x} d'un échantillon de taille 36 prélevé au hasard dans la population des acheteurs d'une première voiture neuve.

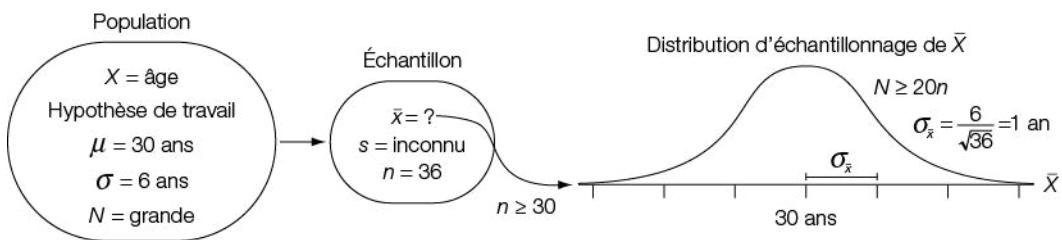
Analogie avec un procès

On peut établir une analogie entre un test d'hypothèse et un procès. Dans un test d'hypothèse, il y a un accusé ($\mu = 30$ ans) et un acte d'accusation ($\mu < 30$ ans). L'accusé, comme dans tout procès au Québec, bénéficie de la présomption d'innocence jusqu'à preuve du contraire. C'est le procureur (le chercheur) qui représente la poursuite : c'est donc lui qui doit apporter des preuves de la culpabilité de l'accusé ($\mu < 30$ ans). Un verdict de culpabilité sera prononcé si le procureur (le chercheur) fournit une preuve convaincante de la culpabilité de l'accusé (une moyenne d'échantillon convaincante).

Sur quel critère sera basée la décision ?

On sait bien qu'il y a peu de chances que la moyenne \bar{x} de l'échantillon prélevé soit égale à la moyenne μ de la population ; il y aura sûrement un écart entre \bar{x} et μ qui sera attribuable au hasard. La question est de savoir jusqu'où peut aller cet écart pour qu'il soit uniquement imputable au hasard de l'échantillonnage. Il faut donc déterminer, à l'aide de la distribution d'échantillonnage de \bar{X} , quelles valeurs la moyenne \bar{x}

a peu de chances de prendre lorsque la moyenne de la population est de 30 ans, avec un écart type de 6 ans. Voici une représentation graphique de la situation.



Le chercheur doit-il rejeter ou non l'hypothèse $\mu = 30 \text{ ans}$ si :

- a) la moyenne de l'échantillon est 29,4 ans ? _____
- b) la moyenne de l'échantillon est 26,7 ans ? _____
- c) la moyenne de l'échantillon est 27,5 ans ? _____

Interprétation des décisions

- a) Dans le premier cas, en se basant sur la distribution d'échantillonnage de \bar{X} , un écart de 0,6 an entre \bar{x} et μ est inférieur à un écart type ($\sigma_{\bar{X}} = 1 \text{ an}$), de sorte qu'on peut sûrement l'attribuer au hasard d'échantillonnage. On doit conclure que cet écart n'est pas assez grand (on dit qu'il est statistiquement non significatif) pour justifier le rejet de l'hypothèse de travail voulant que la moyenne d'âge μ des acheteurs d'une première voiture neuve soit encore de 30 ans aujourd'hui.
- b) Dans le deuxième cas, l'écart entre \bar{x} et μ est de 3,3 ans; donc, \bar{x} est à plus de 3 écarts types de la moyenne μ de la population. Du point de vue statistique, un tel écart est très grand; il a très peu de chances d'être obtenu dans le cas d'une distribution normale de moyenne $\mu = 30 \text{ ans}$. On peut donc, avec très peu de risques, rejeter l'hypothèse de travail voulant que $\mu = 30 \text{ ans}$.
- c) Pour ce qui est du dernier cas, il n'est pas facile de prendre une décision. La moyenne \bar{x} de l'échantillon se trouve à 2,5 écarts types de la moyenne μ de la population; les chances que le hasard donne un tel écart existent, mais elles sont faibles. Cet écart est-il assez grand pour décider de rejeter l'hypothèse de travail ? C'est une situation délicate; il faudrait avoir un critère pour nous aider à prendre une décision : par exemple, connaître un ou des points critiques à ne pas dépasser (la note de 60 % fixée comme frontière entre la réussite et l'échec d'un cours illustre bien cette idée).

Un test d'hypothèse sert à établir une règle permettant de décider du rejet ou du non-rejet d'une hypothèse dans une situation comme celle de la mise en situation.

Hypothèses d'un test

La première étape de la démarche de construction d'un test d'hypothèse consiste à formuler les hypothèses. Il en existe deux types.

H_0 : L'hypothèse nulle

L'hypothèse nulle spécifie la valeur acceptée jusqu'à maintenant comme moyenne μ de la population. C'est cette hypothèse qui est testée. Toute la démarche du test s'effectue en considérant cette hypothèse comme vraie jusqu'à preuve du contraire. Elle prend la forme suivante :

$$H_0: \mu = \mu_0$$

Par exemple, dans la mise en situation, on a $H_0: \mu = 30 \text{ ans}$.

H₁: L'hypothèse alternative ou hypothèse du chercheur

C'est l'hypothèse formulée par le chercheur au sujet d'une modification possible de la moyenne de la population. C'est celle qui sera acceptée si l'on rejette l'hypothèse H₀.

L'hypothèse alternative nous indiquera le type de test à réaliser: bilatéral ou unilatéral. Il y a trois sortes d'hypothèse H₁; c'est le contexte qui nous dira laquelle choisir.

H₁: $\mu \neq \mu_0$ Test bilatéral

H₁: $\mu > \mu_0$ Test unilatéral à droite

H₁: $\mu < \mu_0$ Test unilatéral à gauche

Par exemple, dans la mise en situation, on a H₁: $\mu < 30$ ans (test unilatéral à gauche).

EXEMPLE 1

Un chercheur émet l'hypothèse que l'âge moyen des femmes à leur premier mariage a augmenté depuis la dernière étude menée sur le sujet en 2011, qui avait établi la moyenne à 31 ans. Formuler les hypothèses H₀ et H₁.

Source: Institut de la statistique du Québec. *Les mariages au Québec en 2011*, juin 2014.

EXEMPLE 2

Une association de consommateurs examine un échantillon de 100 contenants de sirop d'érable pour vérifier si le volume moyen de sirop dans les contenants est bien de 540 ml, comme l'indique l'étiquette. Formuler les hypothèses H₀ et H₁.

EXEMPLE 3

Un producteur de sirop d'érable prélève un échantillon de 100 contenants de sirop dans la production d'une journée afin de s'assurer que le volume moyen de sirop est bien égal à 540 ml. Formuler les hypothèses H₀ et H₁.

Seuil de signification d'un test

Le seuil de signification d'un test, noté α , correspond aux risques de se tromper en prenant la décision de rejeter l'hypothèse nulle, c'est-à-dire aux risques de rejeter H₀ alors que cette hypothèse est vraie. On souhaite évidemment que ces risques soient faibles: par exemple, qu'ils ne dépassent pas 1 %. Le seuil de signification est fixé par le chercheur avant d'effectuer le test. Les valeurs de α les plus courantes sont 0,01, 0,05 et 0,10 (qu'on exprime aussi en pourcentage). Un seuil de signification est une probabilité qui s'exprime comme suit:

$$\alpha = P(\text{rejeter l'hypothèse } H_0 \text{ alors que } H_0 \text{ est vraie})$$

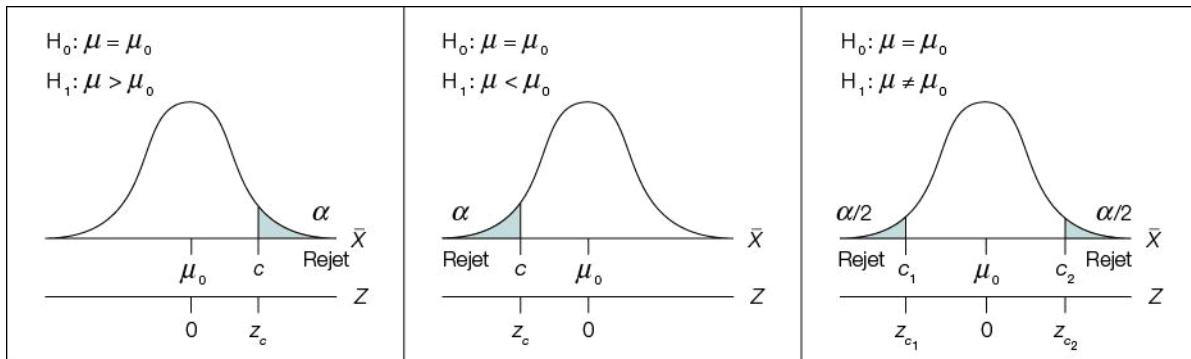
Règle de décision d'un test

Une règle de décision comporte un critère statistique sur lequel on s'appuie pour rejeter l'hypothèse nulle. Pour construire une règle de décision, il faut déterminer un **point critique c** sur l'axe des moyennes \bar{x} , en tenant compte du seuil de signification du test. On définit ensuite une **zone de rejet** de l'hypothèse H_0 et on énonce la règle de décision comme suit :

Règle de décision pour un test d'hypothèse sur une moyenne

Rejeter H_0 si la moyenne \bar{x} de l'échantillon se trouve dans la zone de rejet délimitée par le ou les points critiques.

Voici la représentation graphique de la zone de rejet pour chaque type de test.



Analogie avec un procès (suite)

Si l'on poursuit le parallèle avec un procès, le seuil de signification correspond au risque, toujours présent, de condamner un innocent. Quant à la règle de décision, c'est la loi qui guide les membres du jury au moment de prononcer le verdict sur la culpabilité de l'accusé.

MISE EN

SITUATION (suite)

Doit-on rejeter l'hypothèse selon laquelle l'âge moyen des acheteurs d'une première voiture neuve est encore de 30 ans aujourd'hui, si la moyenne d'âge des 36 personnes de l'échantillon est de 27,5 ans ? Faire un test d'hypothèse au seuil de signification de 0,01.

1. Formulons les hypothèses du test et donnons le seuil de signification.
2. Vérifions la condition d'application de la loi normale et représentons le test sur la courbe.

$$H_0: \mu = 30 \text{ ans}$$

$$H_1: \mu < 30 \text{ ans}$$

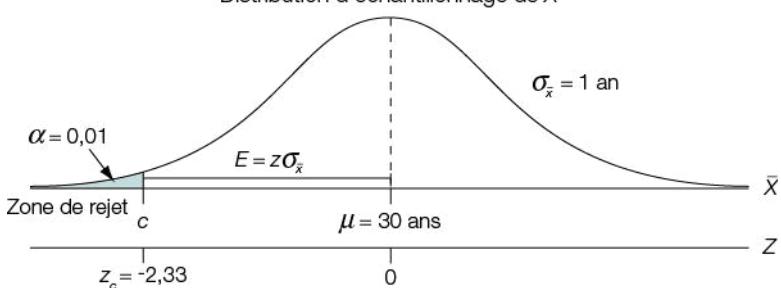
$$\alpha = 0,01$$

On a : $\bar{x} = 27,5$ ans

$$\sigma = 6 \text{ ans}$$

$$n = 36$$

Distribution d'échantillonnage de \bar{X}



3. Déterminons le point critique c .

- L'écart maximal tolérable entre \bar{x} et μ au seuil de 0,01 est : $E = z\sigma_{\bar{x}} = 2,33 \times 1 = 2,3$ ans.
- Le point critique est : $c = 30 - 2,3 = 27,7$ ans.

4. Énonçons la règle de décision.

Rejeter H_0 si la moyenne \bar{x} de l'échantillon est inférieure à 27,7 ans.

5. Décidons et concluons.

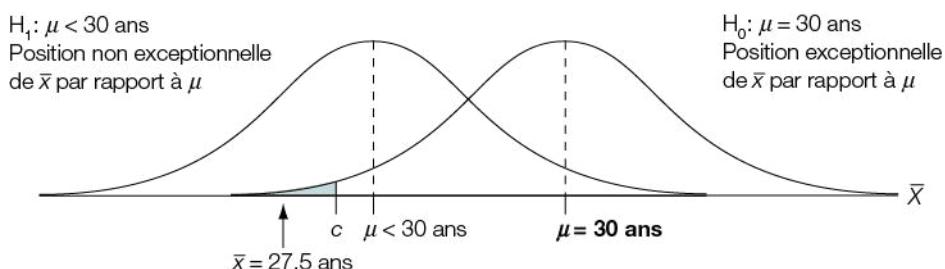
Comme $\bar{x} = 27,5$ ans < 27,7 ans, on rejette H_0 et on accepte H_1 .

Donc, l'âge moyen des acheteurs d'une première voiture neuve est inférieur à 30 ans.

Interprétation de la décision

Il est possible que la moyenne d'âge \bar{x} d'un échantillon s'écarte de plus de 2,33 écarts types de la moyenne d'âge μ de la population, mais comme les chances que cela se produise sont de moins de 1 % (1 fois sur 100), on pense qu'il est plus probable que l'échantillon prélevé provienne en fait d'une population ayant une moyenne d'âge inférieure à 30 ans. C'est pourquoi on décide de rejeter H_0 , en assumant un risque d'au plus 1 % de se tromper en prenant cette décision.

Les courbes suivantes illustrent bien le choix que l'on fait entre une position exceptionnelle de \bar{x} sur la courbe normale, sous l'hypothèse $H_0: \mu = 30$ ans, et une position non exceptionnelle de \bar{x} sur la courbe normale, sous l'hypothèse $H_1: \mu < 30$ ans : on parie sur H_1 avec moins de 1 % de risques de faire le mauvais choix.



Résumons la démarche à suivre pour effectuer un test d'hypothèse :

Démarche à suivre pour construire un test d'hypothèse

- Étape 1 : Formuler les hypothèses du test et donner le seuil de signification.
- Étape 2 : Vérifier les conditions d'application et représenter le test sur la distribution d'échantillonnage de \bar{X} .
- Étape 3 : Déterminer le ou les points critiques.
- Étape 4 : Énoncer la règle de décision du test.
- Étape 5 : Prendre une décision, puis tirer une conclusion en fonction de l'hypothèse énoncée par le chercheur.

EXEMPLE 1

Une machine, précise à 5 ml près, remplit des contenants de sirop d'érable. Le producteur contrôle régulièrement le réglage de la machine, par échantillonnage, afin de s'assurer que le volume de sirop dans les contenants est bien de 540 ml en moyenne, comme l'indique l'étiquette.

- a) Construire une règle de décision, au seuil de signification $\alpha = 0,05$, permettant de vérifier régulièrement le réglage de la machine en utilisant des échantillons de 100 contenants de sirop. Utiliser la précision de la machine comme écart type de la production.

Solution

- b) Quels sont les risques de faire ajuster inutilement la machine avec cette règle de décision ?

- c) Pour un échantillon de 100 contenants de sirop d'érable prélevés au hasard dans la production, on obtient une moyenne de 541,5 ml. L'écart de 1,5 ml seulement entre \bar{x} et μ indique-t-il un mauvais réglage de la machine au seuil de 0,05 ?

- d) Sur la base de la règle de décision établie, donner un exemple de moyenne d'échantillon qui donnerait à penser que le volume moyen de sirop d'érable par contenant est :

- inférieur à 540 ml : $\bar{x} =$ _____
- conforme à la moyenne de 540 ml : $\bar{x} =$ _____

NOTE

Il ne faut pas confondre les points critiques d'un test d'hypothèse bilatéral avec les bornes d'un intervalle de confiance de l'estimation d'une moyenne μ . Dans un test d'hypothèse, la moyenne de la population est supposée connue, et les points critiques sont donnés par : $c_1 = \mu - E$ et $c_2 = \mu + E$, où E est l'écart maximal tolérable. Dans l'estimation d'une moyenne, μ est inconnu, et les bornes de l'intervalle de confiance sont données par : borne inférieure = $\bar{x} - E$ et borne supérieure = $\bar{x} + E$, où E est la marge d'erreur (*voir la représentation graphique d'un intervalle de confiance à la page 206*).

RAPPEL

- Quand l'écart type σ de la population est inconnu, on utilise l'écart type corrigé s de l'échantillon comme estimateur de σ .
- Pour un échantillon de petite taille ($n < 30$) et une population normale d'écart type σ inconnu, on a $E = t\sigma_{\bar{x}}$, où t suit la loi de Student avec $dl = n - 1$.

EXEMPLE 2

On veut tester la durée, en kilomètres, d'une nouvelle semelle de pneus de voiture. Une analyse échantillonnale de 12 pneus a donné une durée moyenne de 53 870 km avec un écart type corrigé de 7 760 km. Au seuil de signification de 0,01, peut-on dire que la nouvelle semelle améliore la durée moyenne des pneus actuels, si celle-ci suit une distribution normale dont la moyenne est de 50 000 km ?

Solution

Interprétation de la décision

L'écart de 3 870 km entre \bar{x} et μ n'est pas assez grand statistiquement pour qu'on rejette l'hypothèse nulle à un seuil de 0,01 : il est probablement attribuable à la variation d'échantillonnage causée par le hasard.

NOTE

Il est important de bien comprendre qu'un test d'hypothèse ne peut pas servir à prouver que l'hypothèse nulle est vraie. Il conduit tout au plus au non-rejet de l'hypothèse nulle, faute d'évidence statistique contredisant celle-ci. Seule une étude complète de la population permettrait d'affirmer que l'hypothèse nulle est vraie ou fausse. Par conséquent, **lorsqu'on affirme que l'hypothèse nulle n'est pas rejetée, cela signifie qu'aucune évidence statistique ne permet de la rejeter.**

Analogie avec un procès (suite)

Il y a également une analogie avec un procès dans le non-rejet de l'hypothèse nulle. Le fait de reconnaître l'accusé non coupable ne prouve pas qu'il soit innocent : on affirme seulement que les preuves ne sont pas suffisantes pour le condamner. Cela explique pourquoi l'on utilise l'expression « la Cour déclare l'accusé non coupable » et non l'expression « la Cour déclare l'accusé innocent ».

EXERCICE DE COMPRÉHENSION | 5.1

En 2005, les adolescents québécois âgés de 12 à 17 ans consacrent en moyenne 9,5 heures par semaine à l'écoute de la radio. En 2012, un chercheur émet l'hypothèse que l'arrivée des baladeurs numériques et des services de musique en ligne a entraîné une diminution du temps d'écoute de la radio chez les adolescents. Pour vérifier cette hypothèse, il prélève un échantillon aléatoire de 64 adolescents et il établit que ces derniers consacrent en moyenne 8,3 heures par semaine à l'écoute de la radio, avec un écart type corrigé de 4,2 heures.

Sources: Données BBM extraites des éditions 2006 et 2014 du *Guide annuel des médias* publié par Infopresse.

- a) La moyenne échantillonnale confirme-t-elle l'hypothèse du chercheur ? Effectuer un test d'hypothèse au seuil de signification de 0,05.

Solution

1. Hypothèses et seuil de signification 2. Condition d'application et représentation graphique
On a _____, donc \bar{X} suit une normale.

3. Point critique

- Écart maximal tolérable : $E =$
- Point critique : $c =$

4. Règle de décision

5. Décision et conclusion

- b) À combien peut-on estimer les risques que la conclusion du test soit fausse ? _____
c) Quelle hypothèse est testée par un test d'hypothèse : H_0 ou H_1 ? _____
d) C'est l'hypothèse _____ qui indique le type de test (unilatéral à droite ou à gauche, ou bilatéral) qu'il faut effectuer.
e) C'est l'hypothèse _____ qui donne la moyenne de la courbe normale utilisée pour représenter la distribution d'échantillonnage de \bar{X} .

EXERCICES 5.1

1. On veut tester l'hypothèse nulle $H_0: \mu = 100$. Pour ce faire, on prélève un échantillon aléatoire de taille 36 dans la population. Supposons que l'on obtienne l'une des 3 valeurs ci-dessous comme moyenne échantillonnale. En sachant que $\sigma = 12$ et en considérant la position de chacune des valeurs sur la distribution d'échantillonnage de \bar{X} , indiquer la valeur qui nécessiterait la construction d'une règle de décision, donc d'un test d'hypothèse, pour décider du rejet ou du non-rejet de l'hypothèse nulle. Justifier votre choix.

$$\bar{x} = 93,3$$

$$\bar{x} = 101,3$$

$$\bar{x} = 104,6$$

2. a) Quelle hypothèse est testée par un test d'hypothèse : H_0 ou H_1 ?
- b) Lequel des trois énoncés suivants définit le seuil de signification ?
- Le point à partir duquel on décide de rejeter H_0 .
 - La zone de rejet de H_0 : si \bar{x} est dans cette région, on doit rejeter H_0 .
 - La probabilité de rejeter l'hypothèse H_0 alors que cette hypothèse est vraie.
- c) Lorsqu'on prend la décision de ne pas rejeter l'hypothèse H_0 , cela constitue une preuve que cette hypothèse est vraie. Commenter cette affirmation.
3. Formuler les hypothèses H_0 et H_1 pour les situations suivantes :
- a) Un fabricant examine un échantillon de 30 bouteilles remplies par une machine afin de vérifier si celle-ci verse bien, en moyenne, 500 ml de jus par bouteille.
- b) Un chercheur émet l'hypothèse que la durée de séjour des touristes dans les hôtels a augmenté à Québec en 2013 par rapport à 2011, où l'on avait observé une moyenne de 2,7 nuitées par personne.
- c) La longueur moyenne d'une tige métallique fabriquée par une machine doit être de 35 mm ; on veut vérifier le réglage de la machine.
- d) Une machine produit des articles dont le diamètre doit être de 6,25 cm. Si le diamètre moyen d'un lot est inférieur à 6,25 cm, le lot doit être détruit. Par contre, si le diamètre moyen est supérieur à 6,25 cm, les articles pourront être vendus au même prix, mais pour un usage différent. On veut vérifier le diamètre moyen des articles.
- e) Un chercheur émet l'hypothèse que l'absentéisme des femmes au travail est moindre quand il y a une garderie sur les lieux du travail. En moyenne,

le nombre de jours d'absence des travailleuses du Québec est de 4,4 journées par année.

4. Une machine remplit des sacs de sucre de façon que le poids de ceux-ci est, en moyenne, de 5 kg avec un écart type de 0,18 kg.

a) On prélève régulièrement un échantillon aléatoire de 50 sacs de sucre dans la production afin de surveiller le réglage de la machine. Construire une règle de décision, précise au centième près, permettant de s'assurer que les sacs contiennent bien, en moyenne, 5 kg de sucre. Utiliser un seuil de signification de 0,05.

b) Les tableaux suivants donnent, pour les 6 derniers échantillons prélevés, le poids moyen des 50 sacs de sucre de chaque échantillon. Y a-t-il un échantillon qui indique que la machine était mal ajustée au moment du prélèvement ?

Lundi		
10 h	13 h	16 h
$\bar{x} = 5,06$ kg	$\bar{x} = 4,98$ kg	$\bar{x} = 4,97$ kg

Mardi		
10 h	13 h	16 h
$\bar{x} = 5,02$ kg	$\bar{x} = 4,96$ kg	$\bar{x} = 4,94$ kg

c) Expliquer ce que signifie un seuil de signification de 0,05 dans le contexte du problème.

5. Le responsable du procédé de fabrication d'une entreprise suggère d'employer un nouvel alliage pour la production de tiges en acier. Il pense ainsi améliorer la résistance moyenne à la rupture, qui est actuellement de 50 kg/cm². Après avoir décidé d'utiliser le nouvel alliage, on désire vérifier si les objectifs sont atteints. On prélève un échantillon aléatoire de 40 tiges dans la production. La résistance moyenne à la rupture de ces tiges est de 54,5 kg/cm² avec un écart type corrigé de 2,4 kg/cm². Est-ce que l'écart observé dans la résistance moyenne à la rupture avant et après l'introduction du nouvel alliage est suffisamment élevé pour conclure, au seuil de signification de 1 %, qu'il y a augmentation significative de la résistance moyenne à la rupture ?

6. Afin d'améliorer le service à la clientèle, une entreprise a informatisé la gestion des stocks. Avant l'informatisation, le temps nécessaire pour répondre à la demande d'un client suivait une loi normale dont la moyenne était de 8,3 minutes et l'écart type, de 3,2 minutes. À la suite de l'informatisation, un

échantillon aléatoire de 25 clients a donné les temps de service suivants, en minutes :

7	9	6	6	3
6	5	7	7	8
10	9	4	3	6
5	7	8	8	3
4	4	6	5	4

Peut-on conclure, au seuil de signification de 5 %, que l'informatisation a permis d'accélérer le service à la clientèle ?

7. En 2000, une étude dresse un portrait statistique de l'adoption internationale au Québec. On y apprend que 7 900 enfants ont été adoptés par des familles québécoises entre 1990 et 1999. La moyenne d'âge des enfants au moment de l'adoption est de 23,3 mois. Depuis ce temps, une nouvelle convention internationale visant à mieux protéger les enfants semble avoir entraîné une augmentation de cette moyenne d'âge. Cette hypothèse est-elle confirmée si un échantillon aléatoire de 48 enfants prélevé parmi les 339 enfants adoptés en 2011 indique une moyenne d'âge de 30,8 mois, avec un écart type corrigé de 29,2 mois ? Utiliser un seuil de signification de 0,05.

Sources: Ministère de la Famille. *Un portrait statistique des familles au Québec – Édition 2005*; Secrétariat à l'adoption internationale du Québec. *Statistiques 2011*.

8. On estime qu'il faut en moyenne 10 minutes pour remplir un formulaire gouvernemental. Peut-on conclure, au seuil de signification de 0,05, qu'il faut en moyenne plus de temps pour remplir le formulaire si un échantillon aléatoire de 25 personnes ont pris les nombres de minutes suivants pour le faire ? On considère que le nombre de minutes requis est distribué normalement.

9,1	11,3	11,9	10,9	11,9
11,6	10,3	11,8	11,0	10,6
12,3	9,6	11,9	10,8	12,9
10,3	10,1	12,2	11,1	10,3
10,0	10,5	9,8	11,8	12,0

9. Dans le but de réduire les pertes en vies et en matériel causées par les incendies, le ministère de la Sécurité publique a fixé à 5 minutes le temps moyen qu'une première équipe de 4 pompiers doit prendre pour se rendre sur le lieu d'une intervention en milieu urbain. Au seuil de signification de 0,05, peut-on considérer que la Ville de Québec dépasse la norme gouvernementale, si l'on a obtenu la distribution suivante pour un échantillon aléatoire de 20 interventions ? On considère que la distribution du temps d'intervention suit une loi normale.

Répartition de 20 interventions selon le temps pris par la première équipe de pompiers pour se rendre sur les lieux

Temps (en min)	[1 ; 3[[3 ; 5[[5 ; 7[[7 ; 9[Total
Nombre	2	7	7	4	20

Source: Service de protection contre l'incendie de la Ville de Québec (SPCIQ). *Schéma de couverture de risques en incendie 2012-2017*, août 2011.

10. La durée de vie moyenne des tubes fluorescents fabriqués par une entreprise est estimée à 1 000 heures. Les techniciens tentent d'améliorer la durée de vie en modifiant la composition du gaz. Un test préliminaire montre que, pour un échantillon de 100 tubes fluorescents modifiés, la durée de vie moyenne est de 1 050 heures.
- Si l'écart type corrigé de l'échantillon est de 168 heures, au seuil de signification de 0,01, peut-on conclure que les néons modifiés durent plus longtemps ?
 - Estimer les risques que la durée de vie moyenne des tubes fluorescents soit toujours de 1 000 heures, et donc que la modification du gaz n'a eu aucun effet.
 - La conclusion serait-elle la même si la moyenne de l'échantillon était de 1 025 heures ? Peut-on conclure que cela prouve que la durée de vie moyenne réelle des tubes fluorescents produits est bien de 1 000 heures ?

5.2 Le test d'hypothèse sur un pourcentage

Dans la section précédente, nous avons appris à valider une hypothèse portant sur la moyenne μ d'une population à l'aide de la moyenne \bar{x} d'un échantillon. Dans la présente section, nous apprendrons à valider une hypothèse portant sur un pourcentage p d'une population en utilisant le pourcentage \hat{p} d'un échantillon. On donne le nom de test d'hypothèse sur un pourcentage à cette procédure.

Démarche à suivre pour construire un test d'hypothèse sur un pourcentage

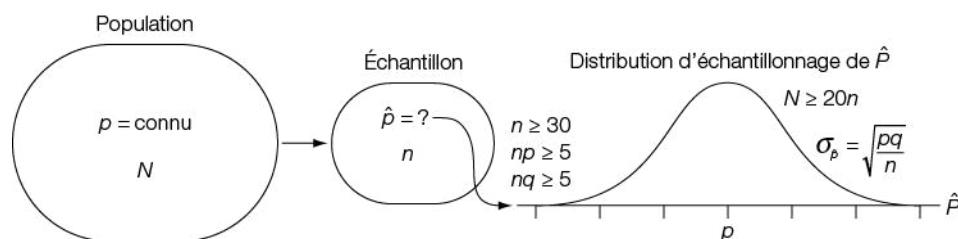
La démarche en cinq étapes suivie pour tester une hypothèse sur une moyenne s'applique, avec la même logique, à un test d'hypothèse sur un pourcentage :

- Étape 1 : Formuler les hypothèses du test et donner le seuil de signification.
- Étape 2 : Vérifier les conditions d'application et représenter le test sur la distribution d'échantillonnage de \hat{P} .
- Étape 3 : Déterminer le ou les points critiques.
- Étape 4 : Énoncer la règle de décision du test.
- Étape 5 : Prendre une décision, puis tirer une conclusion en fonction de l'hypothèse du chercheur.

Rappel des caractéristiques de la distribution d'échantillonnage de \hat{P}

Pour $n \geq 30$, $np \geq 5$ et $nq \geq 5$, la distribution d'échantillonnage de \hat{P} suit une loi normale dont la moyenne et l'écart type sont :

- $\mu_{\hat{P}} = p$
- $\sigma_{\hat{P}} = \begin{cases} \sqrt{\frac{pq}{n}} & \text{si la population est grande } (N \geq 20) \\ \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}} & \text{si la population est petite } (N < 20) \end{cases}$



EXEMPLE

Seulement 20 % des clients d'un magasin acquittent leurs achats par paiement direct. Le propriétaire du magasin organise une campagne de promotion afin d'inciter un plus grand nombre de clients à employer ce mode de paiement. Quelque temps après la fin de la campagne, on veut en vérifier l'efficacité. Dans un échantillon de 150 clients, 42 ont utilisé le paiement direct. Peut-on accepter l'hypothèse selon laquelle la campagne de promotion a été efficace ? Faire un test au seuil de signification de 0,01.

Solution

1. Hypothèses et seuil de signification

$$\begin{aligned} H_0: p &= 20 \% \\ H_1: p &> 20 \% \\ \alpha &= 0,01 \end{aligned}$$

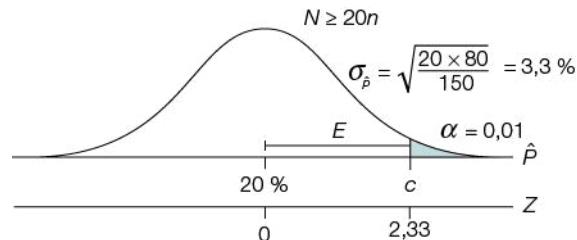
On a : $n = 150$

$$\hat{p} = \frac{42}{150} = 28 \%$$

2. Conditions d'application

On a $n \geq 30$, $np = 30 \geq 5$ et $nq = 120 \geq 5$, donc \hat{P} suit une normale.

Représentation graphique du test



3. Point critique

- Écart maximal tolérable : $E = z\sigma_{\hat{p}} = 2,33 \times 3,3 = 7,7 \%$
- Point critique : $c = 20 \% + 7,7 \% = 27,7 \%$

4. Règle de décision

Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est supérieur à 27,7 %.

5. Décision et conclusion

Comme $\hat{p} = 28 \% > 27,7 \%$, on rejette H_0 .

La campagne de promotion a été efficace : on constate une augmentation du pourcentage de clients qui utilisent le paiement direct.

EXERCICES DE COMPRÉHENSION | 5.2

1. Une machine effectue un mélange de bonbons de différentes couleurs. Pour s'assurer que le produit obtenu contient bien 20 % de bonbons rouges, on prélève régulièrement 500 bonbons au hasard, puis on calcule le pourcentage de bonbons rouges.

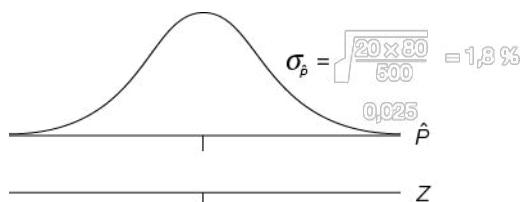
- a) Construire une règle de décision qui permettrait au responsable du contrôle de la qualité de s'assurer du bon fonctionnement de la machine au seuil de signification de 0,05.

Solution

1. Hypothèses et seuil de signification

2. Conditions d'application et représentation graphique

On a $n \geq 30$, $np = 100 \geq 5$ et $nq = 400 \geq 5$, donc \hat{P} suit une normale.



3. Points critiques

Écart maximal tolérable : $E =$

Points critiques : $c_1 =$

$c_2 =$

4. Règle de décision

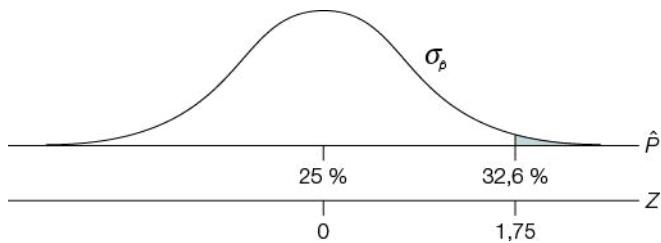
- b) Le tableau suivant donne le nombre de bonbons rouges que contenaient les huit derniers échantillons prélevés par le responsable du contrôle de la qualité. Compléter le tableau.

Numéro de l'échantillon	1	2	3	4	5	6	7	8
Nombre de rouges	105	102	88	80	84	99	120	116
Pourcentage \hat{p}			17,6 %	16,0 %	16,8 %	19,8 %	24,0 %	23,2 %

Quels sont les numéros des échantillons qui indiquent qu'au moment du prélèvement le mélange ne contenait pas 20 % de bonbons rouges ? Dans chaque cas, dire s'il y avait trop ou pas assez de bonbons rouges dans le mélange.

c) À combien peut-on estimer les risques que la conclusion précédente soit fausse ? _____

2. Voici la représentation graphique d'un test d'hypothèse sur un pourcentage avec un échantillon de taille $n = 100$.



a) D'après la représentation graphique :

- i) la valeur du pourcentage p de la population est _____.
- ii) les hypothèses du test sont H_0 : _____ et H_1 : _____.
- iii) le seuil de signification est _____.

b) Quelle est la valeur de l'écart type de cette distribution normale ?

c) Énoncer la règle de décision du test.

d) Donner deux exemples de valeurs de \hat{p} qui permettraient :

- de rejeter H_0 : $\hat{p} = \underline{\hspace{2cm}}$ ou $\hat{p} = \underline{\hspace{2cm}}$.
- de ne pas rejeter H_0 : $\hat{p} = \underline{\hspace{2cm}}$ ou $\hat{p} = \underline{\hspace{2cm}}$.

EXERCICES 5.2

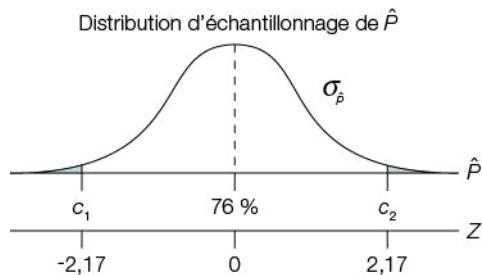
1. Lors d'une discussion, un sociologue affirme que 80 %¹ des Québécois ne seraient pas inquiets à l'idée qu'une personne homosexuelle enseigne dans une école primaire. Vous croyez que ce pourcentage est grandement exagéré. Pour le prouver, vous effectuez un sondage auprès d'un échantillon aléatoire de 570 Québécois.

a) Faut-il rejeter l'affirmation du sociologue, au seuil de signification de 5 %, si 448 répondants disent qu'ils ne seraient pas inquiets à l'idée qu'une personne homosexuelle enseigne dans une école primaire ?

b) La conclusion du test effectué en a) prouve-t-elle que le sociologue a raison ?

1. Pourcentage basé sur un sondage CROP publié dans la revue *L'actualité* en mai 2007.

2. Voici la représentation graphique d'un test d'hypothèse sur un pourcentage.



a) Donner les deux hypothèses du test.

b) Quelle hypothèse est considérée comme vraie jusqu'à preuve du contraire dans un test d'hypothèse ?

- c) Quels sont les risques de se tromper si, dans la conclusion du test, on décide de rejeter H_0 ?
- d) Le pourcentage de 76 % est-il celui de la population ou de l'échantillon?
- e) Sachant que la taille de l'échantillon prélevé pour effectuer ce test est 1 300, calculer la valeur de l'écart type $\sigma_{\hat{p}}$.
- f) Énoncer la règle de décision de ce test.
- g) Donner deux exemples de pourcentages d'échantillon qui permettraient de rejeter H_0 .

3. TROP TARD POUR CHANGER SON VOTE

Trois semaines après le référendum du 30 octobre 1995 sur la souveraineté du Québec, où l'option OUI a perdu avec un pourcentage de 49,4 %, la firme de sondage Léger et Léger a publié un sondage indiquant que 54,8 % des 1 003 personnes de l'échantillon répondraient «Oui» si la question du référendum leur était de nouveau posée. Au seuil de signification de 0,05, peut-on considérer que l'écart entre le résultat du sondage et celui du référendum est assez significatif statistiquement pour affirmer que le pourcentage de Québécois en accord avec la souveraineté a augmenté depuis le référendum?

Source: *Le Journal de Montréal*, 24 novembre 1995.

4. Selon une étude, seulement 65 % des jeunes du secondaire sont conscients des dangers de la divulgation de renseignements personnels sur Internet. Des enseignants décident de mener une campagne de sensibilisation à ce sujet auprès des 1 500 étudiants de la commission scolaire. On mesure l'efficacité de la campagne par un sondage mené auprès d'un échantillon aléatoire de 200 étudiants. Doit-on conclure, au seuil de signification de 5 %, que la campagne a été efficace, si 136 étudiants considèrent qu'il est dangereux de dévoiler de l'information personnelle dans Internet?

Source: HabiloMédias. *Jeunes Canadiens dans un monde branché, phase III. Vie privée en ligne, promotion en ligne*, 2014.

5. Dans le cadre du reboisement des forêts, le gouvernement décide d'expérimenter une nouvelle variété de plants d'épinette pour laquelle le producteur assure un taux de survie à la transplantation de 80 %. Plusieurs milliers de ces plants sont transplantés dans différentes régions du Québec, car on pense que le taux de survie sera différent d'une région à l'autre. Un an plus tard, un échantillon aléatoire de 1 200 épinettes est prélevé dans chaque région parmi les plants transplantés.

- a) Construire une règle de décision permettant de tester l'hypothèse émise par le gouvernement au seuil de signification de 1 %.
- b) Voici, pour quatre régions, le nombre d'arbres de l'échantillon qui étaient encore vivants un an après la transplantation.

Abitibi : 912	Saguenay : 1 020
Côte-Nord : 936	Gaspésie : 984

Y a-t-il une région où le taux de survie a été différent du 80 % prévu? Si oui, le taux de survie semble-t-il plus élevé ou moins élevé que le taux prévu?

6. Une étude sur la compétence des adultes en compréhension de texte révèle que 14,6 % des Canadiens âgés de 16 à 65 ans se situent à un très faible niveau de compétence. Peut-on soutenir l'hypothèse selon laquelle ce pourcentage est plus faible chez les jeunes âgés de 16 à 25 ans si, pour un échantillon aléatoire de 150 jeunes, on en dénombre 14 qui se situent à un très faible niveau de compétence en compréhension de texte? Utiliser un seuil de signification de 5 %.

Source: OCDE et Statistique Canada. *La littératie, un atout pour la vie: Nouveaux résultats de l'Enquête sur la littératie et les compétences des adultes*, 2011.

5.3 Le test d'hypothèse sur l'égalité de deux paramètres

De nombreuses études, tant en administration qu'en psychologie, cherchent à comparer deux populations différentes en utilisant les statistiques de deux échantillons aléatoires. Par exemple, on voudra comparer les internautes féminins et masculins pour vérifier si la proportion d'utilisateurs de médias sociaux y est différente et si les femmes consacrent, en moyenne, plus de temps que les hommes aux médias sociaux. Un test d'hypothèse portant sur l'égalité de deux pourcentages et sur l'égalité de deux moyennes permet de répondre à ce type de question.

Notations

Nous distinguerons les paramètres des populations étudiées en ayant recours aux indices 1 et 2. Il en sera de même pour les statistiques des échantillons prélevés.

Population 1	Population 2
Paramètres: $\mu_1, \sigma_1, p_1, N_1$	Paramètres: $\mu_2, \sigma_2, p_2, N_2$
Statistiques: $\bar{x}_1, s_1, \hat{p}_1, n_1$	Statistiques: $\bar{x}_2, s_2, \hat{p}_2, n_2$

5.3.1 Le test sur l'égalité de deux moyennes

La méthode pour effectuer un test d'hypothèse sur l'égalité de deux moyennes dépend du type des échantillons prélevés. Ceux-ci peuvent être indépendants ou dépendants (appariés). Nous présentons ci-dessous la marche à suivre pour chaque type d'échantillon.

Échantillons indépendants de grande taille ($n \geq 30$)

Deux échantillons sont indépendants lorsque les observations sont indépendantes dans chaque échantillon et entre les deux échantillons. On obtient ce résultat lorsque les échantillons sont prélevés au hasard dans chacune des populations étudiées.

Pour construire un test d'hypothèse sur l'égalité des moyennes μ_1 et μ_2 de deux populations différentes, la procédure suit la même logique que celle utilisée dans les sections précédentes : formulation des hypothèses nulle et alternative, puis établissement d'une règle de décision, reposant sur la distribution d'échantillonnage de $\bar{X}_1 - \bar{X}_2$, qui permet de décider du rejet ou du non-rejet de l'hypothèse nulle.

Hypothèses du test

Hypothèse nulle

$H_0: \mu_1 - \mu_2 = 0 \Leftrightarrow H_0: \mu_1 = \mu_2$ Il n'y a pas de différence entre les moyennes des populations.

Hypothèse alternative ou hypothèse du chercheur

$H_1: \mu_1 - \mu_2 \neq 0 \Leftrightarrow H_1: \mu_1 \neq \mu_2$ Les moyennes des populations sont différentes.

$H_1: \mu_1 - \mu_2 > 0 \Leftrightarrow H_1: \mu_1 > \mu_2$ La moyenne de la population 1 est supérieure à celle de la population 2.

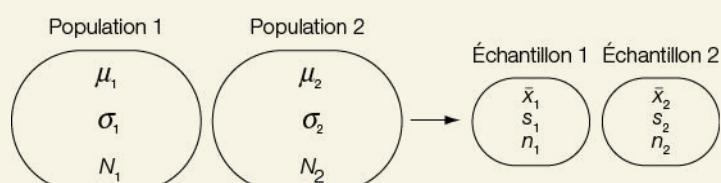
$H_1: \mu_1 - \mu_2 < 0 \Leftrightarrow H_1: \mu_1 < \mu_2$ La moyenne de la population 1 est inférieure à celle de la population 2.

Distribution d'échantillonnage de $\bar{X}_1 - \bar{X}_2$

On utilise la différence des moyennes échantillonnelles \bar{x}_1 et \bar{x}_2 pour tester l'égalité des moyennes μ_1 et μ_2 des populations. On construit la règle de décision du test en se basant sur la loi de probabilité qui s'applique à la différence entre les moyennes de deux échantillons indépendants.

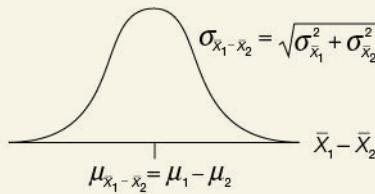
Distribution d'échantillonnage de la différence de deux moyennes

Soit une population 1 de moyenne μ_1 et d'écart type σ_1 , et une population 2 de moyenne μ_2 et d'écart type σ_2 . Si l'on prélève des échantillons aléatoires indépendants de taille n_1 dans la population 1 et de taille n_2 dans la population 2, alors la distribution d'échantillonnage de $\bar{X}_1 - \bar{X}_2$ présente les caractéristiques suivantes :



- Moyenne: $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$
- Écart type: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2}$
- Forme de la distribution:
Normale si $n_1 \geq 30$ et $n_2 \geq 30$ ou si les populations sont normales.

Distribution d'échantillonnage de $\bar{X}_1 - \bar{X}_2$



NOTE

Si les écarts types σ_1 et σ_2 sont inconnus, on utilise les écarts types corrigés s_1 et s_2 comme estimateurs.

EXEMPLE

Dans certaines succursales d'une chaîne de restauration rapide, on utilise un afficheur électronique pour présenter des photos de divers produits. Pour mesurer l'efficacité de ce type de promotion, on prélève un échantillon de 36 jours, et on compare la moyenne quotidienne des ventes du produit vedette de deux succursales, l'une disposant d'un afficheur électronique (A) et l'autre pas (A'). Pour ces 36 jours, la moyenne quotidienne des ventes a été de 170 unités avec un écart type corrigé de 6 unités pour le restaurant disposant d'un afficheur et de 165 unités avec un écart type corrigé de 5 unités pour l'autre restaurant. Au seuil de signification de 0,05, peut-on affirmer que la moyenne quotidienne des ventes du produit vedette est plus élevée pour la succursale utilisant un afficheur électronique ?

Solution

$$\begin{aligned} 1. \quad & H_0: \mu_A - \mu_{A'} = 0 \\ & H_1: \mu_A - \mu_{A'} > 0 \\ & \alpha = 0,05 \end{aligned}$$

On a :

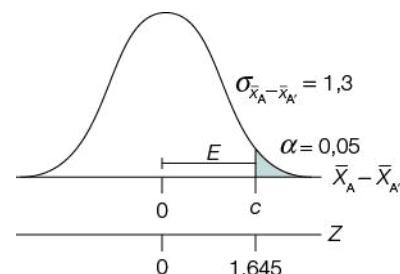
$$n_A = 36, \bar{x}_A = 170, s_A = 6$$

$$n_{A'} = 36, \bar{x}_{A'} = 165, s_{A'} = 5$$

$$\sigma_{\bar{x}_A} = \frac{\sigma_A}{\sqrt{n_A}} \approx \frac{s_A}{\sqrt{n_A}} = \frac{6}{\sqrt{36}} = 1,0$$

$$\sigma_{\bar{x}_{A'}} = \frac{\sigma_{A'}}{\sqrt{n_{A'}}} \approx \frac{s_{A'}}{\sqrt{n_{A'}}} = \frac{5}{\sqrt{36}} = 0,8$$

- $n_A \geq 30$ et $n_{A'} \geq 30$, donc $\bar{X}_A - \bar{X}_{A'}$ suit une normale avec:
 - $\mu_{\bar{x}_A - \bar{x}_{A'}} = \mu_A - \mu_{A'} = 0$
 - $\sigma_{\bar{x}_A - \bar{x}_{A'}} = \sqrt{\sigma_{\bar{x}_A}^2 + \sigma_{\bar{x}_{A'}}^2} = \sqrt{1,0^2 + 0,8^2} = 1,3$



3. Point critique

Écart maximal tolérable : $E = z\sigma_{\bar{x}_A - \bar{x}_{A'}} = 1,645 \times 1,3 = 2,1$ unités

Point critique : $c = 0 + 2,1 = 2,1$ unités

4. Règle de décision

Rejeter H_0 si la différence $\bar{x}_A - \bar{x}_{A'}$ est supérieure à 2,1 unités.

5. Décision et conclusion

Comme $\bar{x}_A - \bar{x}_{A'} = 5$ unités > 2,1 unités, on rejette H_0 .

L'utilisation d'un afficheur électronique est efficace pour faire augmenter les ventes des produits vedettes.

Échantillons dépendants ou appariés

Deux échantillons sont dépendants si les observations sont liées entre elles. C'est notamment le cas lorsque l'on veut mesurer l'effet d'un traitement : pour un même individu, on prend une mesure avant et après le traitement. L'échantillon est alors constitué de paires d'observations (x, y) , où x est la mesure avant le traitement et y , la mesure après le traitement.

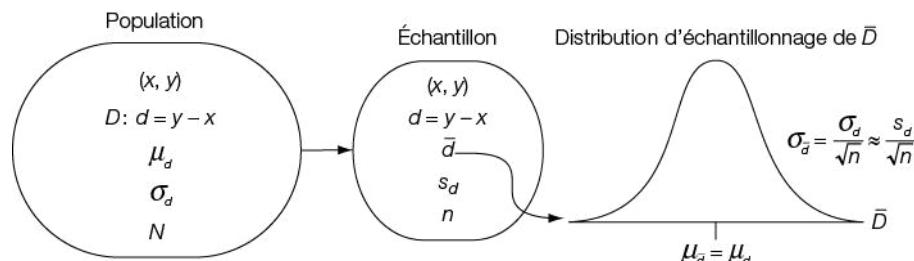
On obtient aussi des échantillons dépendants quand, pour les besoins d'une recherche, on constitue des échantillons appariés où chaque individu du groupe expérimental est associé à un individu du groupe témoin ayant des caractéristiques semblables. C'est particulièrement le cas lorsque l'on apparie de vrais jumeaux.

Variable aléatoire D

La variable aléatoire D , nommée différence, associe à chaque paire possible d'observations (x, y) la différence $d = y - x$. L'espérance de D se note μ_d et son écart type, σ_d .

Distribution d'échantillonnage de \bar{D}

Si, d'une population de taille N composée des différences d associées à des observations dépendantes (x, y) , dont la moyenne est μ_d et l'écart type, σ_d , on prélève un échantillon aléatoire de taille n , alors la distribution d'échantillonnage de la moyenne \bar{D} des différences d a les caractéristiques suivantes :



- Moyenne : $\mu_{\bar{d}} = \mu_d$
- Écart type : $\sigma_{\bar{d}} = \frac{\sigma_d}{\sqrt{n}} \approx \frac{s_d}{\sqrt{n}}$
- Forme de la distribution :
 - si $n \geq 30$, alors \bar{D} suit une loi normale ;
 - si $n < 30$ et population normale, alors $t = (\bar{d} - \mu_d)/\sigma_d$ suit une loi de Student avec $dl = n - 1$.

Hypothèses du test

Si l'on utilise la différence $d = y - x$ pour comparer les observations appariées (x, y) , un test sur l'égalité de deux moyennes devient alors un test sur une moyenne, la moyenne μ_d .

Hypothèse nulle

$$H_0: \mu_d = 0 \Leftrightarrow \mu_y = \mu_x$$

Hypothèse alternative ou hypothèse du chercheur

$$H_1: \mu_d \neq 0 \Leftrightarrow \mu_y \neq \mu_x \quad \text{ou} \quad H_1: \mu_d > 0 \Leftrightarrow \mu_y > \mu_x \quad \text{ou} \quad H_1: \mu_d < 0 \Leftrightarrow \mu_y < \mu_x$$

EXEMPLE

Dans le cadre d'une étude sur le fonctionnement du système nerveux, on a voulu savoir si l'écoute de la musique modifie le temps de réaction à un stimulus visuel chez les adolescents. Pour ce faire, on a choisi 8 adolescents au hasard et l'on a mesuré leur temps de réaction à l'apparition d'une image sur un écran. Deux séries de mesures ont été effectuées : une série dans le calme et une autre alors que le sujet écoutait de la musique douce avec un casque. La forme visualisée était un carré bleu et son nombre d'apparitions était fixé à 20. Le tableau ci-dessous donne le temps de réaction moyen des adolescents, en millisecondes (ms). Au seuil de signification de 0,05, peut-on conclure que le temps de réaction chez les adolescents est influencé par l'écoute de la musique ? (On suppose que la variable aléatoire D de la différence entre y et x suit une loi normale.)

Temps de réaction (en ms)

Sujet	1	2	3	4	5	6	7	8
Sans musique (x)	319	262	293	374	270	265	261	303
Avec musique (y)	299	256	312	357	277	279	253	286
$d = y - x$	-20	-6	19	-17	7	14	-8	-17

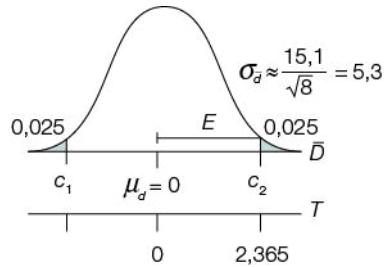
Solution

On effectue un test bilatéral, car on ignore si l'écoute de la musique causera une augmentation ou une diminution du temps moyen de réaction au stimulus visuel chez les adolescents.

1. $H_0: \mu_d = 0$ Le temps de réaction est le même.
 $H_1: \mu_d \neq 0$ Le temps de réaction est différent.
 $\alpha = 0,05$
2. $n < 30$ et D suit une normale, alors
 $E = t\sigma_{\bar{d}}$ où t suit une Student avec $dl = 8 - 1 = 7$.

On a $\bar{d} = -3,5$ et $s_d = 15,1$.

- $\mu_{\bar{d}} = \mu_d = 0$
- $\sigma_{\bar{d}} = \frac{\sigma_d}{\sqrt{n}} \approx \frac{s_d}{\sqrt{n}} = \frac{15,1}{\sqrt{8}} = 5,3$



3. Points critiques

Écart maximal tolérable : $E = t\sigma_{\bar{d}} = 2,365 \times 5,3 = 12,5$ ms

Points critiques : $c_1 = 0 - 12,5 = -12,5$ ms

$$c_2 = 0 + 12,5 = 12,5 \text{ ms}$$

4. Règle de décision

Rejeter H_0 si la moyenne \bar{d} des différences de temps de réaction observées dans l'échantillon est inférieure à -12,5 ms ou supérieure à 12,5 ms.

5. Décision et conclusion

On a $\bar{d} = -3,5$ ms. Comme $-12,5 \text{ ms} < \bar{d} < 12,5 \text{ ms}$, on ne rejette pas H_0 .

Au seuil de signification de 0,05, les données échantillonnelles ne permettent pas de conclure que le temps de réaction à un stimulus visuel chez les adolescents est modifié par l'écoute de la musique.

5.3.2 Le test sur l'égalité de deux pourcentages

La démarche pour construire un test d'hypothèse sur l'égalité de deux pourcentages est semblable à celle qui sert à tester l'égalité de deux moyennes avec des échantillons indépendants. La distinction réside dans la formulation des hypothèses et dans les caractéristiques de la distribution d'échantillonnage de la différence des pourcentages de deux échantillons indépendants.

Hypothèses du test

Hypothèse nulle

$H_0: p_1 - p_2 = 0 \Leftrightarrow H_0: p_1 = p_2$ Il n'y a pas de différence entre les pourcentages des deux populations.

Hypothèse rivale ou alternative

$H_1: p_1 - p_2 \neq 0 \Leftrightarrow H_1: p_1 \neq p_2$ Les pourcentages des populations sont différents.

$H_1: p_1 - p_2 > 0 \Leftrightarrow H_1: p_1 > p_2$ Le pourcentage de la population 1 est supérieur à celui de la population 2.

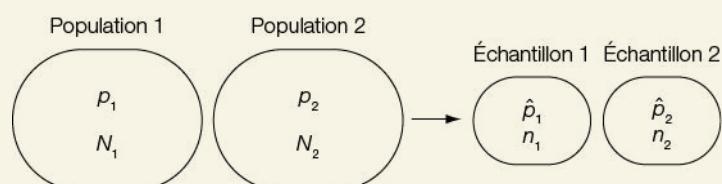
$H_1: p_1 - p_2 < 0 \Leftrightarrow H_1: p_1 < p_2$ Le pourcentage de la population 1 est inférieur à celui de la population 2.

Distribution d'échantillonnage de $\hat{P}_1 - \hat{P}_2$

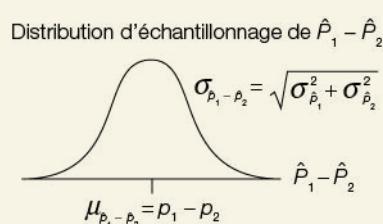
La différence entre les pourcentages échantillonaux \hat{p}_1 et \hat{p}_2 permet de juger de l'égalité des pourcentages respectifs p_1 et p_2 des deux populations. Voici les caractéristiques de la distribution d'échantillonnage de $\hat{P}_1 - \hat{P}_2$.

Distribution d'échantillonnage de la différence de deux pourcentages

Soit une population 1 de pourcentage p_1 et une population 2 de pourcentage p_2 . Si l'on prélève des échantillons aléatoires indépendants de taille n_1 dans la population 1 et de taille n_2 dans la population 2, alors la distribution d'échantillonnage de $\hat{P}_1 - \hat{P}_2$ présente les caractéristiques suivantes:



- Moyenne: $\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$
- Écart type: $\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\sigma_{\hat{P}_1}^2 + \sigma_{\hat{P}_2}^2}$
- Forme de la distribution:
Normale si $n_1 \geq 30$, $n_1\hat{p}_1 \geq 5$, $n_1\hat{q}_1 \geq 5$ et
 $n_2 \geq 30$, $n_2\hat{p}_2 \geq 5$, $n_2\hat{q}_2 \geq 5$



Estimation du pourcentage commun p sous l'hypothèse $H_0: p_1 = p_2$

Pour calculer l'écart type de la distribution d'échantillonnage de $\hat{P}_1 - \hat{P}_2$, il faut connaître les pourcentages p_1 et p_2 des deux populations étudiées ; or, dans un test d'hypothèse sur l'égalité de deux pourcentages, ces pourcentages sont inconnus. Toutefois, si l'on suppose que l'hypothèse nulle est vraie, soit que $p_1 - p_2 = 0$, alors p_1 et p_2 ont une valeur commune, notée p : on a donc $p_1 = p_2 = p$. On estime ponctuellement la valeur de p par le pourcentage \hat{p} du nombre de succès obtenus pour l'ensemble des deux échantillons prélevés.

$$p \approx \hat{p} = \frac{\text{nombre total de succès pour les deux échantillons}}{n_1 + n_2} \times 100 \%$$

EXEMPLE

Une étude menée auprès d'un échantillon de 450 hommes et 500 femmes indique que 17 % des hommes et 13 % des femmes dorment moins de 6,5 heures par nuit. Au seuil de signification de 0,05, peut-on en conclure que le pourcentage de personnes qui dorment moins de 6,5 heures par nuit est plus élevé chez les hommes que chez les femmes ?

Source: Statistique Canada. *Enquête sociale générale, 2006.*

Solution

$$\begin{aligned} 1. \quad & H_0: p_H - p_F = 0 \\ & H_1: p_H - p_F > 0 \\ & \alpha = 0,05 \end{aligned}$$

On a :

$$n_H = 450 \quad \hat{p}_H = 17 \%$$

$$n_F = 500 \quad \hat{p}_F = 13 \%$$

$$\begin{aligned} \bullet \quad p \approx \hat{p} &= \frac{17\% \times 450 + 13\% \times 500}{450 + 500} \\ &= \frac{76,5 + 65}{950} = 14,9 \% \end{aligned}$$

$$\bullet \quad \sigma_{\hat{p}_H} \approx \sqrt{\frac{\hat{p}\hat{q}}{n_H}} = \sqrt{\frac{14,9\% \times 85,1\%}{450}} = 1,7 \%$$

$$\bullet \quad \sigma_{\hat{p}_F} \approx \sqrt{\frac{\hat{p}\hat{q}}{n_F}} = \sqrt{\frac{14,9\% \times 85,1\%}{500}} = 1,6 \%$$

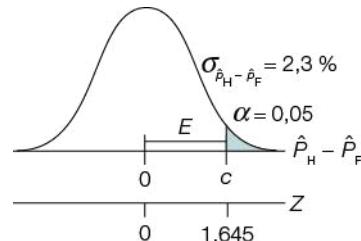
$$2. \quad n_H \geq 30 \quad n_H \hat{p}_H = 76,5 \geq 5 \quad n_H \hat{q}_H = 373,5 \geq 5$$

$$n_F \geq 30 \quad n_F \hat{p}_F = 65 \geq 5 \quad n_F \hat{q}_F = 435 \geq 5$$

$\Rightarrow \hat{P}_H - \hat{P}_F$ suit une normale avec :

$$\bullet \quad \mu_{\hat{p}_H - \hat{p}_F} = p_H - p_F = 0$$

$$\bullet \quad \sigma_{\hat{p}_H - \hat{p}_F} = \sqrt{\sigma_{\hat{p}_H}^2 + \sigma_{\hat{p}_F}^2} = \sqrt{1,7^2 + 1,6^2} = 2,3 \%$$



3. Point critique

- Écart maximal tolérable : $E = z\sigma_{\hat{p}_H - \hat{p}_F} = 1,645 \times 2,3 = 3,8 \%$
- Point critique : $c = 0 + 3,8 \% = 3,8 \%$

4. Règle de décision

Rejeter H_0 si la différence $\hat{p}_H - \hat{p}_F$ est supérieure à 3,8 %.

5. Décision et conclusion

Comme $\hat{p}_H - \hat{p}_F = 17\% - 13\% = 4\% > 3,8\%$, on rejette H_0 .

Au seuil de signification de 0,05, il y a un pourcentage plus élevé d'hommes que de femmes qui dorment moins de 6,5 heures par nuit.

EXERCICES 5.3

1. Une étude auprès de 120 personnes ayant été mariées et de 100 personnes ayant vécu en union libre indique qu'au moment de la rupture du couple, chez les personnes mariées, l'âge moyen était de 35,6 ans avec un écart type corrigé de 12,4 ans alors qu'il était de 30,5 ans avec un écart type de 10,4 ans chez les personnes vivant en union libre. Au seuil de signification de 0,01, peut-on affirmer qu'au moment de la rupture du couple, les personnes vivant en union libre sont en moyenne plus jeunes que celles qui sont mariées ?

Source: Statistique Canada. *Enquête sociale générale*, 2006.

2. Peut-on affirmer que les femmes dorment en moyenne plus longtemps que les hommes si, dans un échantillon aléatoire de 375 femmes, la moyenne d'heures de sommeil est de 8,2 heures avec un écart type corrigé de 1,5 heure alors que, dans un échantillon de 360 hommes, la moyenne est de 8,0 heures avec un écart type de 1,6 heure ? Effectuer un test à un seuil de signification de 5 %.

Source: Statistique Canada. *Enquête sociale générale*, 2006.

3. Les statistiques suivantes permettent-elles d'affirmer que les filles arrivent à l'école généralement mieux préparées que les garçons pour répondre aux exigences et aux attentes du milieu scolaire ?
 À la fin de la maternelle, on a fait passer un test (jeux du Lollipop) visant à mesurer le niveau de préparation à l'école. Les 600 filles de l'échantillon prélevé ont obtenu une note moyenne de 58,7 avec un écart type corrigé de 9,7 et les 590 garçons, une note moyenne de 56,0 avec un écart type corrigé de 9,8. Effectuer un test à un seuil de signification de 0,05.

Source: Institut de la statistique du Québec. *Étude longitudinale du développement des enfants du Québec*.

4. Un échantillon aléatoire de 6 personnes est constitué pour valider un programme d'activités visant à stimuler la mémoire des personnes âgées. Le tableau suivant présente les résultats du test de mémoire passé avant et après le programme d'activités. Au seuil de signification de 0,05, le programme d'activités permet-il d'augmenter la mémoire des personnes âgées ? La distribution de la différence des notes suit une normale.

Résultats au test de mémoire

Sujet	1	2	3	4	5	6
Avant	60	65	56	78	70	55
Après	65	72	53	82	76	60

5. Un psychologue a mis sur pied des ateliers pour aider les cégepiens à gérer le stress. Le niveau de stress est mesuré par un test psychologique au début et à la fin de la formation. On donne ci-dessous les notes de 6 étudiants prélevés au hasard parmi ceux qui ont suivi la formation. Cet échantillon permet-il de conclure, au seuil de signification de 0,01, que les ateliers font diminuer le niveau de stress des étudiants ? La distribution des différences des notes suit une normale.

Notes au test de niveau de stress

Sujet	1	2	3	4	5	6
Avant	28	24	26	20	30	22
Après	21	19	17	18	20	13

6. Des graines de fleurs ont été recouvertes d'une substance naturelle afin de faciliter la germination. Pour tester l'efficacité du produit, on choisit un échantillon de 5 terrains dans des endroits ayant des conditions de cultures différentes. Afin de comparer les deux types de graines dans les mêmes conditions de culture, on divise chaque terrain en 2 parties : la zone expérimentale, où l'on sème des graines enrobées, et la zone témoin, où l'on sème des graines non enrobées. Au seuil de signification de 1 %, peut-on dire que les graines enrobées germent plus rapidement que les graines non enrobées ? La distribution de la différence du temps de germination suit une normale.

Nombre moyen de jours de germination

Terrain	1	2	3	4	5
Non enrobées	20,2	22,3	23,2	18,6	21,4
Enrobées	16,1	18,8	15,2	13,9	16,5

7. Une enquête menée auprès d'un échantillon d'enfants québécois de 4 ans indique que 30,4 % des 780 garçons de l'échantillon écoutent fréquemment la télévision pendant les repas alors que ce pourcentage n'est que de 24,8 % chez les 770 filles. Au seuil de signification de 0,01, ces statistiques échantillonnelles permettent-elles d'affirmer que la proportion de jeunes de 4 ans qui écoutent fréquemment la télévision pendant les repas est plus élevée chez les garçons que chez les filles ?

Source: Institut de la statistique du Québec. *Enquête de nutrition auprès des enfants québécois de 4 ans*, octobre 2005.

8. On émet l'hypothèse que les élèves de 6^e année qui lisent une heure et plus par semaine pour le plaisir sont proportionnellement plus nombreux à réussir l'épreuve obligatoire de mathématique que ceux qui lisent moins. Cette hypothèse est-elle validée si, pour un échantillon aléatoire de 800 élèves, dont 300 lisent une heure et plus par semaine et 500 lisent moins d'une heure, 81 % ont réussi l'épreuve de mathématique dans le 1^{er} groupe et 73 % dans le 2^e groupe ? Utiliser un seuil de signification de 5 %.

Source: Institut de la statistique du Québec. *Les facteurs liés à la réussite à l'épreuve obligatoire de mathématique en sixième année du primaire*, 2010.

9. Un échantillon aléatoire de 1 000 étudiants à temps plein âgés de 15 à 24 ans comprend 600 hommes et

400 femmes. La proportion d'étudiants ayant un emploi est de 37 % chez les hommes et de 46 % chez les femmes. Chez les 222 hommes et 184 femmes qui ont un emploi, la moyenne d'heures de travail par semaine est de 15,7 heures avec un écart type corrigé de 2,5 heures pour les hommes, et de 14,6 heures avec un écart type corrigé de 2,4 heures pour les femmes.

Source: Institut de la statistique du Québec. *Regard statistique sur la jeunesse. État et évolution de la situation des Québécois âgés de 15 à 29 ans*, 2014.

Tester les hypothèses suivantes au seuil de signification de 0,01.

- a) Le taux d'emploi des étudiants est inférieur à celui des étudiantes.
- b) La moyenne d'heures de travail par semaine des étudiants est supérieure à celle des étudiantes.

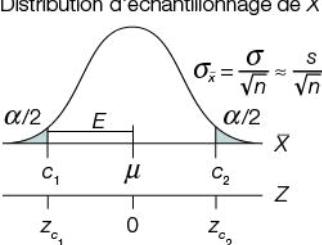
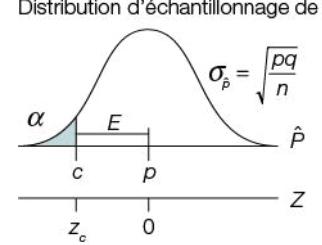
RÉSUMÉ DU CHAPITRE 5

Démarche pour construire un test d'hypothèse

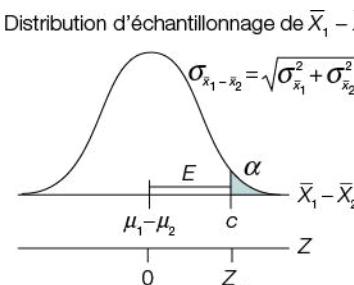
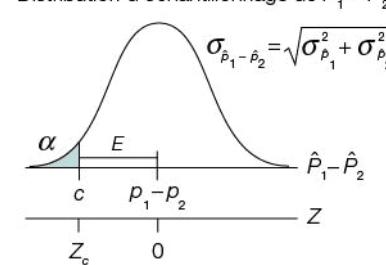
Étape 1 Formuler les hypothèses H_0 et H_1 et préciser le seuil de signification.

Étape 2 Vérifier les conditions d'application et représenter le test sur la distribution d'échantillonnage.

Représentation d'un test sur un paramètre

Test sur une moyenne	Test sur un pourcentage
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ Distribution d'échantillonnage de \bar{X}  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$	$H_0: p = p_0$ $H_1: p < p_0$ Distribution d'échantillonnage de \hat{P}  $\sigma_p = \sqrt{\frac{pq}{n}}$

Représentation d'un test sur l'égalité de deux paramètres

Test sur deux moyennes	Test sur deux pourcentages
<ul style="list-style-type: none"> Échantillons indépendants $H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 > 0$ Distribution d'échantillonnage de $\bar{X}_1 - \bar{X}_2$  $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2}$	$H_0: p_1 - p_2 = 0$ $H_1: p_1 - p_2 < 0$ Distribution d'échantillonnage de $\hat{P}_1 - \hat{P}_2$  $\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\sigma_{\hat{P}_1}^2 + \sigma_{\hat{P}_2}^2}$

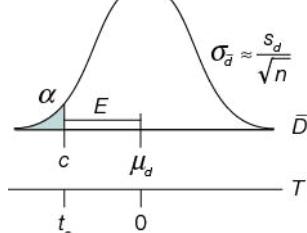
- Échantillons dépendants

$$(x, y) \rightarrow d = y - x$$

Distribution d'échantillonnage de \bar{D}

$$H_0: \mu_d = 0 \Leftrightarrow \mu_y = \mu_x$$

$$H_1: \mu_d < 0 \Leftrightarrow \mu_y < \mu_x$$



- $E = z\sigma_{\bar{D}}$ si $n \geq 30$.

- $E = t\sigma_{\bar{D}}$, où t suit une Student avec $df = n - 1$, si $n < 30$ et la population D suit une normale.

$$p \approx \hat{p} = \frac{\text{total de succès pour les deux échantillons}}{n_1 + n_2}$$

$$\sigma_{\hat{p}_1} = \sqrt{\frac{\hat{p}\hat{q}}{n_1}} \text{ et } \sigma_{\hat{p}_2} = \sqrt{\frac{\hat{p}\hat{q}}{n_2}}$$

Étape 3 Déterminer le ou les points critiques.

Attention ! Dans le cas d'un test sur une moyenne avec un échantillon de petite taille ($n < 30$) et une population normale d'écart type σ inconnu, l'écart tolérable $E = t\sigma_{\bar{x}}$, où t suit la loi de Student avec $dl = n - 1$.

Étape 4 Énoncer la règle de décision du test.

Étape 5 Prendre une décision, puis tirer une conclusion en fonction de l'hypothèse énoncée par le chercheur.

EXERCICES RÉCAPITULATIFS

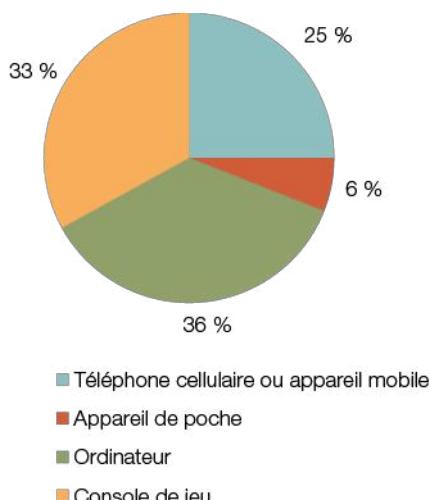
1. POUR LES AMATEURS DE JEUX VIDÉO

En 2013, un sondage est réalisé auprès d'un échantillon aléatoire de 4 183 Canadiens de 6 ans et plus. On dénombre 2 425 amateurs de jeux vidéo dans l'échantillon. Voici quelques statistiques à leur sujet :

Source: Association canadienne du logiciel de divertissement. *Faits essentiels 2013 sur le secteur canadien du jeu vidéo*.

- 1 310 amateurs de jeux vidéo sont des hommes ;
- 46 % disent y jouer plusieurs jours par semaine ;
- la moyenne d'âge des amateurs de jeux vidéo est de 31 ans avec un écart type corrigé de 12,5 ans ;
- à la question « Sur quelle plate-forme jouez-vous le plus souvent à des jeux vidéo ? », on a obtenu les réponses présentées dans le diagramme suivant.

Répartition des amateurs de jeux vidéo de l'échantillon selon la plate-forme la plus souvent utilisée pour jouer



a) Donner une estimation ponctuelle du pourcentage de Canadiens de 6 ans et plus qui jouent à des jeux vidéo en 2013.

b) Les données du sondage permettent-elles de conclure, au seuil de signification de 0,05, que les hommes représentent plus de la moitié des amateurs de jeux vidéo en 2013 ?

c) Une étude réalisée en 2010 indique que les joueurs de jeux vidéo de 6 ans et plus ont en moyenne 33 ans, que 34 % d'entre eux utilisent le plus souvent une console de jeu pour jouer et qu'ils sont 45 % à jouer plusieurs jours par semaine. Au seuil de signification de 0,05, les données du sondage permettent-elles de conclure :

- que les joueurs de jeux vidéo sont plus jeunes en 2013 ?
- que le pourcentage de joueurs qui utilisent le plus souvent une console de jeu est moins élevé en 2013 ?
- que le pourcentage de joueurs qui jouent plusieurs jours par semaine est différent en 2013 ?

2. Chaque mois, une étude est menée auprès d'un échantillon aléatoire de Québécois afin de suivre l'évolution du commerce électronique.

- En janvier 2013, 216 des 800 Québécois sondés ont dépensé une moyenne de 330 \$ pour des achats sur Internet, avec un écart type corrigé de 90 \$.
- En janvier 2014, 230 des 900 Québécois sondés ont dépensé une moyenne de 253 \$ pour des achats sur Internet, avec un écart type corrigé de 95 \$.

Source: CEFRIQ. *Indice du commerce électronique au Québec (ICEQ) en janvier 2014*, 28 février 2014.

Les différences observées entre les statistiques des deux échantillons permettent-elles de conclure, au seuil de signification de 0,05 :

- a) que le pourcentage de Québécois qui font des achats sur Internet est plus faible en janvier 2014 qu'en janvier 2013 ?
 - b) qu'en janvier 2014, le montant moyen des achats des cyberacheteurs est inférieur à celui de janvier 2013 ?
3. On veut tester l'hypothèse voulant que la présence d'un radar photo fasse diminuer le nombre d'accidents. Pour ce faire, on prélève un échantillon aléatoire de 7 endroits où un radar photo a été installé,

puis on dénombre les accidents dans un rayon de 4 km autour du radar sur une période de 5 ans avant et de 5 ans après l'installation du radar. Effectuer le test, au seuil de signification de 0,05. On suppose que la distribution de la différence du nombre d'accidents avant et après l'installation d'un radar photo suit une loi normale.

Nombre d'accidents avant et après la mise en service du radar photo

Zone	1	2	3	4	5	6	7
Avant	91	13	59	39	18	7	78
Après	60	9	21	49	6	5	43

PRÉPARATION À L'EXAMEN

Pour préparer votre examen, assurez-vous d'avoir les compétences suivantes :

	Si vous avez la compétence, cochez.
Le test d'hypothèse sur une moyenne	
• Formuler correctement les hypothèses H_0 et H_1 d'un test unilatéral ou bilatéral.	<input type="radio"/>
• Construire une règle de décision.	<input type="radio"/>
• Énoncer clairement la conclusion du test dans le contexte du problème.	<input type="radio"/>
• Interpréter le seuil de signification du test.	<input type="radio"/>
• Représenter, sur la courbe de la distribution d'échantillonnage de \bar{X} : la moyenne μ de la population, le seuil de signification, le ou les points critiques et la zone de rejet de H_0 .	<input type="radio"/>
Le test d'hypothèse sur un pourcentage	
• Formuler correctement les hypothèses H_0 et H_1 d'un test unilatéral ou bilatéral.	<input type="radio"/>
• Construire une règle de décision.	<input type="radio"/>
• Énoncer clairement la conclusion du test dans le contexte du problème.	<input type="radio"/>
• Interpréter le seuil de signification du test.	<input type="radio"/>
• Représenter, sur la courbe de la distribution d'échantillonnage de \hat{P} : le pourcentage p de la population, le seuil de signification, le ou les points critiques et la zone de rejet de H_0 .	<input type="radio"/>
Le test d'hypothèse sur l'égalité de deux moyennes	
Différencier les échantillons indépendants des échantillons dépendants ou appariés.	<input type="radio"/>
Échantillons indépendants	
• Formuler correctement les hypothèses H_0 et H_1 d'un test unilatéral ou bilatéral.	<input type="radio"/>
• Représenter le test sur la distribution d'échantillonnage de $\bar{X}_1 - \bar{X}_2$.	<input type="radio"/>
• Construire une règle de décision.	<input type="radio"/>
• Énoncer clairement la conclusion du test dans le contexte du problème.	<input type="radio"/>
Échantillons dépendants ou appariés	
• Formuler correctement les hypothèses H_0 et H_1 d'un test unilatéral ou bilatéral.	<input type="radio"/>
• Représenter le test sur la distribution d'échantillonnage de \bar{D} .	<input type="radio"/>
• Construire une règle de décision.	<input type="radio"/>
• Énoncer clairement la conclusion du test dans le contexte du problème.	<input type="radio"/>
Le test d'hypothèse sur l'égalité de deux pourcentages	
• Formuler correctement les hypothèses H_0 et H_1 d'un test unilatéral ou bilatéral.	<input type="radio"/>
• Représenter le test sur la distribution d'échantillonnage de $\hat{P}_1 - \hat{P}_2$.	<input type="radio"/>
• Construire une règle de décision.	<input type="radio"/>
• Énoncer clairement la conclusion du test dans le contexte du problème.	<input type="radio"/>



Chapitre 6

Les tests du khi-deux



OBJECTIFS DU CHAPITRE

- Utiliser un test d'ajustement du khi-deux pour valider une hypothèse portant sur:
 - l'ajustement d'une distribution d'une variable à un modèle théorique spécifié;
 - la représentativité d'un échantillon pour une variable spécifiée;
 - l'ajustement de la distribution d'une variable à une loi normale.
- Étudier la dépendance de deux variables à l'aide du test d'indépendance du khi-deux.

OBJECTIF DU LABORATOIRE

Le laboratoire 5 vise à apprendre à utiliser Excel pour effectuer un test d'indépendance du khi-deux.



Le chapitre 6 est consacré à l'étude de deux tests non paramétriques dont la validation fait intervenir la loi de probabilité du khi-deux: le test d'ajustement et le test d'indépendance.

6.1 Le test d'ajustement du khi-deux

Un test d'ajustement permet de décider, sur la base des données d'un échantillon aléatoire, si la distribution d'une variable d'une population suit une distribution spécifiée. Plus précisément, ce test répond à la question suivante : y a-t-il un bon ajustement entre la distribution d'une variable dans une population et un modèle théorique stipulé (uniforme, proportionnel, normal, binomial ou autre) ?

MISE EN SITUATION

Un de vos amis travaille la fin de semaine dans un magasin d'articles de farces et attrapes. Samedi dernier, alors qu'il servait un client, des adolescents se sont amusés à mélanger les dés pipés et les dés non pipés. Comme les dés sont en apparence tous parfaitement identiques, le seul moyen de les différencier est de les lancer plusieurs fois et de décider s'ils sont pipés d'après les résultats obtenus. Votre ami vous a demandé de tester trois dés. Voici les résultats que vous avez obtenus pour chaque dé en 120 lancers.

Dé 1

Face	1	2	3	4	5	6	Total
Effectifs	39	26	38	5	8	4	120

Dé 2

Face	1	2	3	4	5	6	Total
Effectifs	24	16	22	23	18	17	120

Dé 3

Face	1	2	3	4	5	6	Total
Effectifs	14	16	28	30	18	14	120

❓ En examinant les résultats, pouvez-vous identifier un ou plusieurs dés pipés ?

Dé(s) pipé(s) : _____

Analysons la démarche qui a mené à cette conclusion.

❓ 1. Avez-vous identifié le ou les dés pipés en vous appuyant sur le comportement d'un dé pipé ou d'un dé non pipé ?

Sur le comportement d'un dé _____

2. *A priori*, vous saviez que si le dé lancé n'était pas pipé, le nombre de résultats serait presque le même pour chaque face.

Théoriquement, pour un dé non pipé, la distribution des résultats attendue est :

Distribution théorique des résultats de 120 lancers d'un dé non pipé

Face	1	2	3	4	5	6	Total
Effectifs théoriques	20	20	20	20	20	20	120

Comme on obtient le même nombre de résultats pour chaque face du dé, on dit que l'on a une **distribution uniforme** ou que les résultats se distribuent uniformément.

3. Spontanément, vous avez donc comparé la distribution des résultats observés pour chaque face à la distribution théorique espérée pour un dé non pipé, soit environ le même nombre de résultats pour chaque face.
4. Pour chaque dé testé, vous avez pris une décision en portant un jugement sur l'importance des écarts entre les distributions observée et espérée, et vous êtes arrivé aux conclusions suivantes.

Dé 1

La grandeur des écarts observés semble indiquer que ce dé est pipé.

Dé 2

L'ajustement entre les distributions théorique et observée semble assez bon ; les écarts constatés sont fort probablement attribuables au hasard.

Dé 3

Il est plus difficile de prendre une décision pour ce dé. Peut-on dire, par exemple, que l'écart de 10 que l'on obtient pour la face 4, entre les effectifs espérés (20) et les effectifs observés (30), est trop grand pour être attribué au hasard ? À partir de quelle valeur l'écart est-il trop important pour être attribuable au hasard ?

Un test d'hypothèse, appelé «test d'ajustement du khi-deux», permet d'établir une règle de décision dans une telle situation.

6.1.1 La construction d'un test d'ajustement du khi-deux

La démarche de construction d'un test d'ajustement est analogue à celle qu'on applique spontanément pour tester des dés. Il suffit de l'exprimer sous une forme plus rigoureuse.

Voici les étapes du test d'ajustement du khi-deux qui permet de décider si le dé 3 de la mise en situation est pipé ou non.

Étape 1. Formuler les hypothèses

Il faut d'abord formuler les deux hypothèses du test : l'hypothèse nulle (H_0) et l'hypothèse alternative ou hypothèse du chercheur (H_1).

H_0 : L'hypothèse nulle

C'est l'hypothèse qui est testée. Toute la démarche repose sur le fait qu'on considère cette hypothèse comme vraie jusqu'à preuve du contraire. C'est en se basant sur l'hypothèse nulle qu'on détermine les effectifs théoriques. Dans un test d'ajustement du khi-deux, elle a la forme suivante :

« La variable étudiée se distribue selon un modèle théorique spécifié. »

H_1 : L'hypothèse alternative ou hypothèse du chercheur

C'est l'hypothèse qui est acceptée si l'on rejette l'hypothèse nulle. Elle a la forme suivante :

« La variable étudiée ne se distribue pas selon le modèle théorique spécifié. »

EXEMPLE 1

Les hypothèses du test d'ajustement du khi-deux pour le dé 3 sont :

H_0 : Le dé 3 n'est pas pipé : les résultats se distribuent uniformément.

C'est l'hypothèse qui est testée.

H_1 : Le dé 3 est pipé : les résultats ne se distribuent pas uniformément.

C'est l'hypothèse qui est acceptée si l'on rejette l'hypothèse nulle H_0 .

Étape 2. Calculer les effectifs théoriques selon l'hypothèse nulle H_0

On calcule les effectifs théoriques, c'est-à-dire les effectifs qu'on devrait obtenir pour chaque catégorie en supposant que la distribution des résultats suit le modèle théorique spécifié dans l'hypothèse nulle. Il est important de souligner qu'on utilise toujours les effectifs, jamais les fréquences relatives. On désigne les effectifs théoriques par la lettre T et les effectifs observés par la lettre O .

EXEMPLE 1 (suite)

Effectifs observés et théoriques pour le dé 3 :

Face	1	2	3	4	5	6	Total
Effectifs observés (O)	14	16	28	30	18	14	120
Effectifs théoriques (T)	20	20	20	20	20	20	120

La distribution des effectifs théoriques est basée sur la probabilité d'obtenir une face quelconque du dé, soit 1/6.

Condition d'application d'un test du khi-deux

Pour pouvoir appliquer un test du khi-deux, il faut que l'échantillon aléatoire soit suffisamment grand pour que tous les effectifs théoriques soient supérieurs ou égaux à 5. Si cette condition¹ n'est pas respectée, on regroupe deux ou plusieurs catégories adjacentes de manière à ce qu'elle le soit. Le calcul des effectifs théoriques avant la collecte de données permet de choisir un échantillon ayant la taille voulue pour éviter ce genre de problème. (À titre d'exemples, voir les numéros 9 et 11 des exercices 6.1.)

NOTE

Il est à souligner qu'aucune condition n'est imposée aux effectifs observés pour la validité d'un test du khi-deux ; seuls les effectifs théoriques doivent être égaux ou supérieurs à 5.

Étape 3. Calculer la valeur du khi-deux

L'ajustement entre deux distributions est mesuré par le khi-deux (ou khi carré), que l'on note χ^2 , selon la formule de la page suivante. Plus la valeur du khi-deux est grande, plus les écarts entre les effectifs observés (O) et théoriques (T) sont importants.

1. Il n'y a pas unanimité sur les conditions d'application d'un test du khi-deux. Nous avons retenu ici la règle de Fischer ($T \geq 5$) en raison de sa simplicité. Toutefois, si ce critère n'est pas respecté, et qu'un regroupement de catégories adjacentes cause une perte d'information importante, on vérifie si la règle de Cochran s'applique : les effectifs théoriques de 80 % des catégories doivent être égaux ou supérieurs à 5 et aucune catégorie ne doit avoir un effectif théorique nul.

Valeur du khi-deux

$$\chi^2 = \sum \frac{(O-T)^2}{T}$$

EXEMPLE 1 (suite)

Comme les effectifs théoriques du dé 3 sont tous supérieurs ou égaux à 5, on peut poursuivre le test d'ajustement en calculant la valeur du khi-deux comme suit :

$$\chi^2 = \frac{(14-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(30-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(14-20)^2}{20} = 12,8$$

NOTE

La formule donnant la valeur du khi-deux s'explique ainsi :

- $(14 - 20) = -6$ est l'écart entre l'effectif observé et l'effectif théorique de la face 1.
- $(14 - 20)^2 = 36$ est le carré de l'écart. L'élévation au carré permet d'éliminer les signes négatifs dans le cas d'écart négatifs. Sans cette opération, la somme des écarts, pour les 6 faces, serait égale à 0.
- $(14 - 20)^2/20$ permet de relativiser l'importance des écarts. Par exemple, un écart de +2 pour un effectif espéré de 4 est plus important qu'un écart de +2 pour un effectif espéré de 100 : $(2^2/4 = 1) > (2^2/100 = 0,04)$.
- La somme obtenue en additionnant les résultats pour chaque face donne une vue d'ensemble de l'ajustement entre la distribution observée et la distribution théorique.

Étape 4. Énoncer la règle de décision

Après avoir calculé le χ^2 , il faut décider si une différence d'ajustement de 12,8 entre les effectifs observés et théoriques est trop grande pour être attribuée au hasard d'échantillonnage. Il va de soi que si l'ajustement était parfait, le khi-deux serait égal à 0. La question est donc de savoir à partir de quelle valeur le χ^2 est jugé trop grand pour que le résultat soit attribuable au hasard.

C'est Pearson (*voir la page 275*) qui découvre, en 1904, la loi du hasard s'appliquant à la valeur du χ^2 . Il démontre que la distribution de tous les χ^2 possibles, selon l'hypothèse que la distribution des résultats des échantillons suit une distribution théorique donnée, obéit à une loi de probabilité qu'il nomme **loi du khi-deux**. Plus précisément, il découvre une loi de probabilité pour chaque degré de liberté associé à la distribution étudiée.

Nombre de degrés de liberté

La loi du khi-deux à appliquer dépend du nombre de degrés de liberté, noté dl . Pour un test d'ajustement, on a :

Nombre de degrés de liberté pour un test d'ajustement

$$dl = \text{nombre de catégories} - 1$$

EXEMPLE 1 (suite)

Nombre de degrés de liberté pour le dé 3 : $dl = 6 - 1 = 5$

On peut expliquer l'idée de degrés de liberté à l'aide de cet exemple :

Sachant que la somme des 6 effectifs donne 120, on peut facilement déterminer la valeur d'un de ces 6 effectifs pourvu que l'on connaisse la valeur des 5 autres. On considère alors qu'il y a 5 effectifs libres et un effectif non libre, ou lié. Par exemple, si 5 des 6 effectifs sont 17, 16, 22, 28 et 12, alors le 6^e doit être 25 pour que la somme donne 120.

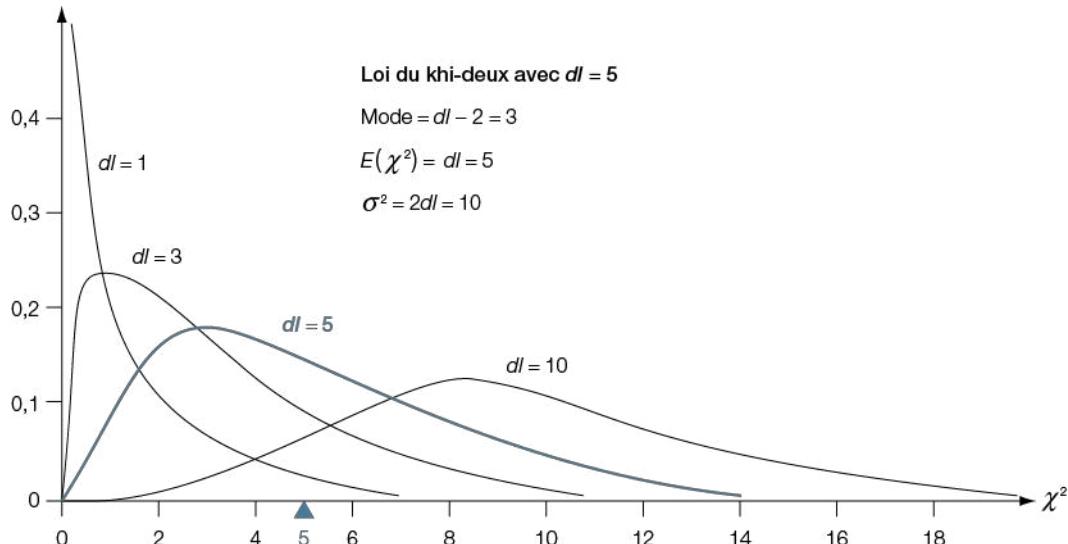
Caractéristiques des lois du khi-deux

- Les valeurs que peut prendre le khi-deux vont de 0 à l'infini : $0 \leq \chi^2 < \infty$.
- L'espérance d'une distribution du khi-deux est égale au nombre de degrés de liberté, et sa variance est égale à deux fois le nombre de degrés de liberté :

$$E(\chi^2) = dl \quad \sigma^2 = 2dl$$

- La forme de la loi du khi-deux dépend du nombre de degrés de liberté. C'est généralement une cloche présentant une dissymétrie qui est d'autant moins prononcée que le nombre de degrés de liberté est grand. Plus ce nombre augmente, plus la forme de la courbe s'approche de la forme de la courbe normale.
- Pour un nombre de degrés de liberté supérieur à 2, la valeur du khi-deux où la courbe atteint un sommet, donc le mode, est égale au nombre de degrés de liberté moins 2 :

$$\text{Mode} = dl - 2 \quad \text{pour } dl > 2$$



Règle de décision

On décide de rejeter ou non l'hypothèse nulle en comparant le khi-deux obtenu à un point critique, appelé **khi-deux critique** et noté χ_c^2 . La règle de décision s'énonce ainsi :

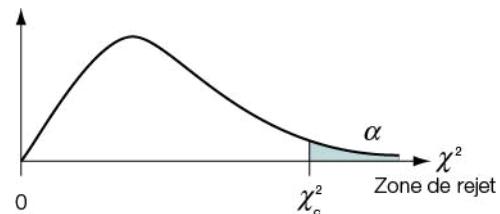
Règle de décision d'un test d'ajustement

Rejeter H_0 si la valeur du khi-deux est supérieure au khi-deux critique.

Rejeter H_0 si $\chi^2 > \chi_c^2$.

Le point critique, que l'on trouve dans la table du khi-deux (voir l'annexe 5, à la page 349), dépend du **seuil de signification**, noté α , et du nombre de degrés de liberté. Le seuil de signification, dont la valeur est généralement 0,05 ou 0,01, est défini ainsi :

$$\alpha = P(\text{rejeter l'hypothèse } H_0 \text{ alors que } H_0 \text{ est vraie})$$



Le seuil de signification correspond donc aux risques de se tromper lorsqu'on prend la décision de rejeter l'hypothèse nulle : on conclut que l'hypothèse H_0 est fausse alors qu'en fait elle est vraie.

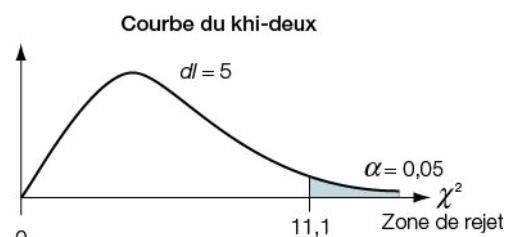
EXEMPLE 1 (suite)

Construisons la règle de décision du test d'hypothèse du dé 3 au seuil de signification de 0,05.

Pour $dl = 5$ et $\alpha = 0,05$, la table du khi-deux donne 11,1 comme khi-deux critique.

La règle de décision s'énonce ainsi :

Rejeter H_0 si $\chi^2 > \chi_c^2 = 11,1$.



Étape 5. Décider et conclure

On décide du rejet ou du non-rejet de l'hypothèse H_0 après avoir appliqué la règle de décision. Par la suite, on tire une conclusion de cette décision dans le contexte du problème.

EXEMPLE 1 (suite)

Pour le dé 3, la valeur du χ^2 est 12,8. Cette différence d'ajustement entre les distributions observée et théorique est-elle trop grande pour être attribuable au hasard ? Le dé 3 est-il pipé ?

Comme $\chi^2 = 12,8 > 11,1$, on rejette H_0 . Le dé 3 est pipé.

Les risques de se tromper en prenant cette décision sont inférieurs à 5 % ; il y a donc 5 % de chances que le dé ne soit pas pipé et que la différence d'ajustement entre les deux distributions soit attribuable au hasard.

NOTE

Pour tester une hypothèse voulant que la distribution d'une variable d'une population suive une loi binomiale $B(n; p)$ ou une loi de Poisson $Po(\lambda)$, on détermine les pourcentages théoriques du test d'ajustement en utilisant la table de probabilité de ces lois. (Pour des exemples, voir les numéros 8 et 9 des exercices 6.1).



Karl Pearson (1857-1936)

Né à Londres, le mathématicien, physicien et historien Karl Pearson s'intéresse également à l'hérédité et à la génétique, car il embrasse la théorie du darwinisme social. Sous l'influence du généticien Francis Galton, il se tourne vers l'étude de la statistique. Il crée plus tard le test du khi-deux, utile pour déterminer si les écarts observés dans un ensemble de variables par rapport aux valeurs théoriques peuvent être attribués ou non à l'échantillonnage. Il est reconnu comme un des fondateurs de la statistique moderne.

EXEMPLE 2

Un chercheur désire comparer la distribution des revenus des familles immigrantes du Québec à celle des revenus de l'ensemble des familles québécoises. Cette dernière distribution est connue grâce au recensement, mais pas celle des revenus des familles immigrantes du Québec. Le chercheur décide donc de procéder par échantillonnage pour faire son étude. En prenant un échantillon aléatoire de 500 familles immigrantes, il obtient la distribution suivante.

Répartition d'un échantillon de 500 familles immigrantes selon la tranche de revenu

Revenu (milliers \$)	Moins de 25	[25; 50[[50; 75[[75; 100[100 et plus	Total
Nombre de familles	44	142	129	65	120	500
Pourcentage	8,8 %	28,4 %	25,8 %	13,0 %	24,0 %	100 %

Répartition des familles selon la tranche de revenu, Québec, 2011

Revenu (milliers \$)	Moins de 25	[25; 50[[50; 75[[75; 100[100 et plus	Total
Pourcentage	6,1 %	26,3 %	23,4 %	16,4 %	27,8 %	100 %

Source: Statistique Canada. Tableau 202-0408, CANSIM, juin 2013.

En comparant les pourcentages des deux distributions, on constate que les familles immigrantes sont moins riches : un plus grand pourcentage de ces familles ont un revenu faible et un plus petit pourcentage ont un revenu élevé.

En fait, cette affirmation est vraie pour les 500 familles immigrantes de notre échantillon, mais est-elle vraie pour l'ensemble de toutes les familles immigrantes du Québec ? Il est en effet possible que les distributions pour l'ensemble de toutes les familles immigrantes et québécoises soient identiques et que les écarts observés ci-dessus soient attribuables à la variation d'échantillonnage causée par le hasard. Un test d'ajustement du khi-deux permet de savoir si c'est le cas.

Effectuer un test d'ajustement du khi-deux, au seuil de signification de 0,01, pour déterminer si la distribution des revenus des familles immigrantes est identique à celle des revenus des familles québécoises.

Solution

1. Formulation des hypothèses

H_0 : La distribution des revenus des familles immigrantes est _____ à celle des revenus des familles québécoises.

H_1 : La distribution des revenus des familles immigrantes est _____ de celle des revenus des familles québécoises.

2. Calcul des effectifs théoriques selon l'hypothèse nulle H_0 .

(Le résultat obtenu ne doit pas être arrondi à l'entier sous prétexte que les données sont entières. L'effectif calculé est un nombre théorique ; on conserve donc une décimale après la virgule.)

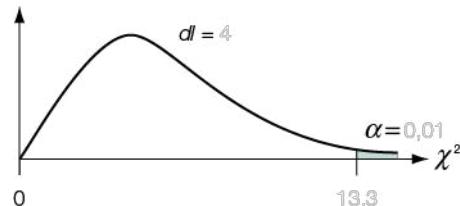
Revenu (milliers \$)	Moins de 25	[25; 50[[50; 75[[75; 100[100 et plus	Total
Effectifs observés (O)	44	142	129	65	120	500
Pourcentages théoriques	6,1 %	26,3 %	23,4 %	16,4 %	27,8 %	100 %
Effectifs théoriques (T)			117,0	82,0	139,0	500

La condition d'application du test du khi-deux est-elle respectée ? _____

3. Calcul du khi-deux

$$\chi^2 = \frac{(\text{---} - \text{---})^2}{117} + \frac{(\text{---} - \text{---})^2}{82} + \frac{(129 - 117)^2}{117} + \frac{(65 - 82)^2}{82} + \frac{(120 - 139)^2}{139} = 14,2$$

4. Règle de décision



5. Décision et conclusion

Au seuil de signification de 0,01, les distributions des revenus des familles immigrantes et québécoises sont différentes : les écarts observés ne sont pas attribuables au hasard. L'analyse des écarts montre que les familles immigrantes sont vraisemblablement moins riches.

Utilisation des touches de mémoire pour calculer le khi-deux

Si votre calculatrice ne permet pas de retourner à une expression pour la corriger, nous vous suggérons de calculer le khi-deux en utilisant la mémoire ; cela vous évitera de devoir tout recommencer si vous faites une erreur en calculant un terme.

Voici comment on utilise la mémoire pour effectuer le calcul suivant :

$$\frac{(6-2)^2}{2} + \frac{(5-4)^2}{4} = 8,25$$

1. Mettre la mémoire à 0 : appuyer sur **0**, sur **STO**, puis sur **M+ (M)** (ou **X → M**).
2. • Calculer le premier terme : $\frac{(6-2)^2}{2} =$ (Pour certaines calculatrices, il ne faut pas appuyer sur **=**.)
• Additionner la réponse au nombre actuellement dans la mémoire : appuyer sur **M+ (ou SUM)**.
3. Refaire la deuxième étape pour chaque terme à additionner.
4. Récupérer la somme des termes dans la mémoire : appuyer sur **RCL M** (ou **RM**). Le nombre 8,25 s'affiche.

EXERCICE DE COMPRÉHENSION | 6.1

Une micro-brasserie s'apprête à lancer sur le marché cinq nouvelles bières : trois blondes (B1, B2, B3) et deux rousses (R1, R2). Dans le cadre d'une étude de marché, on demande à 800 personnes choisies au hasard d'indiquer, après dégustation, laquelle des cinq bières elles préfèrent.

Répartition des 800 personnes de l'échantillon selon la bière qu'elles préfèrent

Bière préférée	B1	B2	B3	R1	R2	Total
Nombre de personnes	156	162	212	110	160	800

- a) Avant d'effectuer l'étude de marché, on avait émis l'hypothèse que les consommateurs préféraient les bières blondes aux bières rousses. Tester cette hypothèse au seuil de signification de 0,01.

Solution

1. Formulation des hypothèses

H_0 :

H_1 :

2. Calcul des effectifs théoriques selon l'hypothèse nulle H_0

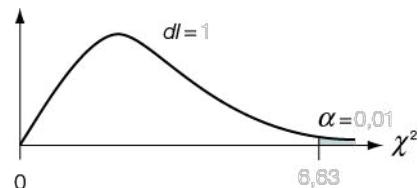
Couleurs des bières			Total
Effectifs observés (O)			
Pourcentages théoriques			
Effectifs théoriques (T)			

La condition d'application du test est-elle respectée ? _____

3. Calcul du khi-deux

4. Règle de décision

5. Décision et conclusion



- b) On avait aussi émis l'hypothèse que le choix des consommateurs se répartirait uniformément entre les trois sortes de bières blondes. Les données échantillonnelles confirment-elles cette hypothèse au seuil de signification de 0,01 ?

Solution

1. Formulation des hypothèses

H_0 :

H_1 :

2. Calcul des effectifs théoriques selon l'hypothèse nulle H_0

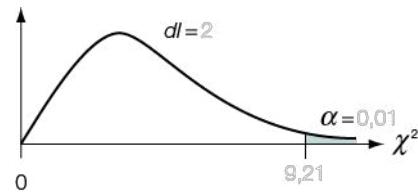
Sortes de bières blondes				Total
Effectifs observés (O)				
Pourcentages théoriques				
Effectifs théoriques (T)				

La condition d'application du test est-elle respectée ? _____

3. Calcul du khi-deux

4. Règle de décision

5. Décision et conclusion



6.1.2 La représentativité d'un échantillon

On dit d'un échantillon qu'il est **représentatif** si les caractéristiques de la population se retrouvent dans les mêmes proportions dans l'échantillon ; l'échantillon est alors une miniaturisation de la population ou un modèle réduit de celle-ci.

Pour qu'un échantillon soit représentatif, il n'est toutefois pas nécessaire qu'il soit tout à fait similaire à la population sous tous les aspects ; il suffit qu'il existe une similitude entre l'échantillon et la population par rapport à certaines caractéristiques jugées importantes pour l'étude statistique en cours. Par exemple, si l'on effectue un sondage afin de connaître les intentions de vote des électeurs aux prochaines élections, la similitude des distributions pour les variables suivantes est importante : sexe, âge, langue maternelle, répartition régionale, etc. Il va de soi que des variables telles le nombre d'enfants ou le fait d'être propriétaire ou non d'une auto ont beaucoup moins d'importance pour ce type de sondage. Le caractère représentatif d'un échantillon est donc purement relatif : il peut être représentatif pour un sujet d'étude donné, mais non pour un autre sujet.

Bien qu'un échantillon choisi au hasard devrait être représentatif, on préfère quand même vérifier sa représentativité quant à certaines variables avant de réaliser une étude. Une telle prudence est particulièrement de mise dans le cas d'un échantillonnage stratifié ou par grappes. Si l'on applique une de ces méthodes d'échantillonnage, il faut parfois consulter plusieurs intervenants afin de constituer la base de données de l'échantillon, ce qui peut entraîner des erreurs. De plus, il faut s'assurer que les grappes ne présentent pas un problème d'homogénéité par rapport à la variable étudiée.

Le test d'ajustement du khi-deux permet de tester la représentativité de l'échantillon pour les variables jugées importantes.

EXEMPLE

On désire mesurer l'effet des nouvelles technologies de communication sur la vie quotidienne des Québécois de 25-64 ans. Pour ce faire, on prélève au hasard un échantillon de 800 personnes dans cette population. Comme on considère que le niveau de scolarité est une variable importante dans ce genre d'étude, on veut s'assurer de la représentativité de l'échantillon pour cette variable avant de procéder à la cueillette des données. Les statistiques présentées dans les deux tableaux suivants permettent-elles d'affirmer que l'échantillon est représentatif des Québécois de 25-64 ans en ce qui concerne le niveau de scolarité, au seuil de signification de 0,05 ?

**Répartition des Québécois de 25-64 ans
selon le plus haut niveau de scolarité atteint, Québec, 2012**

Niveau de scolarité	Aucun diplôme	Diplôme secondaire	Diplôme collégial	Diplôme universitaire	Total
Pourcentage	12,3 %	33,5 %	22,2 %	32,0 %	100,0 %

Source: Statistique Canada. *Enquête sur la population active*, 2013, adapté par l'Institut de la statistique du Québec, juin 2014.

Répartition des 800 répondants selon le niveau de scolarité

Niveau de scolarité	Aucun diplôme	Diplôme secondaire	Diplôme collégial	Diplôme universitaire	Total
Effectifs	91	258	207	244	800

Solution

1. Formulation des hypothèses

H_0 : L'échantillon est représentatif de la population pour le niveau de scolarité.

H_1 : L'échantillon n'est pas représentatif de la population pour le niveau de scolarité.

2. Calcul des effectifs théoriques selon l'hypothèse nulle H_0

Niveau de scolarité	Aucun diplôme	Diplôme secondaire	Diplôme collégial	Diplôme universitaire	Total
O	91	258	207	244	800
% T	12,3 %	33,5 %	22,2 %	32,0 %	100,0 %
T	98,4	268,0	177,6	256,0	800

Aucun effectif théorique n'étant inférieur à 5, la condition d'application du test est respectée.

3. Calcul du khi-deux

$$\chi^2 = \sum \frac{(O - T)^2}{T} = 6,4$$

4. Règle de décision

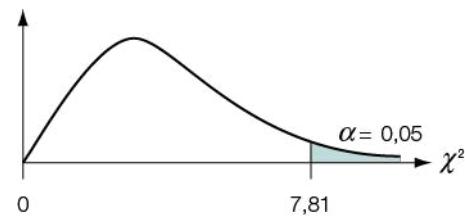
Pour $dl = 4 - 1 = 3$ et $\alpha = 0,05$, on a $\chi^2_c = 7,81$.

Rejeter H_0 si $\chi^2 > \chi^2_c = 7,81$.

5. Décision et conclusion

Comme $\chi^2 = 6,4 < 7,81$, on ne rejette pas H_0 .

Au seuil de signification de 0,05, il n'y a aucune évidence statistique permettant de douter de la représentativité de l'échantillon pour la variable «niveau de scolarité» dans la population. La différence entre les effectifs observés et théoriques est attribuable aux fluctuations d'échantillonnage.



6.1.3 La population est-elle conforme au modèle normal?

En inférence statistique, on émet souvent l'hypothèse que la population suit un modèle normal, c'est-à-dire que les valeurs de la variable étudiée se distribuent selon la loi normale. La plupart du temps, cette normalité de la population se vérifie par échantillonnage. Jusqu'à présent, nous nous sommes contentés de juger, de façon qualitative, de la normalité de la distribution en examinant la forme de l'histogramme. Le test d'ajustement du khi-deux permet de valider scientifiquement cette normalité.

EXEMPLE

Pour dresser le profil statistique des passagers des navires de croisières qui accostent au port de Québec, on prélève un échantillon aléatoire de 1 000 croisiéristes. La moyenne d'âge de ces derniers est de 64,7 ans avec un écart type corrigé de 12,1 ans. Le tableau suivant donne la distribution de l'âge des personnes de l'échantillon. Au seuil de signification de 0,05, ces données permettent-elles d'affirmer que la distribution de l'âge des croisiéristes dans la population suit une distribution normale ?

Répartition des croisiéristes de l'échantillon selon l'âge

Âge (en ans)	[30; 40[[40; 50[[50; 60[[60; 70[[70; 80[[80; 90[Total
Effectifs	24	90	230	310	239	107	1 000

Solution

1. Formulation des hypothèses

H_0 : La distribution de l'âge des croisiéristes de la population suit une loi normale $N(\mu; \sigma^2)$.

H_1 : La distribution de l'âge des croisiéristes de la population ne suit pas une loi normale $N(\mu; \sigma^2)$.

2. Calcul des effectifs théoriques selon l'hypothèse nulle H_0

- Estimation de μ et de σ

Les paramètres μ et σ de la loi normale étant inconnus, on les estime ponctuellement par la moyenne \bar{x} et l'écart type corrigé s de l'échantillon :

$$N(\mu; \sigma^2) \approx N(\bar{x}; s^2) \approx N(64,7; 12,1^2)$$

- Modification de la notation de la première et de la dernière classe

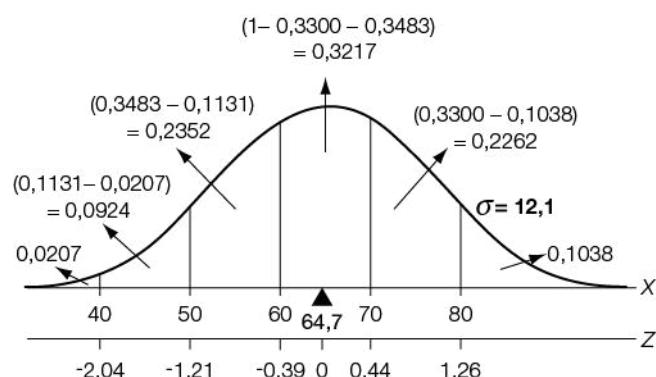
Sous l'hypothèse d'une distribution normale, les valeurs doivent théoriquement se situer dans l'intervalle $]-\infty; +\infty[$; c'est pourquoi nous effectuons les modifications suivantes :

- la classe [30 ans ; 40 ans[devient «moins de 40 ans»;
- la classe [80 ans ; 90 ans[devient «80 ans et plus».

- Calcul des pourcentages théoriques

Sous l'hypothèse que la distribution de l'âge suit une normale $N(64,7; 12,1^2)$, on calcule le pourcentage théorique de chaque classe ainsi :

- on représente la surface dont on cherche l'aire sur la courbe normale $N(64,7; 12,1^2)$;
- on détermine la cote z pour chaque limite de classe, puis l'aire associée à la cote z dans la table de la loi normale centrée réduite. On obtient l'aire recherchée par déduction graphique.



On obtient les pourcentages et les effectifs théoriques suivants pour chaque classe.

Âge (en ans)	Moins de 40	[40; 50[[50; 60[[60; 70[[70; 80[80 et plus	Total
O	24	90	230	310	239	107	1 000
% T	2,07 %	9,24 %	23,52 %	32,17 %	22,62 %	10,38 %	100,00 %
T	20,7	92,4	235,2	321,7	226,2	103,8	1 000

Aucun effectif théorique n'étant inférieur à 5, la condition d'application du test est respectée.

3. Calcul du khi-deux

$$\chi^2 = \sum \frac{(O-T)^2}{T} = 1,95$$

4. Règle de décision

Pour trouver le khi-deux critique associé au seuil de signification, on doit d'abord déterminer le nombre de degrés de liberté.

Nombre de degrés de liberté pour un test d'ajustement à une loi normale

$$dl = \text{nombre de classes} - \text{nombre de paramètres estimés} - 1$$

Comme les deux paramètres μ et σ ont été estimés respectivement par \bar{x} et s , on a :

$$dl = 6 - 2 - 1 = 3$$

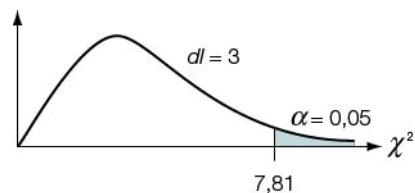
Pour $\alpha = 0,05$ et $dl = 3$, on a $\chi^2_c = 7,81$.

Rejeter H_0 si $\chi^2 > \chi^2_c = 7,81$.

5. Décision et conclusion

Comme $\chi^2 = 1,95 < 7,81$, on ne rejette pas H_0 .

La distribution de l'âge des croisiéristes de la population suit vraisemblablement un modèle normal.



EXERCICE DE COMPRÉHENSION | 6.2

Le tableau suivant donne la distribution des notes d'un échantillon aléatoire de 100 élèves à l'examen d'anglais de 5^e secondaire du ministère de l'Éducation. Au seuil de signification de 0,01, peut-on considérer que la distribution des notes de l'ensemble des élèves à cet examen suit un modèle normal ? Compléter la démarche du test d'hypothèse.

Répartition des élèves de l'échantillon selon la note à l'examen d'anglais du ministère de l'Éducation

Note	Moins de 60	[60; 70[[70; 80[[80; 90[[90; 100]	Total
Effectifs	3	36	25	20	16	100

Solution

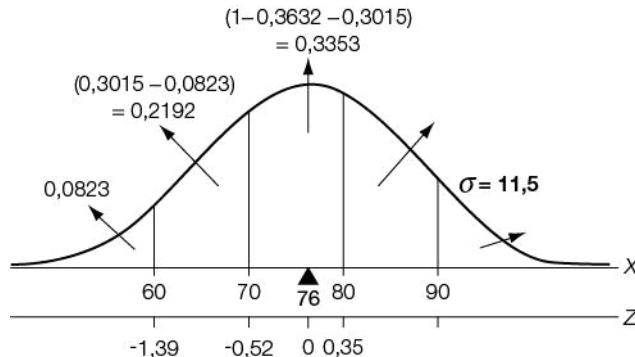
1. Hypothèses

H_0 : La distribution des notes de l'ensemble des élèves _____.

H_1 : La distribution des notes de l'ensemble des élèves _____.

2. Calcul des effectifs théoriques

- $\mu \approx \bar{x} = 76$ et $\sigma \approx s = 11,5$.
- Pour un ajustement au modèle normal, la classe [90 ; 100[devient _____.
- Calcul des pourcentages théoriques
Compléter le graphique et le tableau.



Note	Moins de 60	[60; 70[[70; 80[[80; 90[90 et plus	Total
O	3	36	25	20	16	100
% T	8,2 %	21,9 %	33,5 %			100,0 %
T	8,2	21,9	33,5			100

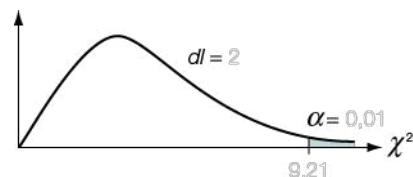
La condition d'application du test est-elle respectée ? Justifier la réponse. _____

3. Calcul du khi-deux

$$\chi^2 = \sum \frac{(O-T)^2}{T} = 17,8$$

4. Règle de décision

Pour $\alpha = 0,01$ et $dl = 2$,
on a $\chi^2_c = 9,21$.



5. Décision et conclusion

EXERCICES 6.1

1. Dans un casino, la surface de deux roulettes apparemment identiques est divisée en trois secteurs de couleurs différentes : rouge, vert et noir. Toutefois, l'une des roulettes est truquée. Pour déterminer laquelle, on fait tourner chaque roulette 100 fois. Voici les résultats obtenus :

Roulette 1 : 30 fois la couleur rouge
35 fois la couleur verte
35 fois la couleur noire

Roulette 2 : 58 fois la couleur rouge
27 fois la couleur verte
15 fois la couleur noire

- a) Pouvez-vous identifier la roulette truquée à l'aide des résultats ? Justifier la réponse.
- b) i) Déterminer quelle roulette est truquée sachant que 60 % de la surface de chaque roulette est rouge, 30 % verte et 10 % noire.
ii) La décision prise en i) repose-t-elle sur la distribution attendue de 100 lancers d'une roulette truquée ou d'une roulette non truquée ?
2. On lance une pièce de monnaie 1 000 fois dans le but de tester si elle est truquée ou non.
- a) Donner les effectifs théoriques si l'on pose comme hypothèse nulle :
 H_0 : La pièce est truquée.
- b) Donner les effectifs théoriques si l'on pose comme hypothèse nulle :
 H_0 : La pièce n'est pas truquée.
- c) Parmi les trois situations suivantes, sans faire de calculs, déterminer celle qui nécessiterait qu'on effectue un test d'ajustement du khi-deux pour décider si la pièce de monnaie est truquée ou non.
- Situation 1 : Effectifs observés :
face : 508 pile : 492
- Situation 2 : Effectifs observés :
face : 575 pile : 425
- Situation 3 : Effectifs observés :
face : 720 pile : 380
3. Un restaurateur offre trois saveurs de crème glacée comme dessert du jour : vanille, chocolat et citron. Pour confirmer son impression que ses clients préfèrent la crème glacée au chocolat, il décide d'effectuer un test d'ajustement du khi-deux en utilisant

les résultats d'un échantillon aléatoire de 45 clients ayant pris le dessert du jour.

- a) Énoncer l'hypothèse nulle H_0 .
b) Donner les effectifs théoriques du test.

4. On entend souvent dire qu'il y a plus de naissances en période de pleine lune. Tester cette affirmation en utilisant les données obtenues pour un échantillon aléatoire de 360 naissances. Utiliser un seuil de signification de 0,05.

Phase de la lune	Nouvelle lune	Premier quartier	Pleine lune	Dernier quartier	Total
Nombre de naissances	76	88	100	96	360

5. ATTENTION AUX CARIBOUS !

De 2004 à 2012, une étude mesure l'impact de l'élargissement de la route 175 en autoroute sur le comportement des caribous de la réserve faunique des Laurentides. Pour déterminer si une période de la journée présente plus de risques d'accident avec un caribou, on analyse un échantillon aléatoire de 105 traversées de la route 175 par un caribou en fonction de l'heure (bien entendu, ce n'est pas toujours le même animal!).

Répartition de 105 traversées de la route 175 par un caribou selon l'heure

Périodes de la journée		Nombre de traversées
Jour	6 h à 10 h	16
	10 h à 14 h	35
	14 h à 18 h	27
Nuit	18 h à 0 h	15
	0 h à 6 h	12
Total		105

Source: Ministère des Ressources naturelles et de la Faune du Québec. *Impacts de la réfection de l'axe routier 73/175 sur le caribou forestier de Charlevoix*, février 2013.

Faire les tests suivants au seuil de signification de 0,01.

- a) Les données échantillonnelles permettent-elles de conclure que les traversées ne se répartissent pas uniformément entre les 5 périodes de la journée ?
- b) Toutes proportions gardées, peut-on conclure qu'il y a plus de traversées le jour que la nuit ?
- c) Tester l'hypothèse voulant que les traversées se répartissent uniformément entre les 3 périodes de jour.

6. On désire sonder les jeunes de 18 à 24 ans sur leurs habitudes de consommation à l'aide d'un échantillon aléatoire de 800 jeunes de ce groupe d'âge. Comme le revenu est important dans ce type d'étude, on veut tester la représentativité de l'échantillon pour cette variable. Sur le site de Revenu Québec, on trouve la distribution suivante pour les revenus des Québécois de 18 à 24 ans :

0 \$: 3,9 %
De 1 \$ à 24 999 \$: 77,2 %
De 25 000 \$ à 49 999 \$: 16,1 %
50 000 \$ et plus : 2,8 %

Source: Revenu Québec. *Le revenu total de particuliers, 2011*, mars 2014.

- a) Formuler les hypothèses H_0 et H_1 .
b) Donner les effectifs théoriques.

7. On prélève un échantillon de 1 200 étudiants universitaires québécois afin d'étudier leurs conditions de vie. On considère qu'il est important que l'échantillon soit représentatif de la population universitaire en ce qui concerne le cycle d'études (baccalauréat, maîtrise, doctorat). Tester cette représentativité, au seuil de signification de 0,05, en utilisant les statistiques suivantes.

Répartition des étudiants de l'échantillon selon le cycle d'études

Cycle d'études	1 ^{er} cycle	2 ^e cycle	3 ^e cycle	Total
Effectifs	78,5 %	16,5 %	5,0 %	100,0 %

Répartition des étudiants universitaires selon le cycle d'études, Québec, 2011

Cycle d'études	1 ^{er} cycle	2 ^e cycle	3 ^e cycle	Total
Effectifs	76,9 %	18,0 %	5,1 %	100,0 %

Source: Ministère de l'Éducation, du Loisir et du Sport, et Ministère de l'Enseignement supérieur. *Statistiques de l'éducation, édition 2011*, 2013.

8. On se demande si la distribution du nombre de filles dans les familles de 4 enfants suit bien une loi binomiale $B(4; 0,5)$. Pour tester cette hypothèse, on prélève un échantillon aléatoire de 1 000 familles de 4 enfants et l'on compte le nombre de filles dans chaque famille. Effectuer le test d'hypothèse au seuil de signification de 0,01. (*On trouve la distribution des fréquences théoriques pour la $B(4; 0,5)$ dans la table de la loi binomiale à la page 337.*)

Répartition des familles de l'échantillon selon le nombre de filles

Nombre de filles	0	1	2	3	4	Total
Nombre de familles	54	283	383	232	48	1 000

9. Une entreprise qui fabrique des tee-shirts émet l'hypothèse que la distribution du nombre de défauts par vêtement suit une loi de Poisson où $\lambda = 0,5$. Tester cette hypothèse, au seuil de signification de 0,05, en utilisant les données d'un échantillon de 900 tee-shirts prélevés au hasard dans la production. (*On trouve la distribution des fréquences théoriques dans la table de la loi de Poisson à la page 342.*)

Répartition des tee-shirts de l'échantillon selon le nombre de défauts

Nombre de défauts	0	1	2	3	4 et plus	Total
Nombre de tee-shirts	568	244	77	6	5	900

10. Afin d'étudier les conditions de vie des personnes âgées habitant à Montréal, on prélève un échantillon aléatoire de 1 200 personnes de 70 ans et plus. Au seuil de signification de 0,05, peut-on affirmer que la distribution de l'âge dans l'échantillon est représentative de celle des Montréalais de ce groupe d'âge ?

Âge (en ans)	Échantillon	Population ¹
De 70 à 74	343	65 715
De 75 à 79	352	57 880
De 80 à 84	232	47 370
85 et plus	273	44 305
Total	1 200	215 270

1. Ville de Montréal. *Profil sociodémographique, Agglomération de Montréal*, édition mai 2014.

11. Dans le but de mettre sur pied un programme de recyclage, une entreprise mène une étude auprès d'un échantillon aléatoire de 180 employés. Ces derniers utilisent en moyenne 52 kg de papier par année avec un écart type corrigé de 11,2 kg. La distribution de la quantité de papier qu'ils utilisent est donnée dans le tableau suivant. Les données échantillonnes permettent-elles d'affirmer que la distribution de la quantité annuelle de papier utilisé par l'ensemble des employés de l'entreprise suit

une loi normale ? Faire un test au seuil de signification de 0,05 (certains effectifs théoriques sont donnés dans le tableau ci-dessous).

Répartition des employés de l'échantillon selon la quantité de papier utilisé en un an

Quantité de papier (en kg)	Effectifs observés	Effectifs théoriques
$20 \leq X < 30$	6	4,5
$30 \leq X < 40$	18	21,1
$40 \leq X < 50$	50	—
$50 \leq X < 60$	64	—
$60 \leq X < 70$	34	33,3
$70 \leq X < 80$	8	—
Total	180	180

12. Le diamètre moyen pour un échantillon aléatoire de 200 tiges produites par une machine est de 1,20 cm avec un écart type corrigé de 0,05 cm. La distribution du diamètre des tiges est présentée

dans le tableau. Les données échantillonnelles permettent-elles d'affirmer que la distribution du diamètre pour l'ensemble des tiges de la production suit un modèle normal ? Faire un test au seuil de signification de 0,05 (certains effectifs théoriques sont donnés dans le tableau ci-dessous).

Répartition des tiges de l'échantillon selon le diamètre

Diamètre (en cm)	Effectifs observés	Effectifs théoriques
$X < 1,12$	7	11,0
$1,12 \leq X < 1,15$	20	—
$1,15 \leq X < 1,18$	46	37,2
$1,18 \leq X < 1,21$	54	—
$1,21 \leq X < 1,24$	37	41,8
$1,24 \leq X < 1,27$	24	26,2
$X \geq 1,27$	12	—
Total	200	200,1

6.2 Le test d'indépendance du khi-deux

Beaucoup d'études visent à déterminer s'il existe un lien entre deux variables, par exemple :

- entre l'argent investi en publicité et le volume des ventes ;
- entre le niveau de scolarité et le revenu ;
- entre le taux de cholestérol dans le sang et les risques de maladies cardiovasculaires.

Le test d'indépendance du khi-deux sert à vérifier s'il existe effectivement un lien entre deux variables.

RAPPEL

Nous avons vu au chapitre 2 que deux événements A et B sont indépendants si la probabilité de réalisation de l'événement A n'est pas modifiée par la réalisation de l'événement B :

$$A \text{ et } B \text{ sont indépendants si } P(A | B) = P(A)$$

MISE EN SITUATION

Dans le cadre d'une étude portant sur la discrimination salariale entre les hommes et les femmes, des chercheurs prélevent un échantillon de 500 personnes travaillant dans l'industrie du textile. Le tableau suivant, que l'on appelle **tableau de contingence** ou **tableau à double entrée**, donne la distribution des travailleurs de l'échantillon selon le salaire et le sexe.

Distribution observée

Répartition des travailleurs de l'échantillon selon le salaire et le sexe

Sexe	Salaire			Total
	Bas (moins de 25 000 \$)	Moyen (25 000 \$ à 40 000 \$)	Élevé (40 000 \$ et plus)	
Femmes	92	154	54	300
Hommes	58	96	46	200
Total	150	250	100	500

❓ Globalement, si l'on ne tient pas compte du sexe, quel pourcentage des travailleurs de l'échantillon ont un salaire bas ? moyen ? élevé ?

Pourcentage	Salaire			Total
	Bas	Moyen	Élevé	
Pourcentage	$P(B) =$	$P(M) =$	$P(E) =$	100 %

❓ S'il n'y a pas de lien entre le salaire et le sexe du travailleur, théoriquement :

- quel pourcentage des femmes de l'échantillon devrait-on trouver à chaque échelon salarial ?
- quel pourcentage des hommes de l'échantillon devrait-on trouver à chaque échelon salarial ?

Distribution théorique (en pourcentages)

Répartition théorique des travailleurs de l'échantillon, par sexe, selon le salaire

Sexe	Salaire			Total
	Bas	Moyen	Élevé	
Femmes	$P(B F) =$	$P(M F) =$	$P(E F) =$	100 %
Hommes	$P(B H) =$	$P(M H) =$	$P(E H) =$	100 %

S'il n'y a pas de lien entre le salaire et le sexe, l'événement B et les événements F et H sont indépendants, d'où les égalités suivantes :

$$P(B | F) = P(B) = 30 \% \quad P(B | H) = P(B) = 30 \%$$

On peut tenir le même raisonnement pour les autres catégories de salaire.

❓ S'il n'y a pas de lien entre le salaire et le sexe du travailleur, théoriquement :

- combien devrait-il y avoir de femmes, parmi les 300 femmes, à chaque échelon salarial ?
- combien devrait-il y avoir d'hommes, parmi les 200 hommes, à chaque échelon salarial ?

Distribution théorique (en effectifs)

Répartition théorique des travailleurs de l'échantillon, par sexe, selon le salaire

Sexe	Salaire			Total
	Bas	Moyen	Élevé	
Femmes				300
Hommes				200

Théoriquement, si la répartition des 300 femmes et des 200 hommes de l'échantillon selon le salaire est semblable à celle du tableau précédent, on peut affirmer qu'il n'y a pas de discrimination salariale entre

les hommes et les femmes. Statistiquement, on peut alors dire que les variables « sexe » et « salaire » sont **indépendantes**, ou encore qu'il n'y a pas de lien entre le salaire et le sexe dans l'industrie du textile. Sinon, on dit que les variables sont **dépendantes**.

En comparant les tableaux des distributions observée et théorique, on constate un écart entre les effectifs observés dans l'échantillon et les effectifs théoriques attendus dans le cas où les variables sont indépendantes. Doit-on en conclure qu'il y a un lien de dépendance entre les variables ? La différence est peut-être entièrement attribuable au hasard de l'échantillonnage : un autre échantillon donnerait-il un meilleur ajustement entre les deux types d'effectifs ?

Il serait utopique de penser obtenir un ajustement parfait entre les effectifs observés et les effectifs théoriques. Il y aura toujours un léger écart dû au hasard de l'échantillonnage ; la question est de savoir à partir de quelle valeur il est jugé trop grand pour qu'on l'attribue entièrement au hasard. Le test d'indépendance du khi-deux sert à établir une règle de décision permettant de trancher la question.

6.2.1 La construction d'un test d'indépendance du khi-deux

Voici la marche à suivre pour construire un test d'hypothèse permettant de déterminer s'il existe une dépendance entre le sexe et le salaire dans l'industrie du textile.

Étape 1. Formulation des hypothèses

H_0 : Les variables Sexe et Salaire sont indépendantes.

H_1 : Les variables Sexe et Salaire sont dépendantes.

On formule toujours l'hypothèse nulle H_0 et l'hypothèse alternative H_1 sous cette forme. Comme pour tout autre test d'hypothèse, on considère que H_0 est vraie jusqu'à preuve du contraire.

Étape 2. Calcul des effectifs théoriques selon l'hypothèse nulle

Il s'agit de construire le tableau des effectifs théoriques que l'on devrait obtenir si l'on considère que l'hypothèse nulle est vraie, c'est-à-dire si les variables sont indépendantes.

Distribution théorique

Sexe	Salaire			Total
	Bas	Moyen	Élevé	
Femmes	90	150	60	300
Hommes	60	100	40	200
Total (% T) ¹	150 (30 %)	250 (50 %)	100 (20 %)	500 (100 %)

1. % T : pourcentage théorique.

On applique le raisonnement suivant pour construire ce tableau :

- Parmi les 500 personnes de l'échantillon, 150 ont un salaire bas, soit 30 %.
Si H_0 est vraie, 30 % des femmes et 30 % des hommes auront un salaire bas.
$$30 \% \times 300 \text{ femmes} = 90 \text{ femmes} \quad 30 \% \times 200 \text{ hommes} = 60 \text{ hommes}$$
- Parmi les 500 personnes de l'échantillon, 250 ont un salaire moyen, soit 50 %.
Si H_0 est vraie, 50 % des femmes et 50 % des hommes auront un salaire moyen.
$$50 \% \times 300 \text{ femmes} = 150 \text{ femmes} \quad 50 \% \times 200 \text{ hommes} = 100 \text{ hommes}$$
- Parmi les 500 personnes de l'échantillon, 100 ont un salaire élevé, soit 20 %.
Si H_0 est vraie, 20 % des femmes et 20 % des hommes auront un salaire élevé.
$$20 \% \times 300 \text{ femmes} = 60 \text{ femmes} \quad 20 \% \times 200 \text{ hommes} = 40 \text{ hommes}$$

Condition d'application du test d'indépendance du khi-deux

Tout comme les autres tests du khi-deux, la condition d'application du test d'indépendance exige que les effectifs théoriques soient tous supérieurs ou égaux à 5. Si cette condition n'est pas respectée, il faut alors regrouper des catégories adjacentes de manière qu'elle le soit. (*Pour un exemple, voir le numéro 5 des exercices 6.2.*)

Étape 3. Calcul du khi-deux

Il faut maintenant comparer les effectifs théoriques (T) et les effectifs observés (O) et mesurer les différences entre ces effectifs. Pour faciliter la comparaison, on place côté à côté les deux types d'effectifs dans un tableau.

Comparaison des distributions théorique et observée				
Sexe	Salaire			Total
	Bas	Moyen	Élevé	
Femmes	92 90	154 150	54 60	300
Hommes	58 60	96 100	46 40	200
Total (% T)	150 (30 %)	250 (50 %)	100 (20 %)	500 (100 %)

Tous les effectifs théoriques étant supérieurs ou égaux à 5, on mesure l'ajustement entre les deux distributions en calculant la valeur du khi-deux.

Valeur du khi-deux

$$\chi^2 = \sum \frac{(O - T)^2}{T}$$

$$\chi^2 = \frac{(92 - 90)^2}{90} + \frac{(58 - 60)^2}{60} + \frac{(154 - 150)^2}{150} + \frac{(96 - 100)^2}{100} + \frac{(54 - 60)^2}{60} + \frac{(46 - 40)^2}{40} = 1,9$$

Étape 4. Énoncé de la règle de décision

Tout comme nous l'avons fait pour les tests d'ajustement, nous décidons de rejeter ou non l'hypothèse nulle en comparant la valeur du χ^2 à celle du point critique χ^2_c . Rappelons que la valeur du point critique dépend du seuil de signification du test (α) et du nombre de degrés de liberté (dl).

Nombre de degrés de liberté

Pour un test d'indépendance entre les variables X et Y , le nombre de degrés de liberté se calcule ainsi :

Nombre de degrés de liberté pour un test d'indépendance

$$dl = (\text{nombre de catégories de la variable } X - 1) \times (\text{nombre de catégories de la variable } Y - 1)$$



Pour la mise en situation, le nombre de degrés de liberté $dl =$ _____

Ce dernier résultat signifie qu'il faut un nombre minimal de 2 données pour être en mesure de retrouver les autres données du tableau si les totaux des lignes et des colonnes sont connus. On a donc 2 données libres et 4 données liées.

Règle de décision

La règle de décision d'un test d'indépendance est identique à celle d'un test d'ajustement :

Règle de décision d'un test d'indépendance

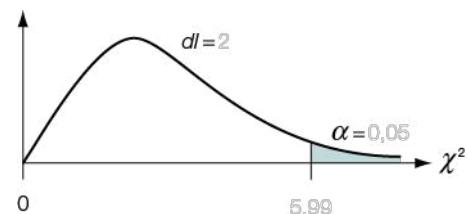
Rejeter H_0 si la valeur du khi-deux est supérieure au khi-deux critique.

Rejeter H_0 si $\chi^2 > \chi_c^2$.

- ❸ Énoncer la règle de décision qui s'applique à la mise en situation, si l'on utilise un seuil de signification de 0,05.

Pour $\alpha = 0,05$ et $dl = 2$, on a $\chi_c^2 = \underline{\hspace{2cm}}$.

Règle de décision



Étape 5. Décision et conclusion

- ❹ Pour la mise en situation, doit-on rejeter ou ne pas rejeter l'hypothèse nulle ?

Comme , on H_0 .

Statistiquement, rien ne permet de croire qu'il existe une discrimination salariale entre les hommes et les femmes dans l'industrie du textile. Le salaire est vraisemblablement indépendant du sexe du travailleur. La différence d'ajustement entre les distributions observée et théorique est due au hasard de l'échantillonnage.

NOTE

Il est important de ne pas inverser l'énonciation des hypothèses H_0 et H_1 . Si, par manque de compréhension de la logique du test, on pose comme hypothèse pour H_0 que les variables sont dépendantes, le non-rejet de cette hypothèse en conclusion conduit à déclarer qu'il existe une discrimination salariale par rapport au sexe dans l'industrie du textile, ce qui est une affirmation lourde de conséquences.

Par ailleurs, lorsqu'on utilise un logiciel statistique (Excel, par exemple) pour effectuer un test du khi-deux, il arrive fréquemment que le logiciel affiche la surface sous la courbe située à droite de la valeur calculée du khi-deux. Dans ce cas, la règle de décision du test s'énonce comme suit pour un seuil de signification de 5 % : « rejeter H_0 si la surface à droite du khi-deux calculé est plus petite que 5 %. »

Dépendance et causalité

Les statistiques permettent d'établir qu'il y a un lien entre deux variables, mais elles ne prouvent jamais qu'il existe un lien de cause à effet, c'est-à-dire qu'une des deux variables est la cause et l'autre, l'effet. Il peut arriver cependant que la dépendance entre deux variables soit attribuable à l'effet simultané d'une troisième variable sur les deux premières.

Les statistiques ont servi à établir qu'il y a une relation entre le cancer du poumon et le tabagisme, mais elles ne prouvent pas que le tabac est la cause principale du cancer du poumon. Ce sont des recherches médicales qui ont démontré que la fumée de tabac est la variable la plus importante dans le développement du cancer. (Pour un exemple, voir le numéro 6 des exercices 6.2.)

EXEMPLE

Les compagnies d'assurance automobile justifient des primes plus élevées pour les jeunes conducteurs par le risque d'accident² jugé plus grand chez ces personnes. Pour vérifier cette hypothèse, une étude est menée auprès d'un échantillon aléatoire de 2 000 conducteurs québécois. Les données obtenues permettent-elles de conclure qu'il y a un lien entre l'âge du conducteur et le risque d'accident? Tester l'hypothèse au seuil de signification de 0,01.

Répartition des conducteurs de l'échantillon selon l'âge et l'implication dans un accident au cours des 12 derniers mois

Âge du conducteur	Implication dans un accident		
	Oui	Non	Total
De 16 ans à 24 ans	14	194	208
De 25 ans à 44 ans	22	652	674
De 45 ans à 64 ans	18	782	800
65 ans et plus	6	312	318
Total	60	1 940	2 000

Source: Société de l'assurance automobile du Québec. *Bilan routier 2012*, juillet 2013.

Solution

1. Formulation des hypothèses

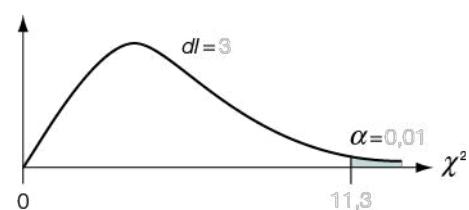
2. Calcul des effectifs théoriques selon l'hypothèse nulle

O T	Implication dans un accident		
	Oui	Non	Total
Âge du conducteur			
De 16 ans à 24 ans	14	194	208
De 25 ans à 44 ans	22	652	674
De 45 ans à 64 ans	18 24,0	782 776,0	800
65 ans et plus	6 9,5	312 308,5	318
Total (% T)	60 ()	1 940 ()	2 000

La condition d'application du test du khi-deux est-elle respectée ? _____

3. Calcul du khi-deux

4. Règle de décision



2. Le risque d'accident correspond au pourcentage de conducteurs impliqués dans un accident.

5. Décision et conclusion

Nature de la dépendance

Lorsqu'il y a un lien entre deux variables, on décrit la nature de la dépendance en faisant ressortir les écarts entre les pourcentages théoriques et observés. Si une tendance se dégage des pourcentages observés, on l'énonce. Nous avons déjà fait ce type d'analyse en probabilité pour interpréter le lien entre deux événements dépendants.

EXEMPLE (suite)

Le texte suivant décrit la nature de la dépendance entre l'âge du conducteur et le risque d'accident.

Globalement, si l'on ne tient pas compte de l'âge du conducteur, le risque d'accident est de 3 % ($60 \div 2\,000$). Or, si l'on tient compte de l'âge du conducteur, le risque d'accident n'est plus le même. Il est de plus du double, soit 6,7 % ($14 \div 208$), chez les conducteurs âgés de 16 à 24 ans, alors que pour les trois autres classes d'âge, de 25 à 44 ans, de 45 à 64 ans et 65 ans et plus, le risque d'accident est de 3,3 %, de 2,3 % et de 1,9 % respectivement.

Une tendance se dégage de ces statistiques : plus le conducteur est jeune, plus le risque d'accident est élevé.

EXERCICE DE COMPRÉHENSION | 6.3

Un sondage est effectué pour connaître les habitudes de lecture des Québécois de 15 ans et plus. Parmi les répondants, 64 % des femmes et 46 % des hommes ont déclaré avoir lu au moins un livre au cours des 12 derniers mois. Le tableau suivant présente la répartition des 800 lecteurs de livres de l'échantillon en fonction du sexe et du nombre de livres lus en un an.

O T Sexe	Nombre de livres lus				Total
	De 1 à 4	De 5 à 9	De 10 à 19	20 et plus	
Femmes	143	105	134	150	532
Hommes	104	52	54	58	268
Total (% T)	247 ()	157 ()	188 ()	208 ()	800

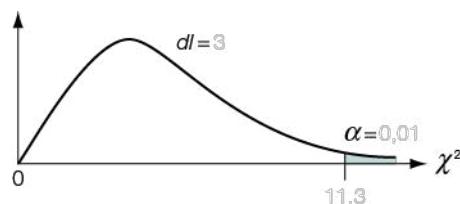
Source: Ministère de la Culture et des Communications. *Les pratiques culturelles au Québec en 2009*, avril 2011.

- a) S'il n'y a pas de lien entre le sexe et le nombre de livres lus chez les lecteurs de livres, quel pourcentage d'hommes devraient lire de 1 à 4 livres par année ?

- b) Faire un test d'indépendance du khi-deux, au seuil de signification de 1 %, pour vérifier s'il existe un lien entre le sexe et le nombre de livres lus chez les lecteurs de livres.
 1. Formulation des hypothèses

- 2. Calcul des effectifs théoriques selon l'hypothèse nulle
 (Inscrire les résultats dans le tableau de la page précédente.)
 La condition d'application du test du khi-deux est-elle respectée ? _____
3. Calcul du khi-deux

4. Règle de décision



5. Décision et conclusion

Nature de la dépendance

Compléter l'analyse suivante.

Globalement, _____ % des lecteurs de livres lisent 10 livres ou plus par année. Or, si l'on tient compte du sexe, ce pourcentage augmente à _____ % chez les femmes et baisse à _____ % chez les hommes.

EXERCICES 6.2

1. On se demande si la performance d'une équipe de soccer est la même lorsqu'elle joue à domicile et lorsqu'elle joue à l'extérieur. On décide donc d'étudier, à l'aide d'un test du khi-deux, s'il y a dépendance entre les variables «lieu où se joue la partie» et «résultat de la partie» avec un échantillon aléatoire de 36 parties.

Résultat de la partie				
Lieu	Victoire	Défaite	Nulle	Total
À domicile				16
À l'extérieur				20
Total	14	18	4	36

- a) Déterminer le pourcentage de parties gagnées par l'équipe au cours des 36 parties.

- b) Si l'on pose comme hypothèse H_0 que les variables sont indépendantes, c'est-à-dire que l'on pense qu'il n'y a pas de lien entre le résultat de la partie et le lieu où se joue celle-ci :
- i) Quel devrait être, sous cette hypothèse, le nombre de parties gagnées à domicile ?
 - ii) Quel devrait être, sous cette hypothèse, le nombre de parties gagnées à l'extérieur ?
- c) Si l'on pose comme hypothèse H_0 que les deux variables sont dépendantes, c'est-à-dire que l'on pense qu'il existe un lien entre le résultat de la partie et le lieu où se joue celle-ci :
- i) Quel devrait être, sous cette hypothèse, le nombre de parties gagnées à domicile ?
 - ii) Quel devrait être, sous cette hypothèse, le nombre de parties gagnées à l'extérieur ?

2. Plusieurs journaux quotidiens sont offerts en version papier et sur Internet. Quelles sont les caractéristiques des personnes qui préfèrent utiliser Internet pour lire des quotidiens ? Pour répondre à cette question, un sondage est effectué auprès d'un échantillon aléatoire de 200 lecteurs de quotidiens. Le tableau suivant présente des statistiques sur l'âge des lecteurs.

Répartition des lecteurs de quotidiens de l'échantillon selon l'âge et le fait qu'ils lisent ou non des quotidiens sur Internet

Âge (en ans)	Lecture de quotidiens sur Internet		
	Oui	Non	Total
34 et moins	27	16	43
De 35 à 54	42	33	75
55 et plus	28	54	82
Total	97	103	200

Source: Ministère de la Culture et des Communications. *Survol: Les pratiques de lecture des Québécois et Québécoises de 2004 à 2009*, n° 24, décembre 2012.

- a) Au seuil de signification de 1 %, peut-on dire qu'il existe un lien entre l'âge et la lecture de quotidiens sur Internet ? S'il y a dépendance, en donner la nature.
- b) S'il n'y avait pas de lien entre l'âge et la lecture de quotidiens sur Internet, quelle proportion de lecteurs de 34 ans ou moins utiliseraient Internet pour lire des quotidiens ? Cette proportion serait-elle différente chez les lecteurs de 55 ans et plus ?
3. Des données du sondage présenté au numéro 2, on a tiré les statistiques suivantes.

Répartition des lecteurs de quotidiens de l'échantillon selon le niveau de scolarité et le fait qu'ils lisent ou non des quotidiens sur Internet

Niveau de scolarité	Lecture de quotidiens sur Internet		
	Oui	Non	Total
Primaire	3	8	11
Secondaire	23	43	66
Collégial	29	28	57
Universitaire	42	24	66
Total	97	103	200

Source: Ministère de la Culture et des Communications. *Survol: Les pratiques de lecture des Québécois et Québécoises de 2004 à 2009*, n° 24, décembre 2012.

- a) Le tableau indique que seulement 3 répondants lisent des quotidiens sur Internet et ont un niveau de scolarité primaire. Doit-on en conclure que la condition d'application du test du khi-deux n'est pas respectée ?

- b) Au seuil de signification de 1 %, les statistiques du tableau permettent-elles d'affirmer qu'il existe un lien entre le niveau de scolarité et la lecture de quotidiens sur Internet ? S'il y a un lien, en donner la nature.

4. Compléter les tableaux suivants, si l'on pose pour hypothèse que le salaire est indépendant du sexe de l'employé.

a)

Sexe	Salaire hebdomadaire			Total
	Moins de 600 \$	600 \$ et plus		
Femmes				100 %
Hommes				100 %
Total	40 %	60 %		100 %

b)

Sexe	Revenu annuel			Total
	Bas	Moyen	Élevé	
Femmes				200
Hommes				
Total	150	225		500

5. Chaque année, l'événement *Défi escaliers de Québec* convie les amateurs à parcourir près de 30 escaliers du Vieux-Québec, en alternant entre un escalier à monter et le suivant à descendre, ce qui totalise plus de 3 000 marches sur 11 kilomètres. On prélève un échantillon aléatoire de 80 coureurs parmi les participants afin de vérifier s'il existe un lien entre l'âge et le temps de course.

Répartition des coureurs de l'échantillon selon l'âge et le temps de course

Âge (en ans)	Temps de course (en min)				Total
	Moins de 75	[75; 90[[90; 105[105 et plus	
Moins de 30	7	12	4	2	25
De 30 à 39	8	7	5	2	22
40 et plus	4	12	9	8	33
Total	19	31	18	12	80

Source: Échantillon prélevé parmi les participants au Défi des escaliers de Québec 2011.

- a) Pour chacune des catégories de coureurs suivantes, dire si la condition d'application du test du khi-deux est respectée.
- Les coureurs de moins de 30 ans qui ont pris de 90 à 105 minutes pour faire le parcours.
 - Les coureurs de 30 à 39 ans qui ont pris 105 minutes ou plus pour faire le parcours.
- b) Pour satisfaire la condition d'application du test du khi-deux, construire un nouveau tableau en groupant les catégories [90; 105[et 105 et plus, puis effectuer le test. Peut-on conclure, au seuil de signification de 0,05, qu'il existe un lien entre l'âge et le temps de course ?

6. Des chercheurs de l'Université de Chicago ont effectué une enquête auprès d'un échantillon aléatoire de 776 États-Uniens afin d'étudier leur attitude face à l'avortement. On a défini trois attitudes possibles : pour, contre ou mixte (les gens pour ou contre dans certains cas seulement). Le tableau donne les effectifs observés et théoriques du test du khi-deux visant à déterminer s'il y a un lien entre la scolarité et l'attitude face à l'avortement.

Répartition des répondants selon la scolarité et l'attitude face à l'avortement

Scolarité (en années)	Attitude face à l'avortement			Total
	Pour	Mixte	Contre	
Moins de 9	31 45,1	23 21,5	56 43,6	110
De 9 à 12	171 179,2	89 85,2	177 173,1	437
Plus de 12	116 93,9	39 44,7	74 90,7	229
Total (% T)	318 (41,0 %)	151 (19,5 %)	307 (39,6 %)	776

Source: Alalouf, Labelle et Ménard. *Introduction à la statistique appliquée*, 2^e éd., Montréal, Addison-Wesley, 1990, 412 p.

- a) La valeur du khi-deux calculée à partir des données du tableau est 17,7. Au seuil de signification

de 5 %, cette valeur permet-elle de conclure que l'attitude face à l'avortement dépend de la scolarité ? S'il y a dépendance, en donner la nature.

- b) L'enquête établit qu'il existe un lien entre deux variables, mais elle ne prouve pas qu'il y a un lien de cause à effet, c'est-à-dire qu'il y a une variable qui est la cause et l'autre, l'effet. Il se peut que la dépendance entre ces deux variables soit attribuable à l'effet simultané d'une troisième variable sur les deux premières (appartenance à une religion, âge, sexe, origine ethnique, etc.). Pour ce qui a trait à l'attitude relative à l'avortement, après analyse de l'échantillon, on a constaté qu'il était composé de 57 % de protestants et de 43 % de catholiques.

On a donc décidé de reprendre le test en séparant ces deux groupes. Voici le résultat que l'on a obtenu pour le khi-deux de chaque groupe :

Catholiques : $\chi^2 = 4,8$

Protestants : $\chi^2 = 17,1$

Pour chacun des deux groupes, dire si l'on doit maintenir la conclusion tirée en a).

Démarche pour construire un test du khi-deux

Étape 1 Formuler les hypothèses H_0 et H_1 du test.

Pour un test d'ajustement à une distribution spécifiée

H_0 : La distribution de la variable dans la population suit une distribution spécifiée.

H_1 : La distribution de la variable dans la population ne suit pas une distribution spécifiée.

Pour la représentativité d'un échantillon

H_0 : L'échantillon est représentatif de la population pour une variable spécifiée.

H_1 : L'échantillon n'est pas représentatif de la population pour une variable spécifiée.

Pour un test d'indépendance

H_0 : Les variables spécifiées sont indépendantes.

H_1 : Les variables spécifiées sont dépendantes.

Étape 2 Calculer les effectifs théoriques selon l'hypothèse nulle H_0 .

Vérifier si les effectifs théoriques sont tous supérieurs ou égaux à 5.

Étape 3 Calculer la valeur du khi-deux.

$$\chi^2 = \sum \frac{(O - T)^2}{T}$$

Étape 4 • Calculer le nombre de degrés de liberté pour le seuil de signification souhaité.

Nombre de degrés de liberté (dl)

– Test d'ajustement et de représentativité :

n^{bre} de catégories – 1

– Test d'ajustement à une loi normale :

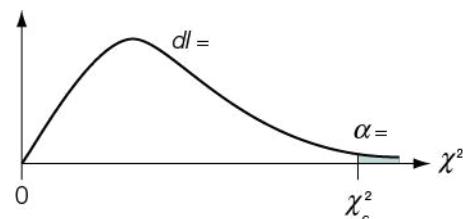
n^{bre} de classes – n^{bre} de paramètres estimés – 1

– Test d'indépendance :

$(n^{bre}$ de catégories pour $X - 1) \times (n^{bre}$ de catégories pour $Y - 1)$

• Énoncer la règle de décision.

Rejeter H_0 si $\chi^2 > \chi^2_c$.



Étape 5 Décider et conclure.

Note : Dans un test d'indépendance, il faut préciser la nature de la dépendance s'il y a lieu.

EXERCICES RÉCAPITULATIFS

1. Afin d'étudier les caractéristiques des accidents routiers, on prélève un échantillon aléatoire de 800 accidents parmi les 104 070 accidents dénombrés au Québec en 2012.

Source: Société de l'assurance automobile du Québec. *Dossier statistique – Bilan 2011, accidents, parc automobile, permis de conduire*, juin 2012.

- a) Des 800 accidents de l'échantillon, 212 ont causé des dommages corporels et 588 ont causé des dommages matériels. En 2012, on a dénombré 29 496 accidents ayant causé des dommages corporels et 74 574 ayant causé des dommages matériels. Au seuil de signification de 0,05, peut-on considérer que l'échantillon est représentatif de l'ensemble des accidents pour la nature des dommages ?
- b) Une analyse des données échantillonnelles a permis de construire le tableau suivant.

Répartition des 800 accidents selon le jour où l'accident s'est produit

Jour	Nombre d'accidents
Lundi	110
Mardi	108
Mercredi	114
Jeudi	126
Vendredi	141
Samedi	109
Dimanche	92
Total	800

- i) Les données échantillonnelles permettent-elles d'affirmer que, toutes proportions gardées, il y a autant d'accidents en semaine (du lundi au vendredi) qu'en fin de semaine (samedi et dimanche) ? Faire le test au seuil de signification de 0,05.
- ii) Au seuil de signification de 5 %, tester l'hypothèse voulant que les accidents se répartissent uniformément du lundi au vendredi.
2. Une enquête auprès d'un échantillon aléatoire d'entreprises révèle que 15,5 % des 990 entreprises branchées à Internet ont une connexion très haute vitesse (100 mégabits par seconde). En considérant les données du tableau et un seuil de signification de 0,01, peut-on dire que le taux de branchement à Internet très haute vitesse dépend de la taille de l'entreprise ? S'il y a dépendance, en donner la nature.

Répartition des entreprises de l'échantillon branchées à Internet selon la taille et l'utilisation d'une connexion très haute vitesse

Nombre d'employés	Connexion très haute vitesse		Total
	Oui	Non	
De 1 à 9	39	311	350
De 10 à 49	40	280	320
De 50 à 249	39	171	210
250 et plus	35	75	110
Total	153	837	990

Source: Institut de la statistique du Québec. *Enquête sur l'intégration d'Internet aux processus d'affaires*, 2012.

3. Une étude visant à dresser le portrait de l'entrepreneuriat au Québec à l'aide d'un échantillon aléatoire de 1 000 entrepreneurs révèle les statistiques suivantes.

Répartition des entrepreneurs de l'échantillon selon l'âge

Âge	Nombre d'entrepreneurs
Moins de 25 ans	12
[25 ans ; 35 ans[129
[35 ans ; 45 ans[284
[45 ans ; 55 ans[323
[55 ans ; 65 ans[204
65 ans et plus	48
Total	1 000

Source: Ministère de l'Économie, de l'Innovation et des Exportations. *Le renouvellement de l'entrepreneuriat au Québec : un regard sur 2013 et 2018*, 2010.

Un test est effectué pour vérifier si la distribution de l'âge dans la population des entrepreneurs suit une loi normale.

- a) Donner l'hypothèse nulle du test.
- b) Sachant que la moyenne et l'écart type corrigé de l'âge dans l'échantillon sont respectivement 47,2 ans et 11,1 ans, donner les effectifs théoriques des deux classes suivantes :
- i) [35 ans ; 45 ans[.
 - ii) [45 ans ; 55 ans[.
- c) Énoncer la règle de décision du test au seuil de signification de 0,01.
- d) Les données échantillonnelles permettent-elles de conclure que la distribution de l'âge des entrepreneurs suit une loi normale si la valeur du khi-deux est 10,4 ?

PRÉPARATION À L'EXAMEN

Pour préparer votre examen, assurez-vous d'avoir les compétences suivantes :

Test d'ajustement du khi-deux

Appliquer la marche à suivre pour construire un test d'ajustement à une distribution spécifiée:

- Formuler les hypothèses;
- Calculer les effectifs théoriques et vérifier la condition d'application;
- Calculer la valeur du khi-deux;
- Énoncer la règle de décision;
- Décider et conclure en tenant compte du contexte.

Si vous avez
la compétence,
cochez.

Test d'indépendance du khi-deux

Appliquer la marche à suivre pour construire un test d'indépendance:

- Formuler les hypothèses;
- Calculer les effectifs théoriques et vérifier la condition d'application;
- Calculer la valeur du khi-deux;
- Énoncer la règle de décision;
- Décider et conclure en tenant compte du contexte;
- Quand il y a dépendance entre les variables, en décrire la nature.

7

Chapitre

La corrélation et la régression linéaire

OBJECTIFS DU CHAPITRE

Mesurer et exprimer sous forme d'équation la dépendance entre deux variables quantitatives.

OBJECTIFS DU LABORATOIRE

Le laboratoire 6 vise à apprendre à utiliser Excel pour étudier la corrélation et la régression linéaire entre deux variables.



Le présent chapitre est consacré à l'étude de la dépendance entre deux variables quantitatives. Contrairement à ce que nous avons fait dans le cas où au moins une des variables était qualitative, nous ne nous limitons pas à affirmer que les variables sont dépendantes ; nous mesurons la force de cette dépendance, et même, nous exprimons la relation entre les variables sous la forme d'un modèle mathématique qui permet d'estimer, pour une valeur donnée d'une des variables, la valeur correspondante de l'autre variable.

7.1 La corrélation linéaire

MISE EN SITUATION

On veut étudier le lien entre le poids d'une personne et le taux d'alcool dans le sang. Pour ce faire, on mesure le taux d'alcool dans le sang de sept hommes et de sept femmes qui ont consommé trois bières. Voici les résultats obtenus pour chaque sexe.

Hommes

Taux d'alcool dans le sang en fonction du poids après la consommation de trois bières

X: Poids	57 kg (125 lb)	68 kg (150 lb)	80 kg (176 lb)	86 kg (190 lb)	91 kg (200 lb)	102 kg (225 lb)	114 kg (251 lb)
Y: Taux d'alcool (en mg/100 ml)	103	87	76	70	65	59	52

Source: Éduc'alcool. *Boire, conduire, choisir. L'alcool au volant. L'alcool et la loi (0.08)*, 2014.

Femmes

Taux d'alcool dans le sang en fonction du poids après la consommation de trois bières

X: Poids	45 kg (100 lb)	52 kg (115 lb)	57 kg (125 lb)	68 kg (150 lb)	73 kg (161 lb)	80 kg (176 lb)	91 kg (200 lb)
Y: Taux d'alcool (en mg/100 ml)	152	133	120	101	99	87	76

Source: Éduc'alcool. *Boire, conduire, choisir. L'alcool au volant. L'alcool et la loi (0.08)*, 2014.

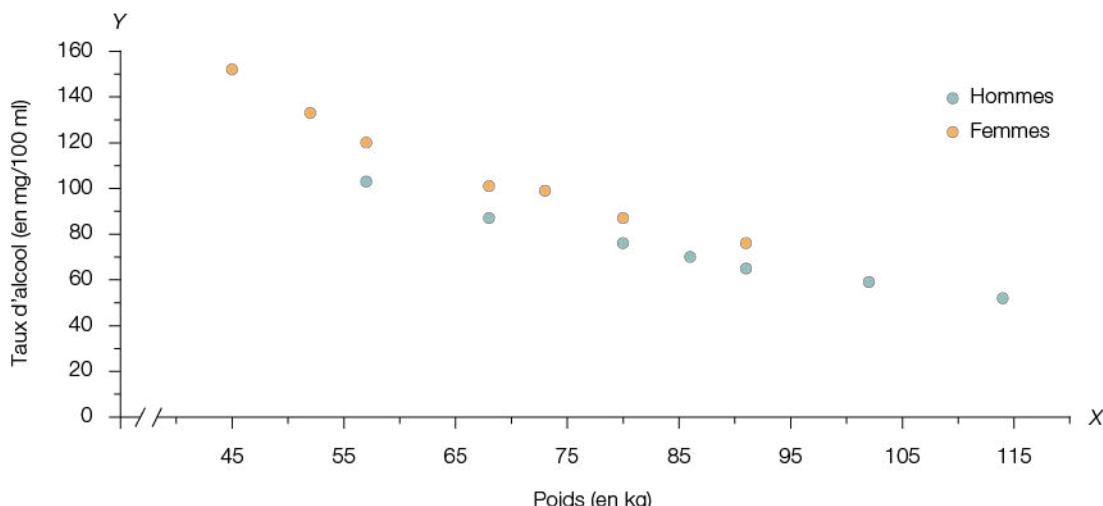
7.1.1 Le diagramme de dispersion (ou nuage de points)

On donne le nom de **diagramme de dispersion** ou **nuage de points** à la représentation graphique dans le plan cartésien de l'ensemble des paires de données (x, y) provenant de l'étude de deux variables quantitatives X et Y .

EXEMPLE

Le graphique de la page suivante présente, sur un même système d'axes, le diagramme de dispersion des données des tableaux de la mise en situation.

Taux d'alcool dans le sang en fonction du poids, par sexe, après la consommation de trois bières



Source: Éduc'alcool. Boire, conduire, choisir. L'alcool au volant. L'alcool et la loi (0.08), 2014.

Variable indépendante et variable dépendante

Il est souhaitable de respecter la convention, largement appliquée, selon laquelle on attribue la lettre X à la variable indépendante et la lettre Y à la variable dépendante (dont la valeur semble dépendre de celle de l'autre variable).

Dans la mise en situation, nous avons attribué la lettre X à la variable «poids» et la lettre Y à la variable «taux d'alcool dans le sang», car il semble que le taux d'alcool dépende du poids du consommateur.

7.1.2 La corrélation

On dira qu'il y a une **corrélation**, ou **dépendance**, entre deux variables quantitatives X et Y si elles ont généralement tendance à varier toutes deux dans le même sens ou en sens contraires. Voici ce qui caractérise une corrélation entre les variables X et Y :

La forme

Linéaire : Les points du diagramme de dispersion ont tendance à se rapprocher d'une droite. C'est ce type de corrélation que nous étudions (exemples : mise en situation et graphiques 1, 2 et 6 à la page suivante).

Non linéaire : Les points du diagramme de dispersion ont tendance à se rapprocher d'une courbe (exemples : graphiques 3 et 4).

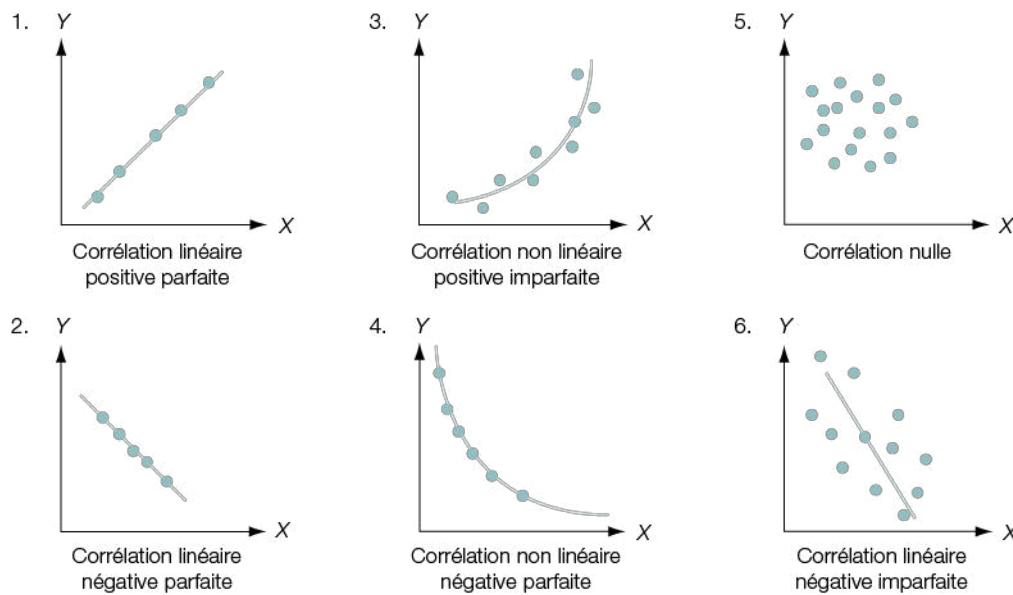
Le sens

Positif : Les deux variables varient dans le même sens : quand les valeurs de la variable X augmentent, celles de la variable Y augmentent aussi (exemples : graphiques 1 et 3).

Négatif : Les deux variables varient en sens contraires : quand les valeurs de la variable X augmentent, celles de la variable Y diminuent (exemples : mise en situation et graphiques 2, 4 et 6).

L'intensité

- Parfaite : Les points du diagramme de dispersion sont parfaitement alignés, dans le cas d'une corrélation linéaire, ou tous situés sur la courbe dans le cas d'une corrélation non linéaire. Une dépendance parfaite permet de déterminer, pour chaque valeur de la variable X , la valeur exacte de la variable Y qui lui est associée, et vice versa (exemples : graphiques 1, 2 et 4).
- Imparfaite : On constate une tendance moins forte des points du diagramme de dispersion à s'aligner ou à prendre la forme d'une courbe. Dans ce cas, on peut tout au plus estimer approximativement la valeur de la variable Y correspondant à une valeur donnée de la variable X (exemples : mise en situation et graphiques 3 et 6).
- Nulle : On ne peut dégager aucune tendance des points à s'approcher d'une droite ou d'une courbe. On dit que les deux variables sont **indépendantes**. Il est donc impossible d'estimer la valeur de la variable Y correspondant à une valeur donnée de la variable X , et vice versa (exemple : graphique 5).



7.1.3 Le coefficient de corrélation linéaire

Le diagramme de dispersion permet une analyse qualitative de la tendance à une relation linéaire entre les variables X et Y . Le **coefficient de corrélation linéaire**, ou **coefficient de Pearson**, permet de mesurer quantitativement la force de la corrélation (ou de la dépendance) linéaire entre les deux variables. On note ce coefficient par la lettre r et on le calcule à l'aide de la formule suivante :

Coefficient de corrélation linéaire

$$r = \frac{\sum xy - n \bar{x} \bar{y}}{(n-1)s_x s_y}$$

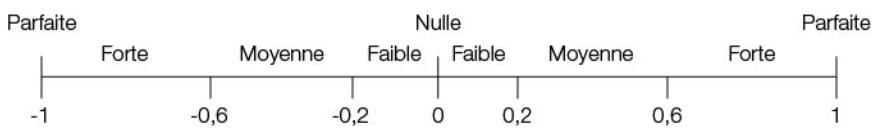
où • $\sum xy$ représente la somme des produits de chaque valeur de la variable X par la valeur correspondante de la variable Y

- n correspond au nombre de couples (x, y)

- \bar{x} représente la moyenne des valeurs de la variable X
- \bar{y} représente la moyenne des valeurs de la variable Y
- s_x est l'écart type corrigé de la variable X
- s_y est l'écart type corrigé de la variable Y

Propriétés du coefficient de corrélation linéaire

1. Le coefficient de corrélation linéaire est un nombre, sans unités, compris entre -1 et 1 : $-1 \leq r \leq 1$.
2. La corrélation linéaire est parfaite et positive pour $r = 1$, parfaite et négative pour $r = -1$ et nulle pour $r = 0$.
3. Dans le cas d'une corrélation linéaire positive, plus la valeur de r est près de 1, plus la corrélation entre X et Y est forte. Il en est de même pour une corrélation linéaire négative : plus la valeur de r est près de -1, plus la corrélation entre X et Y est forte. Le schéma suivant donne une idée de la force d'une corrélation en sciences humaines.



NOTE

Un coefficient de corrélation linéaire égal à 0 n'implique pas nécessairement que les variables X et Y sont indépendantes. La corrélation linéaire nulle indique seulement qu'il n'y a pas de dépendance linéaire entre les variables. Ces dernières pourraient néanmoins entretenir une relation de dépendance non linéaire. Seul le nuage de points peut nous garantir qu'il n'y a aucune autre forme de dépendance entre les variables X et Y . (Pour un exemple, voir le numéro 5 des exercices 7.1.)

EXEMPLE 1

Dans chaque cas, représenter la variable indépendante par X et la variable dépendante par Y . Dire si la corrélation entre les deux variables est positive ou négative.

a) _____ : Coût des dommages matériels causés par un séisme

_____ : Intensité du séisme

Corrélation _____

b) _____ : Prix d'un produit

_____ : Nombre de produits vendus

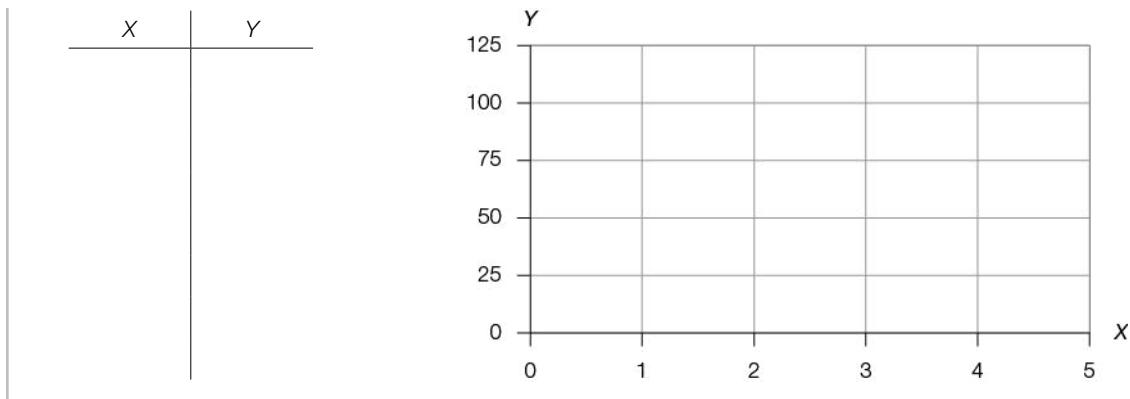
Corrélation _____

c) _____ : Nombre de billets de spectacle achetés par un client

_____ : Coût total de l'achat

Corrélation _____

Quelle est la forme du diagramme de dispersion des deux variables si l'on suppose que chaque billet coûte 25 \$? Peut-on estimer la valeur du coefficient de corrélation ? Peut-on établir une équation mathématique reliant X et Y ?



EXEMPLE 2

Dans la mise en situation, on obtient les résultats suivants pour le taux d'alcool sanguin chez les hommes. Calculer le coefficient de corrélation entre les variables X et Y et commenter. Pour obtenir une plus grande précision, conserver au moins deux décimales dans les calculs intermédiaires.

Hommes

Taux d'alcool dans le sang en fonction du poids après la consommation de trois bières

X: Poids	57 kg	68 kg	80 kg	86 kg	91 kg	102 kg	114 kg
Y: Taux d'alcool (en mg/100 ml)	103	87	76	70	65	59	52

Solution

$$\sum xy =$$

$$n = \quad \bar{x} =$$

$$s_x =$$

$$\bar{y} =$$

$$s_y =$$

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{(n-1)s_x s_y} =$$

Interprétation

Pour une même quantité d'alcool consommé, il y a une très forte corrélation linéaire entre le poids d'un homme et son taux d'alcool dans le sang. Le fait que la corrélation est négative indique la tendance suivante : plus le poids de l'homme est élevé, moins le taux d'alcool dans le sang est élevé.

Dépendance et causalité

On doit faire preuve de prudence lorsqu'on interprète les résultats. Il n'y a pas nécessairement une relation de cause à effet entre deux variables dépendantes. Une corrélation entre X et Y peut résulter de différents types de liaisons : X peut être la cause de Y , Y peut être la cause de X , les deux variables peuvent être causées par un facteur externe Z ou par un mélange de ces rapports.

On peut établir une corrélation positive entre l'âge mental d'un enfant et la longueur de ses pieds, puisque, quand une variable augmente, l'autre augmente également. On ne devrait toutefois pas en conclure que la capacité d'un enfant à résoudre des problèmes a un lien avec la longueur de ses pieds. En fait, c'est la croissance de l'enfant qui entraîne l'augmentation de ces deux variables simultanément.

7.2 La régression linéaire

Lorsqu'il existe une relation logique entre deux variables X et Y , il est intéressant de l'exprimer sous la forme d'un modèle mathématique qui sert à estimer la valeur de Y correspondant à une valeur donnée de X . C'est ce qu'on appelle l'**analyse de régression**. Une telle analyse permet, par exemple, de répondre aux questions suivantes :

- À quel prix peut-on espérer vendre une maison dont l'évaluation municipale est de 200 000 \$?
- Quelle moyenne peut espérer obtenir un étudiant, à sa première session au cégep, s'il a obtenu une moyenne de 75 % en 5^e secondaire ?

Nous limiterons notre étude de la régression à celle de type linéaire.

7.2.1 La droite de régression

Lorsque le diagramme de dispersion indique qu'il existe une corrélation linéaire entre deux variables, on exprime mathématiquement cette relation par l'équation d'une droite. On appelle droite de régression la droite qui représente le mieux le nuage de points. On considère que la droite qui s'ajuste le mieux aux points est celle pour laquelle la valeur D , égale à la somme des carrés des écarts entre chaque point du diagramme de dispersion et la droite, est minimale.

$$D = d_1^2 + d_2^2 + d_3^2 + \dots + d_n^2$$

À l'aide de cette méthode, que l'on appelle **méthode des moindres carrés**, on en arrive à trouver la pente de la droite, notée b , et son ordonnée à l'origine, notée a , ce qui nous donne l'équation suivante :

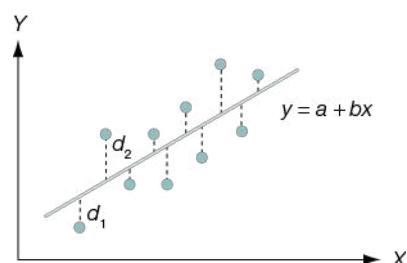
Équation de la droite de régression

$$y = a + bx$$

On calcule les valeurs de a et de b ainsi :

$$a = \bar{y} - b \bar{x}$$

$$b = \frac{\sum xy - n \bar{x} \bar{y}}{(n-1) s_x^2}$$



Utilité de la droite de régression

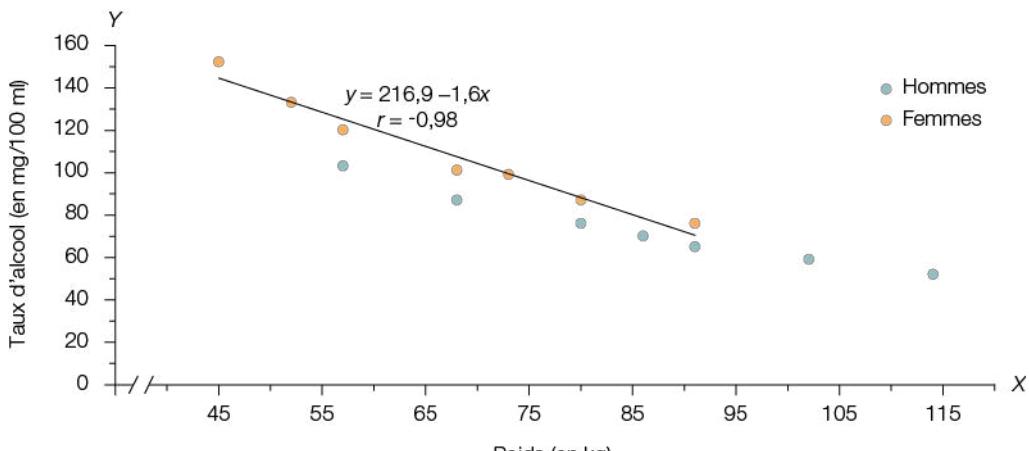
On emploie la droite de régression pour estimer la valeur y associée à une valeur x donnée. Il suffit de remplacer x dans l'équation $y = a + bx$ pour obtenir la valeur y correspondante. On peut aussi faire l'inverse, soit trouver la valeur x correspondant à une valeur y donnée.

MISE EN

SITUATION (suite)

- a) On reprend ici les données de la mise en situation de la section précédente. On a inscrit dans le graphique le coefficient de corrélation et l'équation de la droite de régression du taux d'alcool en fonction du poids chez les femmes. Pour compléter le graphique, trouvons l'équation de la droite de régression pour les hommes, puis traçons cette droite sur le diagramme de dispersion.

Taux d'alcool dans le sang en fonction du poids, par sexe, après la consommation de trois bières



Hommes

Taux d'alcool dans le sang en fonction du poids après la consommation de trois bières

X: Poids	57 kg	68 kg	80 kg	86 kg	91 kg	102 kg	114 kg
Y: Taux d'alcool (en mg/100 ml)	103	87	76	70	65	59	52

On a: $\sum xy = 41748 \quad \bar{x} = 85,43 \quad s_x = 19,42$
 $n = 7 \quad \bar{y} = 73,14 \quad s_y = 17,39$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{(n-1)s_x^2} =$$

$$a = \bar{y} - b\bar{x} =$$

Équation de la droite de régression: $y = a + bx$

Pour tracer la droite, il suffit de placer sur le graphique deux points appartenant à la droite et de les relier. Trouvons deux points de la droite de régression :

- b) On sait qu'il est illégal au Québec de conduire avec un taux d'alcool dans le sang supérieur à 80 mg/100 ml. Selon le modèle mathématique, un homme de 62 kg (137 lb) qui a consommé trois bières peut-il prendre le volant sans enfreindre la loi ?

NOTE

Au Québec, pour un titulaire de permis de conduire âgé de 21 ans ou moins, il est interdit de conduire un véhicule routier s'il y a présence d'alcool dans son organisme.

- c) Considérant votre poids et votre sexe, quel serait théoriquement votre taux d'alcool sanguin après la consommation de trois bières ? (On trouve l'équation de la droite de régression pour les femmes dans le graphique de la page précédente).

(Si vous ignorez votre poids en kilogrammes, utilisez l'équivalence suivante : 1 lb \approx 0,45 kg.)

- d) Selon le modèle mathématique, à combien peut-on estimer le poids d'un homme qui a un taux d'alcool sanguin de 68 mg/100 ml après la consommation de trois bières ?

NOTE

- Comme on peut toujours trouver la moyenne et l'écart type corrigé d'une série de données, le calcul des valeurs de a et de b de la droite de régression sera toujours possible, mais une estimation qui serait faite en utilisant la droite de régression trouvée pour deux variables qui n'ont aucun lien logique entre elles n'aurait pas de sens. La droite de régression traduisant la corrélation entre l'âge mental d'un enfant et la longueur de ses pieds ne serait d'aucune utilité.
- Il est préférable de s'assurer qu'il y a une bonne corrélation linéaire entre les variables avant de faire une estimation avec la droite de régression.
- Il serait risqué de faire des estimations pour des valeurs de la variable X trop éloignées de l'étendue des valeurs observées. (*Pour un exemple, voir le numéro 8 des exercices 7.1.*)
- Une corrélation positive donne une valeur de b positive, et une corrélation négative produit une valeur de b négative.

7.2.2 Le coefficient de détermination

En élevant le coefficient de corrélation au carré, on obtient le coefficient de détermination (r^2). Exprimé en pourcentage, il indique la part de la variation de la variable Y expliquée par la relation entre les variables X et Y . Si tous les points ne sont pas situés sur la droite de régression, c'est que d'autres facteurs influent sur la variation de la variable Y ; on dira que $(1 - r^2)$, exprimé en pourcentage, est la part de la variation de la variable Y qui est attribuable à ces facteurs.

EXEMPLE

Dans la mise en situation portant sur le taux d'alcool dans le sang en fonction du poids après la consommation de trois bières, la valeur du coefficient de corrélation r entre les variables est -0,98, tant chez les hommes que chez les femmes. Calculer et interpréter le coefficient de détermination.

Solution

Coefficient de détermination : $r^2 = (-0,98)^2 = 0,96$

Interprétation

On peut soutenir que le poids d'une personne explique 96 % de la variation du taux d'alcool dans le sang après la consommation de trois bières. Par conséquent, 4 % de la variation du taux d'alcool est attribuable à d'autres facteurs, par exemple la vitesse de consommation, la fatigue, le stress ou l'état de santé du consommateur.

EXERCICES DE COMPRÉHENSION | 7.1

1. Afin d'estimer les coûts de chauffage d'un immeuble, on note quotidiennement le nombre de litres de mazout consommé et la température extérieure moyenne.

- a) Désigner la variable indépendante par X et la variable dépendante par Y .

_____ : Consommation de mazout (en litres)

_____ : Température extérieure moyenne (en Celsius)

La corrélation entre X et Y est-elle positive ou négative ? _____

- b) Compléter l'interprétation du coefficient de corrélation si sa valeur est -0,95.

Interprétation

Il y a une [indiquer la force] _____ corrélation linéaire entre la température extérieure et la quantité de mazout consommé. Le fait que la corrélation est négative indique la tendance suivante : _____ la température extérieure est élevée, _____ on consomme de mazout.

- c) Compléter l'interprétation du coefficient de détermination.

Interprétation

Le coefficient de détermination est _____. On peut estimer que la variation de la température extérieure explique _____ % de la variation quotidienne de mazout consommé. Par conséquent, _____ % de cette variation est attribuable à d'autres facteurs (force des vents, ensoleillement, etc.).

2. Quel montant le gouvernement doit-il débourser par étudiant pour une année de cégep ? Le tableau ci-dessous donne l'évolution de la dépense de fonctionnement des cégeps par étudiant dans le réseau collégial public de 2004 à 2010. On donne le nom de **série chronologique** à ce type de série.

Évolution de la dépense de fonctionnement des cégeps par étudiant, Québec, 2004-2010

X: Année	2004-2005	2005-2006	2006-2007	2007-2008	2008-2009	2009-2010
Y: Dépense par étudiant (en dollars courants)	8 832 \$	9 085 \$	9 453 \$	9 413 \$	9 772 \$	9 877 \$

Source: Ministère de l'Éducation, du Loisir et du Sport, et Ministère de l'Enseignement supérieur. *Indicateurs de l'éducation – Édition 2012, 2013.*

- a) En assignant $x = 1$ à l'année 2004-2005, $x = 2$ à l'année 2005-2006 et ainsi de suite, trouver l'équation de la droite de régression de cette série statistique ($r = 0,97$). (Dans le cas des séries chronologiques, la droite de régression est souvent nommée **droite de tendance**.)

- b) À l'aide de la droite de régression, estimer la dépense gouvernementale moyenne par étudiant pour l'année 2013-2014 si la tendance s'est maintenue.

7.2.3 L'utilisation du mode statistique de la calculatrice (deux variables)

Le mode statistique de la calculatrice permet d'obtenir rapidement le coefficient de corrélation (r) ainsi que l'ordonnée à l'origine (a) et la pente (b) de la droite de régression. Pour savoir comment utiliser le mode statistique de la calculatrice pour traiter simultanément les données de deux variables, consultez le guide d'utilisation qui s'applique au modèle de votre calculatrice.

Guide d'utilisation de la calculatrice scientifique de base (deux variables)

Modèle Sharp EL-531W ou équivalent

CHOISIR LE MODE STATISTIQUE À DEUX VARIABLES

Appuyer sur **MODE**, sélectionner l'option **STAT** en appuyant sur **1**, puis l'option **LINE** en appuyant de nouveau sur **1**.

(Pour certains modèles, on sélectionne le mode **STAT**, puis on choisit l'option **1: $a + bx$** ou **REG**.)

ENTRER LES DONNÉES DES VARIABLES X ET Y

Entrer les coordonnées du premier couple de points (x, y) du tableau ainsi :

- saisir la valeur x , puis appuyer sur **STO** (**x, y**);
- saisir la valeur y , puis appuyer sur **M+** (**data**).

Procéder de la même façon pour entrer les coordonnées des autres couples du tableau.

AFFICHER LES VALEURS CHERCHÉES

- Pour le coefficient de corrélation, appuyer sur **RCL**, puis sur **÷** (**r**).
- Pour la valeur a de la droite de régression, appuyer sur **RCL**, puis sur **| (a)**.
- Pour la valeur b de la droite de régression, appuyer sur **RCL**, puis sur **| (b)**.

On peut aussi obtenir $\sum xy$, \bar{x} , \bar{y} , σ_x , σ_y , s_x ou s_y en appuyant sur **RCL**, puis sur le bouton associé à la mesure désirée.

Attention ! Certaines calculatrices utilisent la forme $y = ax + b$ au lieu de $y = a + bx$ pour l'équation de la droite de régression. Dans ce cas, a est la pente de la droite et b , son ordonnée à l'origine.

Guide d'utilisation de la calculatrice graphique (deux variables)

Modèle TI-84 Plus ou équivalent

CHOISIR LE MODE STATISTIQUE

- Appuyer sur **STAT**.
- Placer le curseur sur le menu **EDIT**, puis sur **1: EDIT** et appuyer sur **ENTER**.

ENTRER LES DONNÉES DES VARIABLES X ET Y

- Entrer la première valeur de la variable *X* dans la colonne **L1**, puis appuyer sur **ENTER**.
Faire de même pour les autres valeurs de *X*.
- Entrer la première valeur de la variable *Y* dans la colonne **L2**, puis appuyer sur **ENTER**.
Faire de même pour les autres valeurs de *Y*.

AFFICHER LES VALEURS CHERCHÉES

- Appuyer sur **STAT**.
- Placer le curseur sur le menu **CALC**, puis sur **4: LIN REG (ax + b)** et appuyer sur **ENTER**.
- Pour afficher les valeurs cherchées, appuyer sur les boutons **2nd ; 1 (L1) ; , ; 2nd ; 2 (L2)**, puis sur **ENTER**.
On peut aussi obtenir $\sum xy$, \bar{x} , \bar{y} , σ_x , σ_y , s_x ou s_y ainsi : appuyer sur **STAT**, placer le curseur sur le menu **CALC**, puis sur **2: VAR STATS** et appuyer sur **ENTER**.
(Si les valeurs de *a*, de *b* et de *r* ne s'affichent pas, activer **DIAGNOSTIC ON** dans **CATALOG** avant d'effectuer les étapes précédentes.)

EXERCICES 7.1

- La corrélation entre les variables suivantes est-elle positive, négative ou nulle ?
 - Revenu personnel et total des impôts personnels à payer.
 - Poids de l'homme et poids de la femme dans un couple.
 - Âge de l'homme et âge de la femme dans un couple.
- a) Quelle serait la valeur du coefficient de corrélation *r*, calculé à partir d'un nuage de points ne contenant que deux points ?
b) Indiquer la variable indépendante *X* et la variable dépendante *Y*.
____ : Épaisseur de la glace couvrant un lac
____ : Température extérieure
- Des chercheurs de l'Université du Michigan ont voulu savoir si le risque de mourir d'un cancer du poumon était directement lié à l'âge auquel une personne cesse de fumer. Ils ont mené une étude auprès d'un échantillon de 900 000 personnes dont

50 % étaient d'anciens fumeurs des deux sexes. L'étude a permis d'établir, pour chaque sexe, le taux de mortalité due à un cancer du poumon par 100 000 ex-fumeurs selon l'âge auquel ils ont cessé de fumer.

- a) Le tableau suivant donne les taux obtenus pour les hommes de l'échantillon. Les chercheurs peuvent-ils affirmer que le taux de mortalité chez les ex-fumeurs dépend de l'âge auquel ils ont cessé de fumer ? Justifier la réponse en fonction de la valeur du coefficient de corrélation.

Taux de mortalité due à un cancer du poumon par 100 000 ex-fumeurs masculins en fonction de l'âge auquel ils ont cessé de fumer

Âge (en ans)	35	45	55	60	65
Taux (en ‰)	90	150	240	340	500

Source: *Le Soleil*, 1993.

- b) Calculer et interpréter le coefficient de détermination.

4. Les statistiques du tableau suivant témoignent de la progression du nombre d'emplois dans l'industrie des jeux vidéo au Québec.

Évolution du nombre d'emplois dans l'industrie des jeux vidéo, Québec, 2008-2012

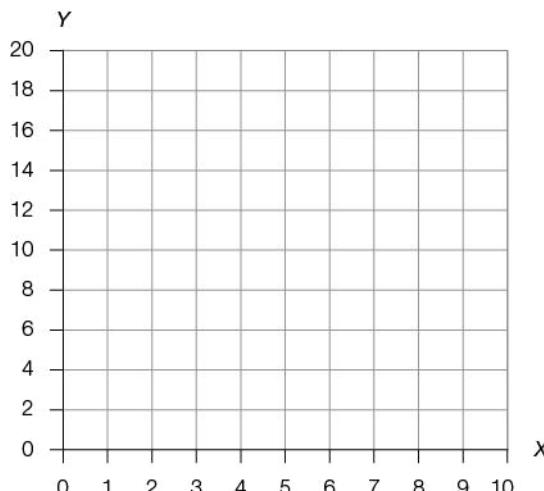
Année	2008	2009	2010	2011	2012
Nombre d'emplois	5 466	6 593	6 902	8 266	8 980

Source: TECHNOCompétences. *L'emploi dans l'industrie du jeu électronique au Québec en 2012*, avril 2013.

- a) En assignant $x = 0$ à l'année 2008, $x = 1$ à l'année 2009 et ainsi de suite, trouver l'équation de la droite de régression ($r = 0,99$).
- b) Si la tendance se maintient, combien d'emplois y aura-t-il dans l'industrie des jeux vidéo en 2015 ?
- c) Selon le modèle mathématique, combien d'emplois y avait-il dans cette industrie en 2006 ?
- d) Établir la signification des valeurs de a et de b de l'équation de la droite de régression dans le contexte de ce problème.
5. a) Calculer le coefficient de corrélation linéaire pour les données suivantes :

X	1	2	3	4	5	6	7	8	9
Y	18	11	6	3	2	3	6	11	18

- b) Les variables sont-elles indépendantes ?
- c) Tracer le nuage de points sur le système d'axes suivant. Que faut-il en conclure ?



6. Lors des parties de hockey junior, on vend des objets promotionnels portant le logo de l'équipe locale. Au cours des 5 derniers mois, on a augmenté sensiblement le prix de vente des casquettes. Le tableau

suivant présente l'effet de ces augmentations de prix sur le volume des ventes.

Volume des ventes en fonction du prix

Mois	Prix	Volume des ventes
Novembre	28 \$	450
Décembre	32 \$	380
Janvier	35 \$	250
Février	38 \$	220
Mars	40 \$	180

- a) Calculer et interpréter le coefficient de corrélation.
- b) Calculer et interpréter le coefficient de détermination.
- c) Donner l'équation de la droite de régression.
- d) Estimer le volume mensuel des ventes si le prix est de 42 \$.
- e) Au mois d'avril, on veut écouter les 310 casquettes qu'il reste en stock. Quel prix de vente devrait-on retenir pour atteindre cet objectif ?
7. Cinq clients d'Hydro-Québec ont payé les montants suivants, au dollar près, pour le nombre de kilowattheures (kWh) consommés.

Montant de la facture d'électricité en fonction du nombre de kilowattheures consommés en 60 jours

Consommation (en kWh)	1 600	800	1 250	1 700	940
Montant à payer (en \$)	114	69	94	120	77

Source: Hydro-Québec. Données basées sur la tarification de 2014.

- a) Calculer et interpréter le coefficient de corrélation.
- b) À quoi ressemblerait le diagramme de dispersion ?
- c) Calculer et interpréter le coefficient de détermination.
- d) Trouver l'équation de la droite de régression.
- e) Estimer le montant qu'il faudrait payer pour une consommation de 1 400 kilowattheures.
- f) Combien de kilowattheures a consommé un client dont la facture s'élève à 88 \$?
- g) Établir la signification des valeurs de a et de b de l'équation de la droite de régression dans le contexte de ce problème.
8. Dans le bilan 2010-2011 de Recyc-Québec, on indique l'évolution du taux de récupération des matières recyclables à la suite de la collecte sélective des déchets effectuée par les municipalités. Voici quelques statistiques.

Évolution du taux de récupération des matières recyclables par la collecte sélective des déchets résidentiels, Québec, 2002-2010

Année	2002	2004	2006	2008	2010
Taux de récupération	20 %	23 %	32 %	56 %	59 %

Source: Recyc-Québec. *Bilan 2010-2011 de la gestion des matières résiduelles au Québec*, 2012.

- a) Poser X : (Année – 2000) et Y : Taux de récupération (en %), puis déterminer l'équation de la droite de régression. (On a $r = 0,96$.)
 - b) Utiliser le modèle mathématique construit pour estimer le taux de récupération en 2012.
 - c) Peut-on utiliser la droite de régression pour estimer le taux de récupération en 2018 ? Justifier la réponse.
-

Corrélation linéaire

On mesure la force de la dépendance linéaire entre deux variables quantitatives au moyen du coefficient de corrélation linéaire (r). Plus la valeur de celui-ci est proche de -1 ou de +1, plus la dépendance linéaire est forte.

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

Le **coefficient de détermination** est le carré du coefficient de corrélation, soit r^2 . Exprimé en pourcentage, il indique la part de la variation de la variable dépendante Y qui s'explique par la variation de la variable indépendante X .

Régression linéaire

S'il y a un lien logique entre deux variables et une bonne corrélation linéaire, la droite de régression peut servir de modèle mathématique de la relation entre les variables. Cette droite permet d'estimer la valeur d'une variable correspondant à une valeur donnée de l'autre variable.

Équation de la droite de régression : $y = a + bx$

On détermine les valeurs de a et de b ainsi : $a = \bar{y} - b\bar{x}$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{(n-1)s_x^2}$$

EXERCICE RÉCAPITULATIF

Un chauffeur de taxi a noté la distance parcourue et le coût de ses sept dernières courses. Voici les résultats.

Coût d'une course en taxi en fonction de la distance parcourue

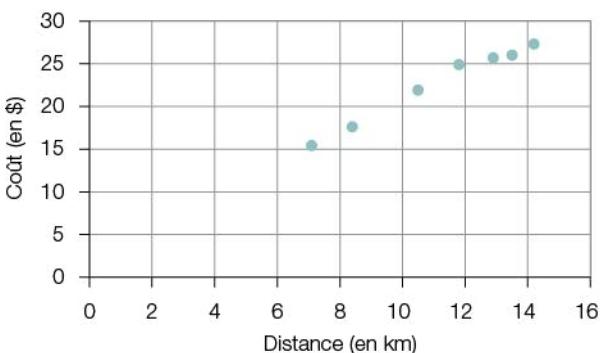
Distance (en km)	14,2	12,9	8,4	10,5	11,8	7,1	13,5
Coût (en \$)	27,3	25,7	17,6	21,9	24,9	15,4	26,0

Source: Commission des transports du Québec. Données basées sur la tarification de 2014.

- Désigner la variable indépendante par X et la variable dépendante par Y .
- Calculer et interpréter le coefficient de corrélation.
- Calculer et interpréter le coefficient de détermination.
- Donner l'équation de la droite de régression.
- Dans le contexte de ce problème, donner une signification aux valeurs a et b de la droite de régression.

- Estimer le coût d'une course de 16 km.
- Estimer la distance parcourue si le coût d'une course est de 30 \$.
- Tracer la droite de régression sur le diagramme de dispersion suivant.

Coût d'une course en taxi en fonction de la distance parcourue



PRÉPARATION À L'EXAMEN

Pour préparer votre examen, assurez-vous d'avoir les compétences suivantes :

	Si vous avez la compétence, cochez.
Corrélation linéaire	
• Tracer un diagramme de dispersion.	<input type="radio"/>
• Connaître les caractéristiques d'une corrélation : forme, sens et intensité.	<input type="radio"/>
• La formule étant donnée, calculer et interpréter le coefficient de corrélation r .	<input type="radio"/>
• Connaître les propriétés du coefficient de corrélation linéaire.	<input type="radio"/>
• Calculer et interpréter le coefficient de détermination.	<input type="radio"/>
Régression linéaire	
• Différencier les variables indépendante X et dépendante Y d'un problème.	<input type="radio"/>
• Les formules étant données, trouver l'équation de la droite de régression.	<input type="radio"/>
• Estimer une valeur de la variable dépendante Y ou indépendante X à l'aide de la droite de régression.	<input type="radio"/>

Chapitre 8

Les séries chronologiques



OBJECTIF DU CHAPITRE

Analyser l'évolution d'une variable dans le temps à l'aide d'une série chronologique.



Les séries chronologiques permettent de suivre l'évolution d'une variable dans le temps. Elles révèlent souvent l'ampleur des changements sociaux ou économiques qui se produisent au fil des ans. Par exemple, l'étude du nombre de naissances ou de mariages au Québec durant une période de 30 ans montre le changement de perception qui s'est opéré chez les Québécois par rapport à la famille et au mariage. De même, l'étude du taux de chômage durant quelques années fournit un aperçu de l'évolution du marché du travail dans une région donnée.

8.1 La définition et la représentation d'une série chronologique

Définition

On donne le nom de **série chronologique** à la succession de valeurs que prend une variable dans le temps (au fil des ans, des mois, des jours, des heures, etc.).

Représentation graphique

On représente graphiquement une série chronologique de l'une ou l'autre des façons suivantes :

- au moyen d'une série de points reliés par des segments de droite dans le plan cartésien. L'abscisse de chacun de ces points représente le temps et l'ordonnée, la valeur prise par la variable au temps considéré. On donne le nom de **chronogramme** à ce graphique (*voir l'exemple 1 ci-dessous*);
- au moyen d'un diagramme à rectangles. Pour utiliser ce type de graphique, il faut toutefois que le nombre de périodes de temps considérées ne soit pas trop élevé (*voir l'exemple 2 de la page suivante*).

EXEMPLE 1

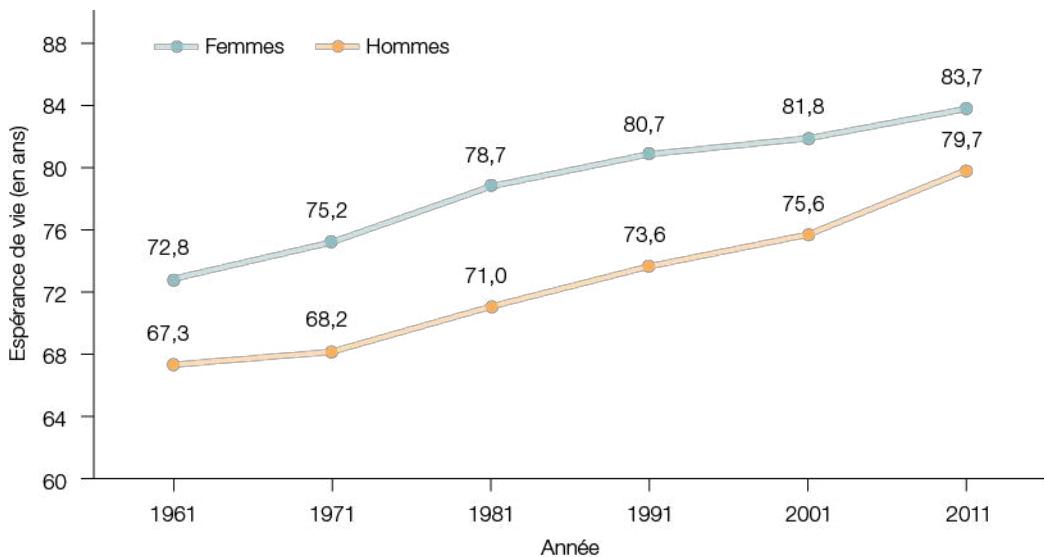
Les données de la série chronologique donnant l'espérance de vie des hommes et des femmes de 1961 à 2011 sont présentées dans le tableau et le chronogramme suivants. Analyser ces données.

Évolution de l'espérance de vie à la naissance selon le sexe, Québec, 1961-2011

Sexe	Année					
	1961	1971	1981	1991	2001	2011
Femmes	72,8 ans	75,2 ans	78,7 ans	80,7 ans	81,8 ans	83,7 ans
Hommes	67,3 ans	68,2 ans	71,0 ans	73,6 ans	75,6 ans	79,7 ans

Source: Institut de la statistique du Québec, 2013.

Évolution de l'espérance de vie à la naissance, Québec, 1961-2011



Source: Institut de la statistique du Québec, 2013.

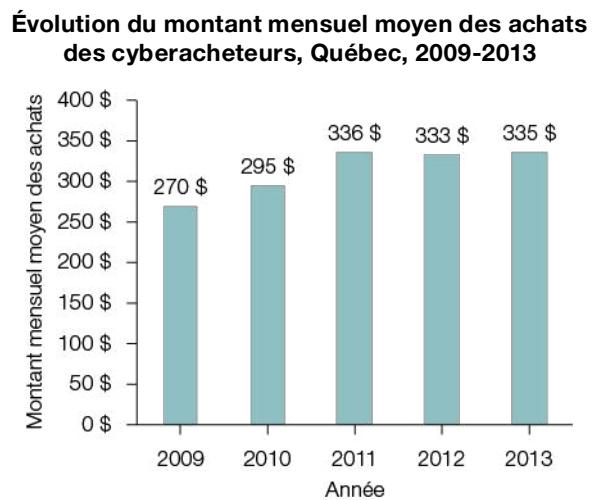
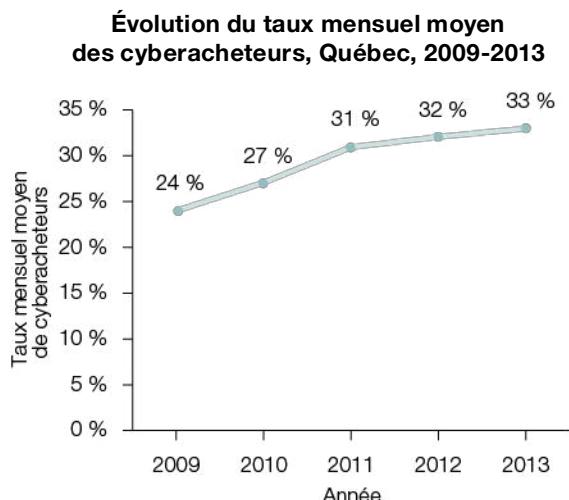
Analyse des données

L'espérance de vie à la naissance a connu une hausse remarquable en 50 ans. Chez les femmes, elle est passée de 72,8 ans en 1961 à 83,7 ans en 2011, soit une augmentation de 10,9 ans ; chez les hommes, elle est passée de 67,3 ans en 1961 à 79,7 ans en 2011, soit une augmentation de 12,4 ans.

De 1961 à 2011, l'espérance de vie des femmes a toujours été plus élevée que celle des hommes. Si l'écart entre les sexes a atteint sa plus grande valeur en 1981, soit 7,7 ans, il n'est plus que de 4 ans en 2011.

EXEMPLE 2

Le chronogramme et le diagramme à rectangles suivants donnent respectivement le taux mensuel moyen des internautes qui effectuent des achats en ligne et la moyenne mensuelle du montant de leurs achats de 2009 à 2013. Analyser ces données.



Source: CEFRIQO. NETendances 2013: Le commerce électronique en pleine croissance au Québec, vol. 4, n° 10.

Analyse des données

De 2009 à 2011, le taux de cyberacheteurs passe de 24 % à 31 %, une augmentation de 7 points de pourcentage. Pendant la même période, la moyenne mensuelle de leurs achats passe de 270 \$ à 336 \$, une augmentation de 66 \$. Par la suite, de 2012 à 2013, le taux mensuel de cyberacheteurs augmente d'à peine 1 point de pourcentage par année et leurs achats se situent légèrement sous la moyenne mensuelle de 336 \$ observée en 2011.

8.2 Les composantes d'une série chronologique

Cette section présente les caractéristiques, ou composantes, propres à certaines séries chronologiques : la tendance à long terme, les cycles, la saisonnalité ainsi que les variations uniques et aléatoires.

Tendance à long terme

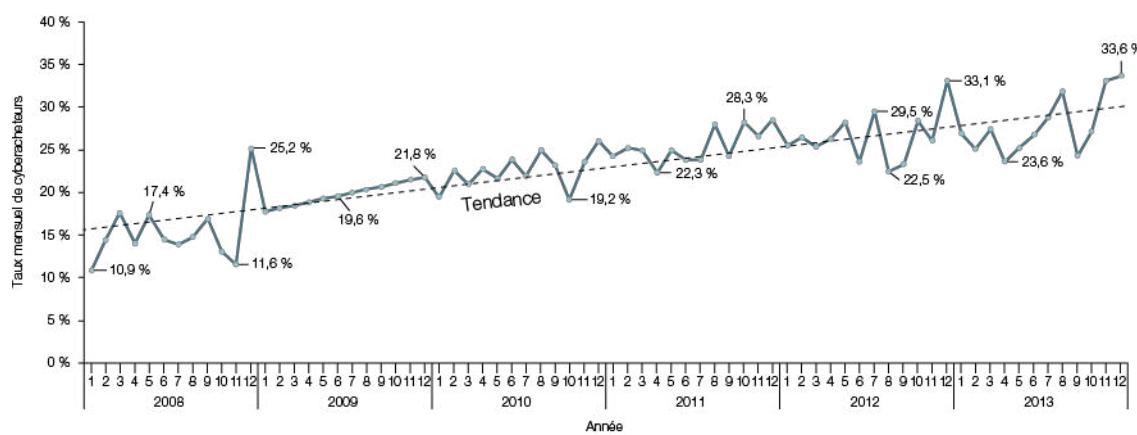
La tendance représente l'allure générale de la série chronologique ; elle traduit le comportement général de la variable à long terme. Une tendance peut être **croissante**, **décroissante** ou **stable**. De plus, si les points de la série décrivent *grossost modo* une droite, on dit que la série a une tendance **linéaire** ; sinon, on dit que la tendance est **non linéaire** ou **curviligne**.

NOTE

Si l'on veut faire une analyse valable de la tendance à long terme d'une série, il est préférable d'obtenir des données couvrant un grand nombre de périodes ; autrement, on risque de confondre un cycle (*voir la page suivante*) avec la tendance de la série.

EXEMPLE 1

Évolution du taux mensuel de cyberacheteurs, Québec, 2008-2013

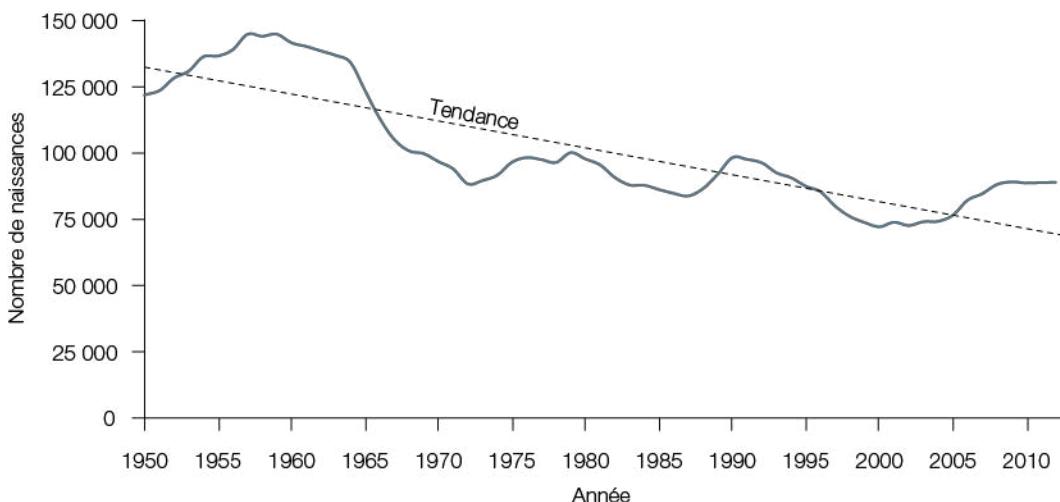


Source: CEFARIO. *Indice du commerce électronique au Québec (ICEQ)*, 2014.

La série chronologique illustrée ci-dessus présente une tendance linéaire croissante : *grossost modo*, le pourcentage mensuel de cyberacheteurs chez les internautes tend à augmenter de façon linéaire avec le temps.

EXEMPLE 2

Évolution du nombre de naissances par année, Québec, 1950-2012



Source: Institut de la statistique du Québec.

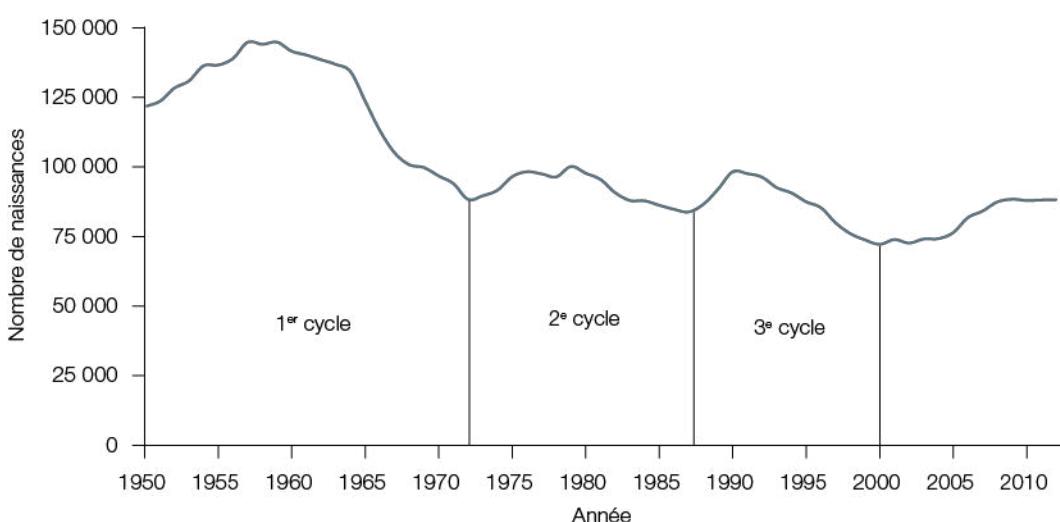
La série chronologique illustrée ci-dessus présente une tendance curviligne décroissante : globalement, de 1950 à 2012, le nombre de naissances a diminué au Québec.

Cycles

On dit qu'une série chronologique est **cyclique** si l'on observe, sur une période de quelques années, une suite de hausses et de baisses de durée variable. La durée d'un cycle est généralement imprévisible. En ce sens, il n'est pas possible de dire où l'on se trouve actuellement (au jour le jour) dans un cycle. Cela prend du recul pour analyser les données et le déterminer. On le voit bien dans le graphique suivant : la dernière section n'est pas identifiée comme un cycle.

EXEMPLE

Évolution du nombre de naissances par année, Québec, 1950-2012



Source: Institut de la statistique du Québec.

La série chronologique illustrée à la page précédente comporte trois cycles, soit trois intervalles de temps où une hausse des naissances est suivie par une baisse de celles-ci :

- 1^{er} cycle : le nombre de naissances augmente de 1950 à 1959, puis il diminue jusqu'en 1972 ;
- 2^e cycle : le nombre de naissances augmente de 1972 à 1979, puis il diminue jusqu'en 1987 ;
- 3^e cycle : le nombre de naissances augmente de 1987 à 1990, puis il diminue jusqu'en 2000.

Quelle est la cause de ces cycles ? Voilà un beau sujet d'étude pour un chercheur !

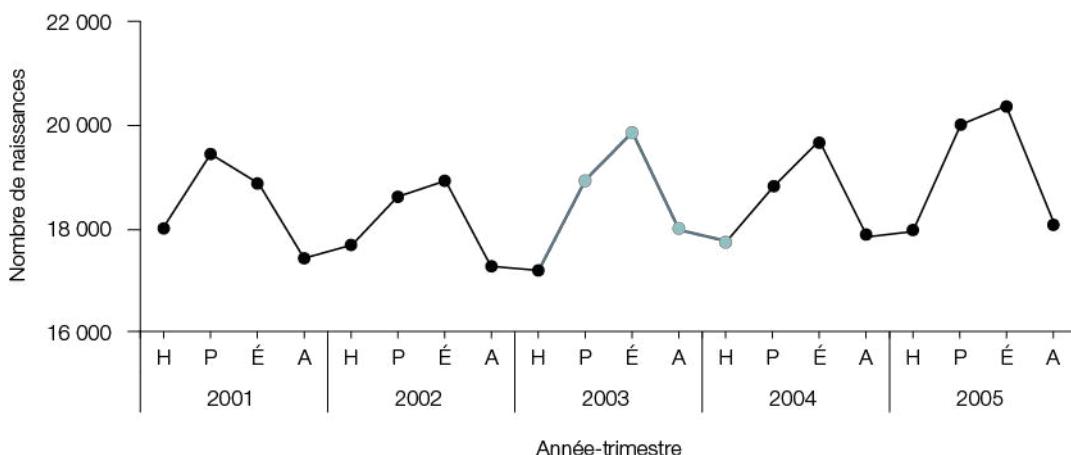
Saisonnalité

On dit qu'une série montre de la **saisonnalité** lorsqu'on observe un changement périodique de la variable, l'intervalle de variation étant une journée, une semaine, une année, etc. De telles variations sont souvent attribuables aux conditions climatiques (le taux de chômage est toujours plus élevé l'hiver que l'été ; il fait toujours plus froid la nuit que le jour), aux coutumes propres à une population et aux fêtes religieuses (le volume des ventes est toujours plus grand les jours précédent Noël, Pâques, la rentrée scolaire, etc.). La représentation graphique d'une série chronologique permet de déceler assez facilement l'existence d'un phénomène saisonnier, comme l'illustre l'exemple suivant.

EXEMPLE

Nous avons étudié dans l'exemple précédent la tendance du nombre de naissances par année au Québec de 1950 à 2012. Demandons-nous maintenant si cette variable est soumise à un phénomène de saisonnalité. Autrement dit, y a-t-il un lien entre le nombre de naissances et les saisons ? Pour répondre à cette question, il faut compter et porter sur un graphique le nombre de naissances par trimestre (hiver, printemps, été, automne). Voici la représentation du nombre de naissances par trimestre de 2001 à 2005.

Évolution du nombre de naissances par trimestre, Québec, 2001-2005



Source: Institut de la statistique du Québec.

La section en bleu met en évidence la forme de la saisonnalité : il s'agit d'une saisonnalité annuelle sur un cycle de quatre périodes. Le nombre de naissances est toujours plus élevé durant les trimestres d'été et du printemps, et plus faible durant ceux d'automne et d'hiver.

Variations uniques et aléatoires

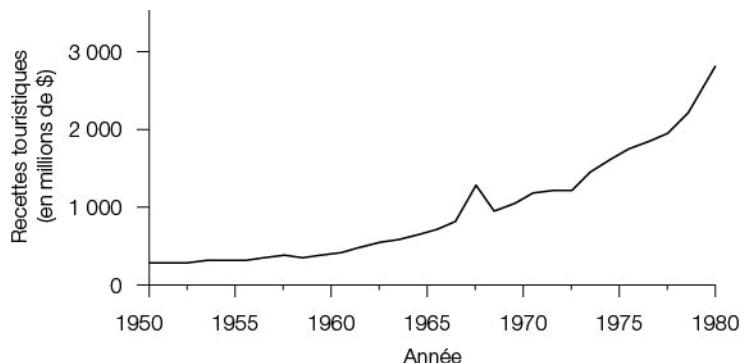
Les variations uniques sont des phénomènes qui «brisent» parfois la régularité de la tendance, des cycles et des changements saisonniers. Elles résultent d'événements importants limités dans le temps, mais qui ont des répercussions majeures, tels une grève, des Jeux olympiques, un tremblement de terre, etc.

Une série chronologique peut aussi présenter des variations non expliquées, irrégulières et imprévisibles sur de très courtes périodes, mais qui ont des répercussions limitées. De telles variations sont dites aléatoires. En pratique, on considère comme aléatoires toutes les variations d'une série chronologique qui ne peuvent être attribuées à un phénomène cyclique, saisonnier, de tendance ou unique.

EXEMPLE

La série chronologique des recettes touristiques au Canada offre un exemple de variation unique. On observe une augmentation fulgurante (56 %) des recettes touristiques en 1967, à cause de l'Exposition universelle de Montréal (Expo 67), qui a attiré un nombre exceptionnel de visiteurs.

Évolution des recettes touristiques, Canada, 1950-1980



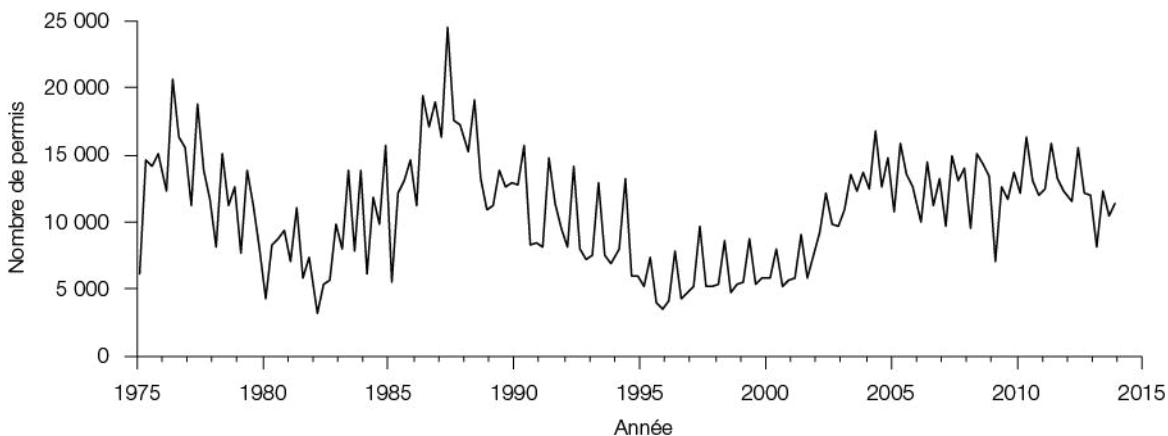
Source: Statistique Canada.

8.3 Le lissage d'une série chronologique

MISE EN SITUATION

Certaines séries chronologiques possèdent simultanément plusieurs caractéristiques. Bien que ce ne soit pas évident au premier coup d'œil, c'est le cas de la série chronologique suivante, qui présente une tendance curviligne, des cycles et une saisonnalité annuelle.

Évolution du nombre de permis de bâtir par trimestre, Québec, 1975-2013

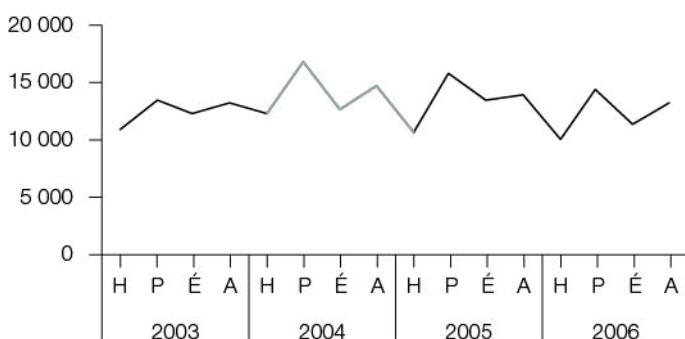


Source: Statistique Canada. Tableau 026-0001, CANSIM.

Étude de la saisonnalité

On met en évidence la saisonnalité du nombre de permis de bâtir par trimestre en effectuant un gros plan sur une partie du graphique de la série. En retenant uniquement les années 2003-2006, on obtient le graphique ci-contre. On y observe en bleu une saisonnalité annuelle sur un cycle de quatre périodes: chaque année, c'est au printemps que le nombre de permis de bâtir délivrés est le plus élevé, et à l'hiver qu'il est le plus faible.

Nombre de permis de bâtir par trimestre, Québec, 2003-2006



Étude de la tendance à long terme et des cycles

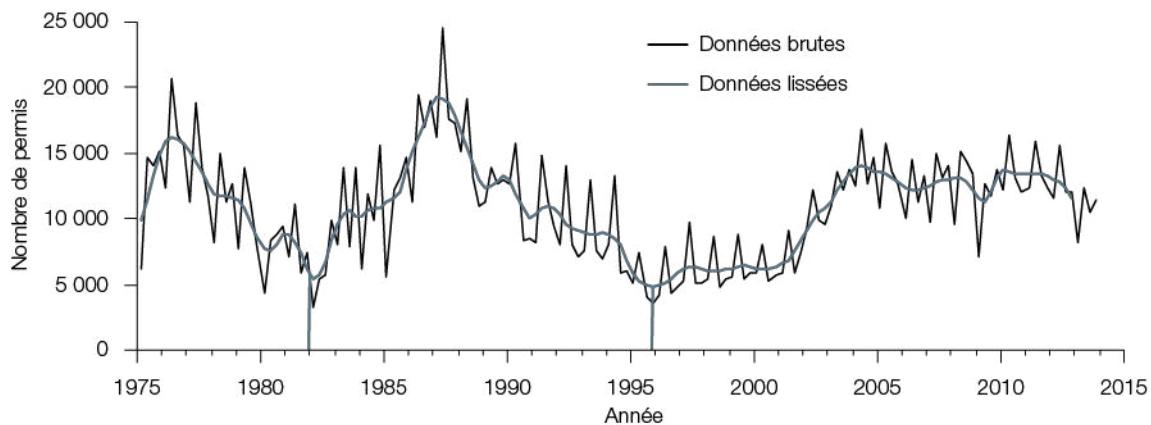
Pour mieux faire ressortir la tendance à long terme et les cycles du nombre de permis de bâtir par trimestre, nous allons effectuer un lissage de la série.

Méthodes de lissage d'une série chronologique

Le lissage d'une série chronologique a pour objectif d'éliminer du graphique les variations en dents de scie attribuables aux changements saisonniers et aléatoires de la variable. La courbe lisse ainsi obtenue facilite l'analyse des variations de la variable. Il existe plusieurs techniques de lissage, dont le **lissage par moyenne mobile** (pondérée ou non), surtout utilisé pour des séries chronologiques qui ont une saisonnalité, et le **lissage exponentiel**.

Avant d'explorer les techniques de lissage d'une série chronologique, observons le résultat du lissage par moyenne mobile pondérée de la série chronologique donnant le nombre de permis de bâtir.

Évolution du nombre de permis de bâtir par trimestre, Québec, 1975-2013



La courbe lissée met en évidence les caractéristiques suivantes :

- la série a une tendance curviligne stable ;
- la série comporte deux cycles : 1^{er} cycle, de 1975 à 1982 ;
2^e cycle, de 1982 à 1996.

8.3.1 Le lissage par moyenne mobile

On obtient le lissage d'une série chronologique par moyenne mobile en remplaçant chaque donnée de la série brute (série 1) par la moyenne, pondérée ou non, de données appartenant à un voisinage.

Nous allons effectuer le lissage de la série chronologique du nombre de permis de bâtir par trimestre de deux façons: avec une moyenne mobile de rayon 1 (série 2) et avec une moyenne mobile pondérée de rayon 2 (série 3). Pour expliquer la procédure, nous nous limiterons aux trimestres de 2002 à 2006.

Moyenne mobile de rayon 1

Effectuons un premier lissage en remplaçant chaque donnée x_i de la série par la moyenne de trois données consécutives, soit x_{i-1} , x_i et x_{i+1} . On dit que l'on retient une fenêtre de rayon $r = 1$ pour calculer la moyenne mobile, le centre de cette fenêtre étant x_i .

Par exemple, on remplace le nombre de permis de bâtir délivrés durant le trimestre d'été 2003 par la moyenne du nombre de permis délivrés au cours du printemps, de l'été et de l'automne 2003 (*voir le tableau présenté ci-contre*).

Été 2003

$$\frac{13\,596 + 12\,266 + 13\,734}{3} = 13\,198,7$$

 Calculer la moyenne mobile pour le trimestre d'hiver 2005 et la reporter dans le tableau 1.

Hiver 2005

Tableau 1

Moyenne mobile de rayon 1

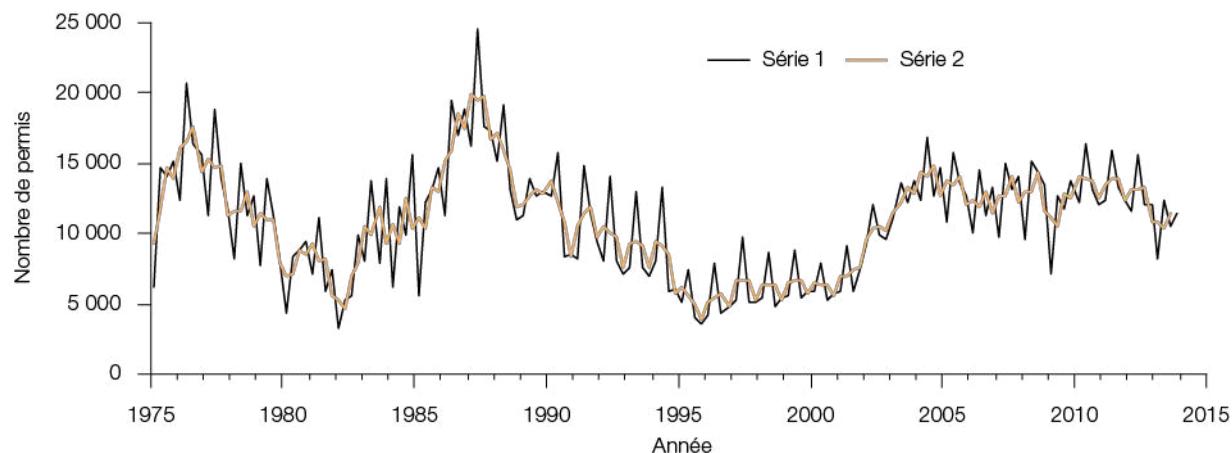
Année	Trimestre	Nombre de permis	Moyenne $r = 1$
		Série 1	Série 2
2002	Hiver	9 194	—
	Printemps	12 115	10 395,7
	Été	9 878	10 542,3
	Automne	9 634	10 153,3
	Hiver	10 948	11 392,7
	Printemps	13 596	12 270,0
	Été	12 266	13 198,7
	Automne	13 734	12 808,0
	Hiver	12 424	14 329,3
2003	Printemps	16 830	13 975,0
	Été	12 671	14 743,7
	Automne	14 730	12 725,0
	Hiver	10 774	—
	Printemps	15 797	13 386,3
2004	Été	13 588	14 023,3
	Automne	12 685	12 110,3
	Hiver	10 058	12 398,7
	Printemps	14 453	11 941,0
	Été	11 312	13 017,0
2005	Automne	13 286	—
	Hiver	10 058	12 398,7
	Printemps	14 453	11 941,0
	Été	11 312	13 017,0
2006	Automne	13 286	—
	Hiver	10 058	12 398,7
	Printemps	14 453	11 941,0
	Été	11 312	13 017,0

NOTE

On ne remplace pas les première et dernière données de la série 1, car une fenêtre centrée sur ces données ne peut pas contenir trois données de la série.

Le graphique qui suit montre que la courbe de la série chronologique construite (série 2) est plus lisse que celle de la série originale (série 1).

Évolution du nombre de permis de bâtir par trimestre, Québec, 1975-2013



Moyenne mobile pondérée de rayon 2

En augmentant le rayon de la fenêtre, on améliore le lissage de la série, mais il ne faut pas exagérer, car on risque de faire disparaître les variations significatives de la tendance de la série. Il est donc préférable de choisir un rayon qui délimite une fenêtre couvrant ou dépassant légèrement la saisonnalité dans le cas d'une série chronologique qui présente une saisonnalité.

Comme la série 1 a une saisonnalité annuelle de quatre périodes, un rayon égal à 2 donne dans le cas présent une fenêtre couvrant la saisonnalité.

En incluant la donnée centrale de la fenêtre, on prend la moyenne mobile de cinq données consécutives de la série, alors que le cycle saisonnier en contient quatre ; pour tenir compte de cette différence, on assigne un poids différent aux données de la fenêtre. Bien qu'il y ait plusieurs façons d'effectuer cette pondération, elles reposent toutes sur le même principe : les données au centre de la fenêtre doivent avoir plus de poids que celles qui sont situées aux extrémités.

Nous utiliserons la pondération suivante : les deux données situées aux extrémités de la fenêtre (x_{i-2} et x_{i+2}) auront un poids de 0,5, et les autres (x_{i-1} , x_i et x_{i+1}) un poids de 1.

La série 3 du tableau 2 donne les moyennes pondérées obtenues de cette façon. Voici quelques exemples des calculs qui ont mené aux données de cette série.

Tableau 2

Moyenne mobile pondérée de rayon 2

Année	Trimestre	Nombre de permis	Moyenne $r = 2$
		Série 1	Série 3
2002	Hiver	9 194	—
	Printemps	12 115	—
	Été	9 878	10 424,5
	Automne	9 634	10 828,9
2003	Hiver	10 948	11 312,5
	Printemps	13 596	12 123,5
	Été	12 266	12 820,5
	Automne	13 734	13 409,3
2004	Hiver	12 424	13 864,1
	Printemps	16 830	14 039,3
	Été	12 671	13 957,5
	Automne	14 730	13 622,1
2005	Hiver	10 774	—
	Printemps	15 797	13 466,6
	Été	13 588	13 121,5
	Automne	12 685	12 864,0
2006	Hiver	10 058	12 411,5
	Printemps	14 453	12 202,1
	Été	11 312	—
	Automne	13 286	—

Été 2003

$$\frac{10\,948 \times 0,5 + 13\,596 + 12\,266 + 13\,734 + 12\,424 \times 0,5}{4} = 12\,820,5$$

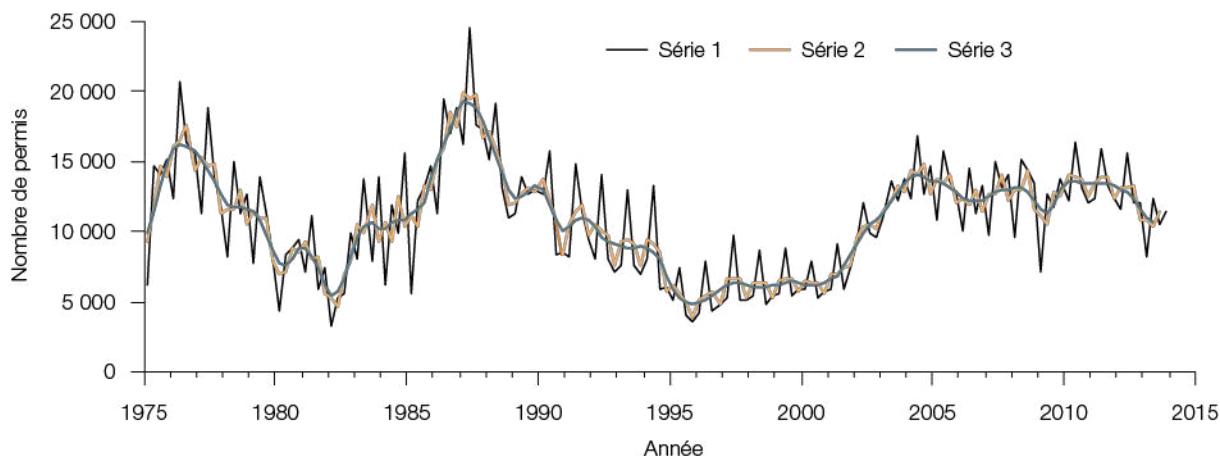
(Le nombre 4 au dénominateur correspond à la pondération totale.)

- ?
- Calculer la moyenne mobile pondérée pour le trimestre d'hiver 2005 et la reporter dans le tableau 2.

Hiver 2005

Une superposition des courbes des séries 1, 2 et 3 montre bien que la courbe de la série 3 est plus lisse que celle de la série 2.

Évolution du nombre de permis de bâtir par trimestre, Québec, 1975-2013



8.3.2 Le lissage exponentiel

Contrairement au lissage par moyenne mobile, qui tient compte des données des périodes adjacentes à la période lissée, le lissage exponentiel prend uniquement en compte les périodes précédentes. Ce type de lissage consiste à attribuer un poids α à la donnée de la période lissée et un poids $(1 - \alpha)$ à la donnée lissée de la période précédente. Plus la valeur de α est grande, plus on accorde un poids important à la période à lisser par rapport aux périodes précédentes.

Voici comment on effectue un lissage exponentiel de la série 1 du nombre de permis de bâtir en utilisant $\alpha = 0,40^1$. Pour faciliter le lissage, on numérote les périodes de 1 à n , la valeur n étant le nombre total de périodes étudiées.

1. Pour déterminer la valeur α qui donne le meilleur ajustement pour la variable étudiée, on compare à l'aide d'Excel les écarts entre les données brutes et les données lissées pour différentes valeurs α , puis l'on choisit celle qui donne la plus petite somme pour les carrés des écarts (méthode des moindres carrés).

Lissage exponentiel ($\alpha = 0,40$)

On désigne la donnée de la période i par x_i et la donnée lissée par L_i .

On pose : $L_1 = x_1$

$$L_i = 0,40x_i + (1 - 0,40)L_{i-1}$$

D'où : $L_1 = 9\,194$

$$\begin{aligned} L_2 &= 0,40 \times 12\,115 + 0,60 \times 9\,194 \\ &= 10\,362,4 \end{aligned}$$

$$\begin{aligned} L_3 &= 0,40 \times 9\,878 + 0,60 \times 10\,362,4 \\ &= 10\,168,6 \end{aligned}$$

$$\begin{aligned} L_4 &= 0,40 \times 9\,634 + 0,60 \times 10\,168,6 \\ &= 9\,954,8 \end{aligned}$$

?

Calculer la donnée lissée pour l'hiver 2005 et la reporter dans le tableau 3.

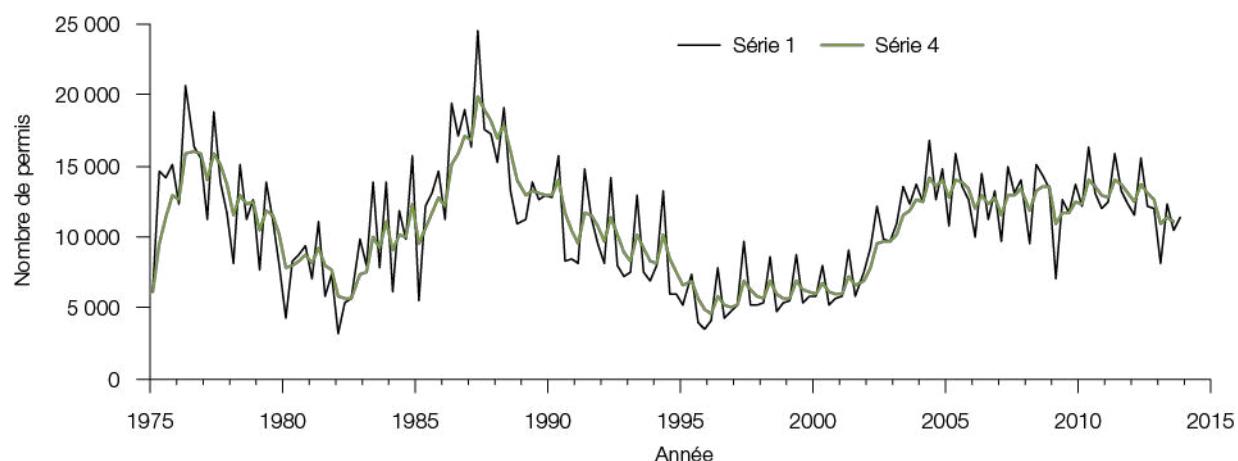
Le graphique ci-dessous permet de comparer les courbes des séries 1 et 4. On observe que le lissage exponentiel réagit plus rapidement aux modifications de la tendance à court terme que le lissage par moyenne mobile. Le lissage exponentiel est particulièrement utilisé pour faire des prévisions à court terme pour des séries chronologiques qui ne présentent ni tendance ni saisonnalité. On se sert alors de la donnée lissée L_i pour prédire la prochaine observation x_{i+1} ou, plus précisément, la prochaine valeur lissée L_{i+1} .

Tableau 3

Lissage exponentiel avec $\alpha = 0,40$

Année	Trimestre	Nombre de permis	$\alpha = 0,40$
		Série 1	Série 4
2002	1. Hiver	9 194	9 194
	2. Printemps	12 115	10 362,4
	3. Été	9 878	10 168,6
	4. Automne	9 634	9 954,8
	5. Hiver	10 948	10 352,1
	6. Printemps	13 596	11 649,6
	7. Été	12 266	11 896,2
	8. Automne	13 734	12 631,3
	9. Hiver	12 424	12 548,4
	10. Printemps	16 830	14 261,0
	11. Été	12 671	13 625,0
	12. Automne	14 730	14 067,0
2005	13. Hiver	10 774	_____
	14. Printemps	15 797	13 968,7
	15. Été	13 588	13 816,4
	16. Automne	12 685	13 363,8
2006	17. Hiver	10 058	12 041,5
	18. Printemps	14 453	13 006,1
	19. Été	11 312	12 328,5
	20. Automne	13 286	12 711,5

Évolution du nombre de permis de bâtir par trimestre, Québec, 1975-2013



8.4 Les séries désaisonnalisées

Une série dont on a éliminé l'effet de saisonnalité est appelée **série désaisonnalisée**. Il y a plusieurs façons d'obtenir une telle série; nous appliquons ici la **méthode du rapport à la moyenne mobile pondérée**.

MISE EN

SITUATION

Désaisonnalisons les données du nombre de permis de bâtir au Québec de 2002 à 2006. Ces données et les valeurs lissées obtenues avec une moyenne mobile pondérée de rayon 2 sont rappelées dans les trois premières colonnes du tableau 4 de la page suivante. Les trois dernières colonnes affichent les résultats des calculs présentés ci-dessous.

Pour désaisonnaliser des données, il faut d'abord calculer l'**indice de saisonnalité**. Ce dernier mesure l'influence de chaque saison sur la tendance à long terme. On obtient l'indice de saisonnalité, puis les données désaisonnalisées, en procédant comme suit.

- On calcule le rapport $\frac{\text{Donnée brute}}{\text{Donnée lissée}}$ pour chaque trimestre.

Par exemple, le rapport pour le printemps 2003 est $\frac{13\,596}{12\,123,5} = 1,121$.

Comme 12 123,5 correspond à la moyenne pondérée de cinq trimestres consécutifs centrés sur le printemps 2003, ce rapport indique que le nombre de permis délivrés au printemps 2003 correspond à 112,1 % de la moyenne des trimestres environnants, soit 12,1 % de plus que celle-ci.

- On calcule la moyenne des rapports de 2002 à 2006 pour le trimestre étudié. Cette moyenne est l'indice de saisonnalité pour le trimestre.

Par exemple, pour le printemps, on a les rapports suivants :

2002 : 1,220 2003 : 1,121 2004 : 1,199 2005 : 1,173 2006 : 1,184

$$\text{Indice de saisonnalité du printemps} = \frac{1,220 + 1,121 + 1,199 + 1,173 + 1,184}{5} = 1,179$$

Interprétation de l'indice de saisonnalité du printemps

De 2002 à 2006, le nombre de permis délivrés au printemps correspond environ à 117,9 % de la moyenne trimestrielle annuelle, soit 17,9 % de plus que celle-ci.

Nombre de permis délivrés au printemps $\approx 117,9\% \times$ moyenne trimestrielle de l'année

- On estime la moyenne trimestrielle de l'année, sur la base du printemps, en l'isolant dans l'équation précédente, ce qui donne le résultat suivant.

$$\begin{aligned}\text{Moyenne trimestrielle de l'année basée sur le printemps} &\approx \frac{\text{nombre de permis délivrés au printemps}}{117,9 \%} \\ &\approx \frac{\text{nombre de permis délivrés au printemps}}{1,179}\end{aligned}$$

Ce quotient est la **valeur désaisonnalisée** du nombre de permis délivrés au printemps pour l'année considérée. Cette valeur exclut l'effet de saisonnalité de la série, ce qui permet de comparer les valeurs désaisonnalisées de différentes périodes.

Par exemple, la valeur désaisonnalisée du nombre de permis délivrés au printemps 2003 est :

$$\text{Moyenne trimestrielle de l'année basée sur le printemps 2003} \approx \frac{13\,596}{1,179} = 11\,532$$

Tableau 4

Indices de saisonnalité et données désaisonnalisées du nombre de permis de bâtir, Québec, 2002-2006

Année	Trimestre	Nombre de permis	Moyenne mobile pondérée, $r = 2$	Donnée brute Donnée lissée	Indice de saisonnalité	Données désaisonnalisées
2002	Hiver	9 194	9 158,3 ¹	1,004	0,894	10 284
	Printemps	12 115	9 934,0 ¹	1,220	1,179	10 276
	Été	9 878	10 424,5	0,948	0,955	10 343
	Automne	9 634	10 828,9	0,890	1,013	9 510
2003	Hiver	10 948	11 312,5	0,968	0,894	12 246
	Printemps	13 596	12 123,5	1,121	1,179	11 532
	Été	12 266	12 820,5	0,957	0,955	12 844
	Automne	13 734	13 409,3	1,024	1,013	13 558
2004	Hiver	12 424	13 864,1	0,896	0,894	13 897
	Printemps	16 830	14 039,3	1,199	1,179	14 275
	Été	12 671	13 957,5	0,908	0,955	13 268
	Automne	14 730	13 622,1	1,081	1,013	14 541
2005	Hiver	10 774	13 607,6	0,792	0,894	12 051
	Printemps	15 797	13 466,6	1,173	1,179	13 399
	Été	13 588	13 121,5	1,036	0,955	14 228
	Automne	12 685	12 864,0	0,986	1,013	12 522
2006	Hiver	10 058	12 411,5	0,810	0,894	11 251
	Printemps	14 453	12 202,1	1,184	1,179	12 259
	Été	11 312	12 239,3 ²	0,924	0,955	11 845
	Automne	13 286	12 270,5 ²	1,083	1,013	13 115

1. Valeur basée sur les trimestres de 2001.

2. Valeur basée sur les trimestres de 2007.

Utilité d'une série désaisonnalisée

Dans le cas de données brutes, à cause de la saisonnalité, il est difficile de comparer les trimestres entre eux pour en dégager une tendance à court terme. Par contre, avec les données désaisonnalisées, on peut facilement mettre une telle tendance en évidence.

Par exemple, les données brutes du tableau 4 indiquent qu'on a délivré 13 596 permis au trimestre du printemps 2003 et 12 266 permis au trimestre suivant; doit-on en conclure que la tendance est à la baisse ? Bien au contraire, les données désaisonnalisées indiquent plutôt une tendance à la hausse puisque la moyenne trimestrielle est passée de 11 532 à 12 844 permis. On observe aussi une tendance à la hausse pendant 5 trimestres consécutifs, soit du printemps 2003 au printemps 2004.

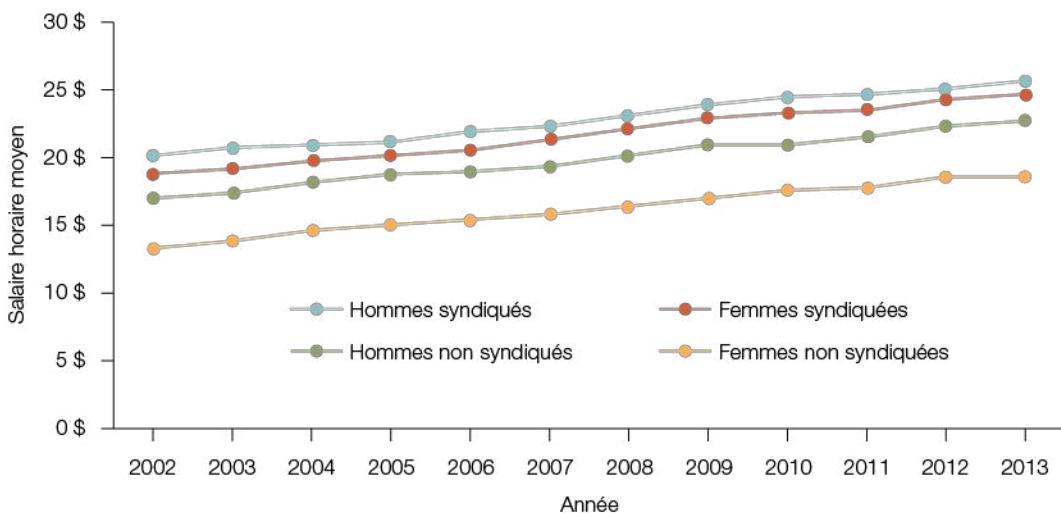
NOTE

La technique présentée pour désaisonnaliser les données s'applique aussi à des données lissées avec la méthode exponentielle.

EXERCICES 8.1

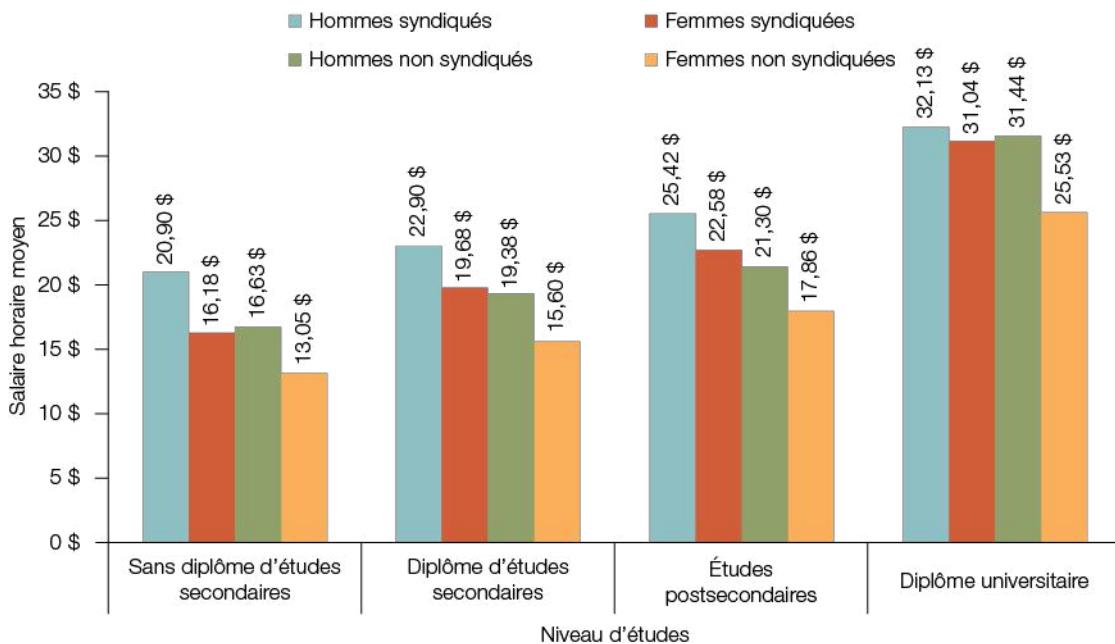
1. Les graphiques suivants présentent des statistiques sur le salaire horaire moyen des employés québécois.

Évolution du salaire horaire moyen des employés selon le sexe et l'adhésion ou non à un syndicat, Québec, 2002-2013



Source: Statistique Canada. *Enquête sur la population active*, 2013, adapté par l’Institut de la statistique du Québec, juin 2014.

Salaire horaire moyen des employés selon le sexe et l'adhésion ou non à un syndicat, par niveau d'études, Québec, 2013



Source: Statistique Canada. *Enquête sur la population active*, 2013, adapté par l’Institut de la statistique du Québec, juin 2014.

- Lequel des deux graphiques représente les données d'une série chronologique ?
- Vrai ou faux ? Selon les statistiques du 1^{er} graphique, de 2002 à 2012 :
 - peu importe le sexe, le salaire horaire moyen est plus élevé chez les employés syndiqués que chez les employés non syndiqués.
 - le salaire horaire moyen des femmes est toujours inférieur à celui des hommes.

iii) l'écart entre le salaire horaire moyen des hommes syndiqués et des femmes non syndiquées est environ deux fois plus grand que l'écart observé entre le salaire horaire moyen des hommes syndiqués et des hommes non syndiqués.

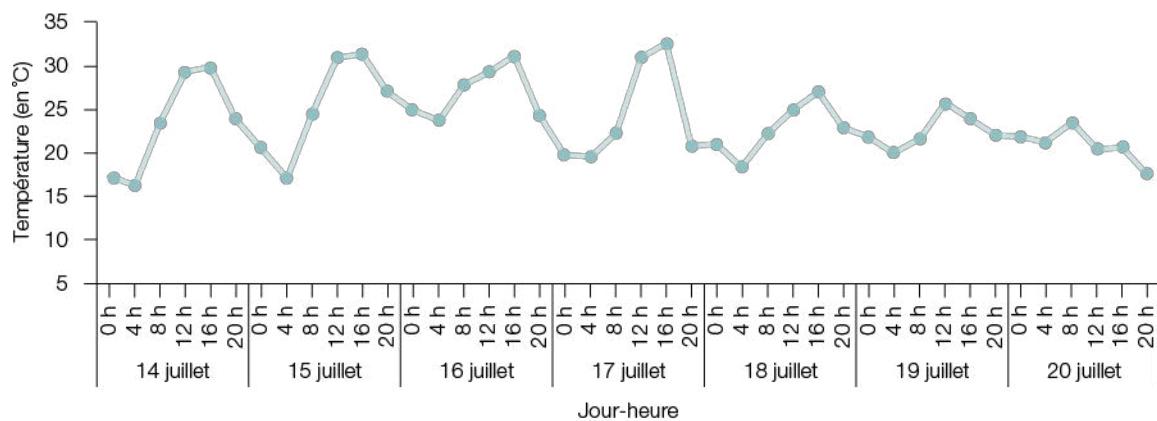
c) Vrai ou faux ? En 2012, selon les statistiques du 2^e graphique :

- i) quelle que soit la catégorie d'employés, les personnes sans diplôme d'études secondaires ont le salaire horaire moyen le plus bas.
- ii) quel que soit le niveau d'études des employés, les femmes non syndiquées ont le salaire horaire moyen le plus bas.
- iii) quel que soit le niveau d'études des employés, le salaire horaire moyen des femmes syndiquées est supérieur à celui des hommes non syndiqués.

2. Pour les séries chronologiques suivantes :

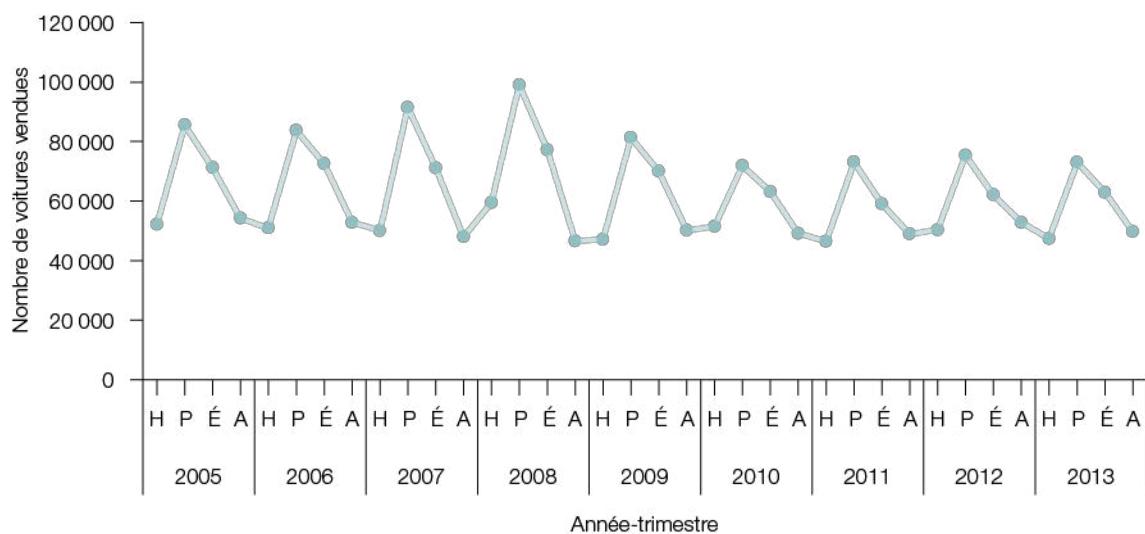
- tracer la forme graphique de la saisonnalité ;
- donner le cycle de la saisonnalité et son nombre de périodes ;
- décrire la variation saisonnière de la variable.

a) **Évolution de la température prise aux quatre heures, ville de Québec, semaine du 14 juillet 2013**



Source: Environnement Canada. *La météo au Canada. Rapports de données quotidiennes*.

b) **Évolution du nombre trimestriel de voitures neuves vendues, Québec, 2005-2013**



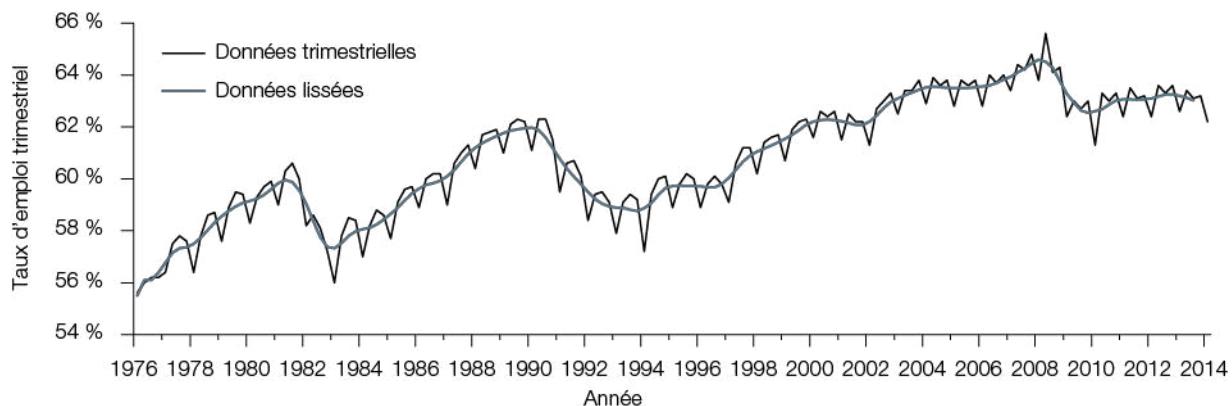
Source: Statistique Canada. *Tableau 079-0003, CANSIM*.

3. Soit le chronogramme suivant.

a) Donner la tendance à long terme du taux d'emploi trimestriel depuis 1976.

b) Indiquer les cycles du taux d'emploi trimestriel de 1976 à 2014.

Évolution du taux d'emploi¹ trimestriel, Canada, 1976-2014



1. Pourcentage des personnes qui ont un emploi parmi les Canadiens de 15 ans et plus.

Source: Statistique Canada. Tableau 282-0001, CANSIM.

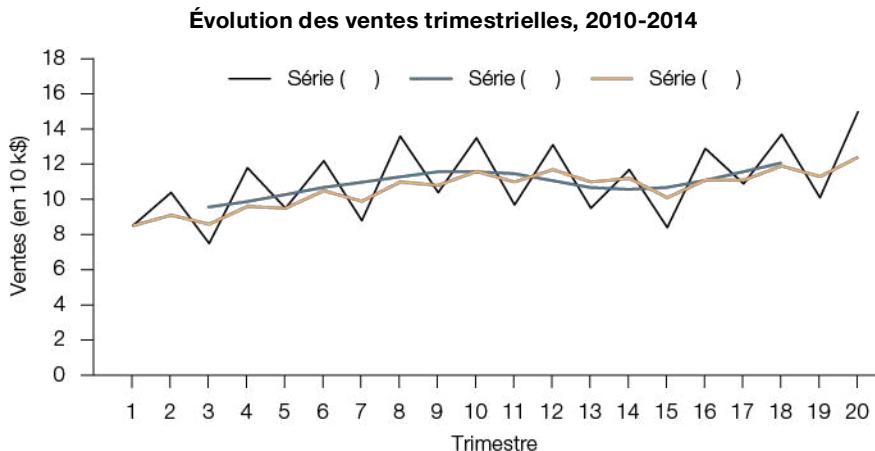
4. Le tableau suivant donne les ventes trimestrielles d'une entreprise sur une période de 5 ans (série 1). On a lissé ces données de deux façons :

- avec une moyenne mobile pondérée de rayon 2 (série 2);
- avec un lissage exponentiel où $\alpha = 0,38$ (série 3).

a) Compléter le tableau.

Année	Trimestre	Ventes (en 10 k\$)	Moyenne mobile pondérée $r = 2$	Lissage exponentiel $\alpha = 0,38$
		Série 1	Série 2	Série 3
2010	1. Hiver	8,5	—	—
	2. Printemps	10,4	—	9,2
	3. Été	7,5	9,7	8,6
	4. Automne	11,8	10,0	9,8
2011	5. Hiver	9,5	10,4	9,7
	6. Printemps	12,2	—	—
	7. Été	8,8	11,1	10,0
	8. Automne	13,6	11,4	11,4
2012	9. Hiver	10,4	—	—
	10. Printemps	13,5	11,7	12,0
	11. Été	9,7	11,6	11,1
	12. Automne	13,1	11,2	11,9
2013	13. Hiver	9,5	10,8	11,0
	14. Printemps	11,7	10,7	11,3
	15. Été	8,4	—	—
	16. Automne	12,9	11,2	11,2
2014	17. Hiver	10,9	11,7	11,1
	18. Printemps	13,7	12,2	12,1
	19. Été	10,1	—	11,3
	20. Automne	15,0	—	12,7

b) Compléter la légende du graphique qui suit en indiquant le numéro de la série.



c) Quel type de lissage fait le mieux ressortir la tendance de la série chronologique ?

5. a) Le tableau suivant donne l'indice de saisonnalité et les données désaisonnalisées de la série chronologique étudiée au numéro 4. Le compléter.

Année	Trimestre	Ventes (en 10 k\$)	Moyenne mobile pondérée, $r = 2$	Donnée brute Donnée lissée	Indice de saisonnalité	Données désaisonnalisées
2010	1. Hiver	8,5	—	—	—	—
	2. Printemps	10,4	—	—	—	—
	3. Été	7,5	9,7	0,773	0,796	9,4
	4. Automne	11,8	10,0	—	—	—
2011	5. Hiver	9,5	10,4	—	—	—
	6. Printemps	12,2	10,8	1,130	1,125	10,8
	7. Été	8,8	11,1	0,793	0,796	11,1
	8. Automne	13,6	11,4	1,193	—	—
2012	9. Hiver	10,4	11,7	0,889	—	—
	10. Printemps	13,5	11,7	1,154	1,125	12,0
	11. Été	9,7	11,6	0,836	0,796	12,2
	12. Automne	13,1	11,2	1,170	—	—
2013	13. Hiver	9,5	10,8	0,880	—	—
	14. Printemps	11,7	10,7	1,093	1,125	10,4
	15. Été	8,4	10,8	0,778	0,796	10,6
	16. Automne	12,9	11,2	1,152	—	—
2014	17. Hiver	10,9	11,7	0,932	—	—
	18. Printemps	13,7	12,2	1,123	1,125	12,2
	19. Été	10,1	—	—	—	—
	20. Automne	15,0	—	—	—	—

b) Interpréter l'indice de saisonnalité du trimestre d'été.

c) En 2011, les ventes du trimestre d'été se sont élevées à 88 000 \$, alors qu'elles avaient été de 122 000 \$ au trimestre précédent. Ces chiffres indiquent-ils une tendance à la baisse ou à la hausse ?

d) À partir de quel trimestre les ventes ont-elles commencé à baisser ? À quel trimestre y a-t-il eu une remontée ?

Série chronologique

Une série chronologique donne l'évolution d'une variable dans le temps. Elle peut comporter une ou plusieurs des caractéristiques suivantes : une tendance à long terme de forme linéaire ou curviligne, des variations cycliques, saisonnières, uniques ou aléatoires.

Lissage d'une série chronologique

On lisse une série chronologique pour obtenir une série moins dentelée que la série brute. Le lissage permet de mettre en évidence la tendance et les cycles de la série en masquant les cycles saisonniers.

Lissage par moyenne mobile de rayon r

Pour une série qui présente une saisonnalité, on choisit une valeur pour r qui fait en sorte qu'une fenêtre de rayon r centrée sur la donnée à lisser couvre ou dépasse légèrement la saisonnalité.

- Moyenne non pondérée: $L_i = \frac{x_{i-r} + \dots + x_i + \dots + x_{i+r}}{2r+1}$
- Moyenne pondérée: $L_i = \frac{0,5x_{i-r} + x_{i-r+1} + \dots + x_i + \dots + x_{i-r-1} + 0,5x_{i+r}}{2r}$ (pondération totale)

Lissage exponentiel

- $L_1 = x_1$
- $L_i = \alpha x_i + (1 - \alpha) L_{i-1}$

NOTE

Plus la valeur de α est grande, plus on accorde un poids important à la période à lisser par rapport aux périodes précédentes.

Série désaisonnalisée

Une série désaisonnalisée exclut l'effet de saisonnalité de la série brute, de sorte que ses données sont comparables d'une période à l'autre. Voici la marche à suivre pour désaisonnaliser une série :

- Calculer le rapport $\frac{\text{Donnée brute}}{\text{Donnée lissée}}$ pour chaque période de la série.
- Calculer l'indice de saisonnalité pour chaque période du cycle saisonnier. Cet indice correspond à la moyenne des rapports calculés en 1 sur la période de temps considérée.
- Calculer les données désaisonnalisées en divisant les données brutes par l'indice de saisonnalité pour la période étudiée.

EXERCICES RÉCAPITULATIFS

Les exercices qui suivent sont consacrés à l'analyse du taux de chômage chez les jeunes Québécois de 15 à 24 ans. Le taux de chômage est le pourcentage de personnes sans travail dans la population active, cette dernière étant constituée des travailleurs et des chômeurs âgés de 15 à 24 ans.

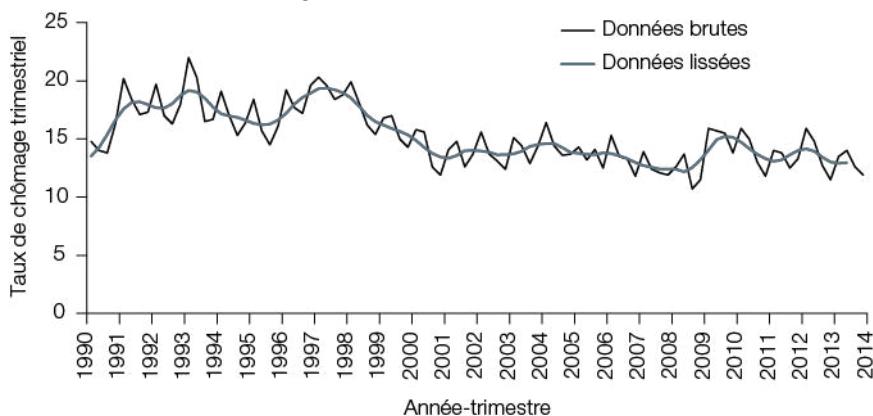
Taux de chômage trimestriel brut et désaisonnalisé chez les 15 à 24 ans, Québec, 2008-2013

Année	Trimestre	Taux brut (%)	Moyenne mobile pondérée, $r = 2$	Donnée brute Donnée lissée	Indice de saisonnalité	Taux désaisonnalisé
2008	Hiver	12,6	—	—	—	—
	Printemps	13,7	—	—	—	—
	Été	10,7	12,538	0,853	—	—
	Automne	11,5	13,200	0,871	0,901	12,8
2009	Hiver	15,9	14,050	1,132	1,090	14,6
	Printemps	15,7	14,938	1,051	1,059	14,8
	Été	15,5	15,225	1,018	—	—
	Automne	13,8	15,138	0,912	0,901	15,3
2010	Hiver	15,9	14,738	1,079	1,090	14,6
	Printemps	15,0	14,175	1,058	1,059	14,2
	Été	13,0	13,688	0,950	—	—
	Automne	11,8	13,300	0,887	0,901	13,1
2011	Hiver	14,0	13,088	1,070	1,090	12,8
	Printemps	13,8	13,213	1,044	1,059	13,0
	Été	12,5	—	—	—	—
	Automne	13,3	14,000	0,950	0,901	14,8
2012	Hiver	15,9	14,150	1,124	1,090	14,6
	Printemps	14,8	13,950	1,061	1,059	14,0
	Été	12,7	13,425	0,946	—	—
	Automne	11,5	13,025	0,883	0,901	12,8
2013	Hiver	13,5	12,913	1,045	1,090	12,4
	Printemps	14,0	12,950	1,081	1,059	13,2
	Été	12,6	—	—	—	—
	Automne	11,9	—	—	—	—

Source: Statistique Canada. *Tableau 282-0001, CANSIM*.

1. a) Compléter le tableau.
 b) De 2008 à 2013, quel est le taux de chômage trimestriel le plus élevé ? le plus faible ?
 c) Compléter l'énoncé. Pour ___ des 6 années considérées, c'est à l'automne que le taux de chômage est le plus bas.
2. a) En 2011, le taux de chômage est passé de 14,0 % au trimestre d'hiver à 13,8 % au trimestre suivant. Est-ce un indice d'une tendance à la baisse du chômage ?
 b) Le taux de chômage est passé de 13,8 % au trimestre d'automne 2009 à 15,9 % à l'hiver 2010. Est-ce un indice d'une tendance à la hausse du chômage ?
 c) Interpréter l'indice de saisonnalité du trimestre d'hiver.
3. a) Si l'on avait utilisé une moyenne mobile non pondérée de rayon 2 pour lisser la série chronologique, quelle valeur aurait-on obtenue comme taux de chômage moyen pour le trimestre d'été 2011 ?
 e) Si l'on avait effectué un lissage exponentiel, avec $\alpha = 0,36$, quelle valeur aurait-on obtenue pour chacun des quatre trimestres de l'année 2008 ?
- a) Le graphique suivant présente le taux de chômage trimestriel chez les Québécois de 15 à 24 ans, de 1990 à 2013. Pour faciliter l'analyse, on a effectué un lissage de la série avec une moyenne mobile pondérée de rayon 2. Décrire la tendance à long terme du taux de chômage trimestriel pour la période étudiée.

Évolution du taux de chômage trimestriel chez les 15 à 24 ans, Québec, 1990-2013

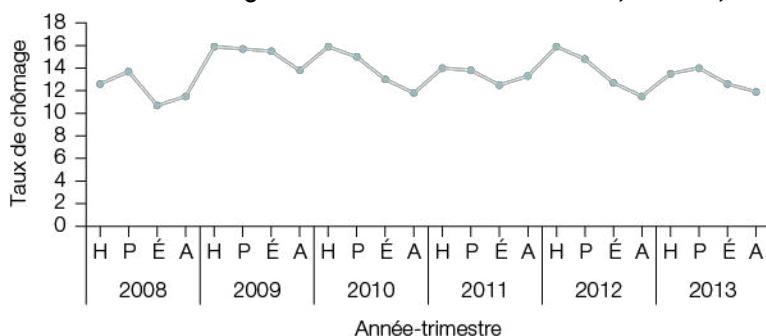


Source: Statistique Canada. Tableau 282-0001, CANSIM.

- b) À l'aide du graphique suivant, étudier la saisonnalité du taux trimestriel de chômage chez les Québécois de 15 à 24 ans de 2008 à 2013 :

- tracer la forme graphique de la saisonnalité ;
- donner le cycle de la saisonnalité et son nombre de périodes ;
- décrire la variation saisonnière de la variable.

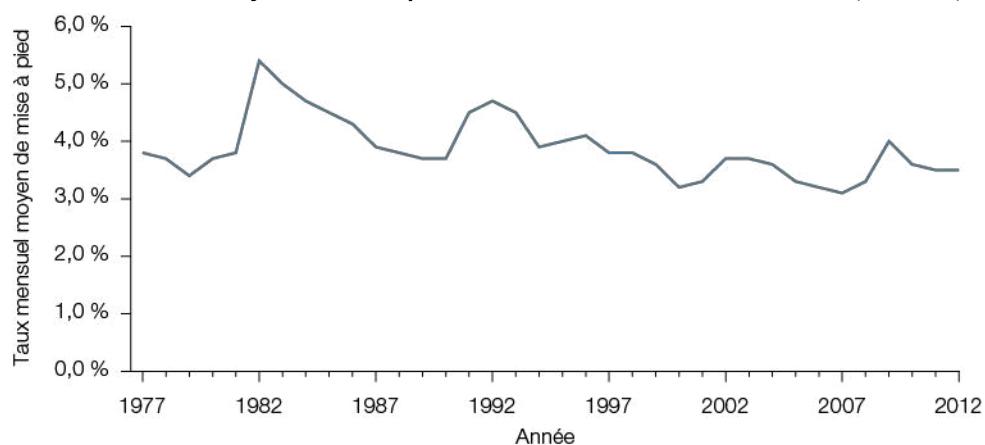
Évolution du taux de chômage trimestriel chez les 15 à 24 ans, Québec, 2008-2013



Source: Statistique Canada. Tableau 282-0001, CANSIM.

4. Indiquer les cycles de la série chronologique représentée par le chronogramme suivant.

Évolution du taux mensuel moyen de mise à pied¹ chez les travailleurs de 15 à 24 ans, Canada, 1977-2012



1. Les personnes mises à pied sont celles qui étaient en emploi au cours d'un mois puis sans emploi (soit en chômage, soit hors de la population active) le mois suivant, et qui avaient déclaré qu'une mise à pied avait été à l'origine de la fin de leur emploi.

Source: Statistique Canada. Aperçus économiques : La dynamique du chômage chez les jeunes Canadiens, n° 24, 2013.

PRÉPARATION À L'EXAMEN

Pour préparer votre examen, assurez-vous d'avoir les compétences suivantes :

Si vous avez la compétence, cochez.	
Série chronologique	
• Reconnaître une série chronologique et savoir la représenter.	<input type="radio"/>
• Déterminer la tendance, les cycles et la saisonnalité d'une série chronologique.	<input type="radio"/>
Lissage d'une série chronologique	
• Lisser une série chronologique en utilisant une moyenne mobile pondérée ou non.	<input type="radio"/>
• Lisser une série chronologique en effectuant un lissage exponentiel.	<input type="radio"/>
Série désaisonnalisée	
• Construire une série désaisonnalisée.	<input type="radio"/>
• Analyser la tendance d'une série à l'aide des données désaisonnalisées.	<input type="radio"/>

ANNEXE 1

Table de la loi binomiale

n	x	p									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1	0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
3	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
	1	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
8	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
	2	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094

n	x	p									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
3	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039
9	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	
11	0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
	1	0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054
	2	0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0513	0,0269
	3	0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806
	4	0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611
	5	0,0001	0,0025	0,0132	0,0388	0,0803	0,1321	0,1830	0,2207	0,2360	0,2256
	6	0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256
	7	0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611
	8	0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806
	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005	
12	0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
	1	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
	2	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
	3	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
	4	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
	5	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
	6	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
	7	0,0000	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934

n	x	p									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
8	8	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
	9	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0068	0,0161
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	
13	0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
	1	0,3512	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016
	2	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095
	3	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349
	4	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873
	5	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571
	6	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095
	7	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095
	8	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571
	9	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873
	10	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
14	0	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
	1	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009
	2	0,1229	0,2570	0,2912	0,2501	0,1802	0,1134	0,0634	0,0317	0,0141	0,0056
	3	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222
	4	0,0037	0,0349	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611
	5	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222
	6	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833
	7	0,0000	0,0002	0,0019	0,0092	0,0280	0,0618	0,1082	0,1574	0,1952	0,2095
	8	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833
	9	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222
	10	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
	2	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
	3	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
	4	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1268	0,0780	0,0417
	5	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916
	6	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
	7	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1964
	8	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
	9	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527
	10	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032

n	x	p									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
14	0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
16	0	0,4401	0,1853	0,0743	0,0281	0,0100	0,0033	0,0010	0,0003	0,0001	0,0000
	1	0,3706	0,3294	0,2097	0,1126	0,0535	0,0228	0,0087	0,0030	0,0009	0,0002
	2	0,1463	0,2745	0,2775	0,2111	0,1336	0,0732	0,0353	0,0150	0,0056	0,0018
	3	0,0359	0,1423	0,2285	0,2463	0,2079	0,1465	0,0888	0,0468	0,0215	0,0085
	4	0,0061	0,0514	0,1311	0,2001	0,2252	0,2040	0,1553	0,1014	0,0572	0,0278
	5	0,0008	0,0137	0,0555	0,1201	0,1802	0,2099	0,2008	0,1623	0,1123	0,0667
	6	0,0001	0,0028	0,0180	0,0550	0,1101	0,1649	0,1982	0,1983	0,1684	0,1222
	7	0,0000	0,0004	0,0045	0,0197	0,0524	0,1010	0,1524	0,1889	0,1969	0,1746
	8	0,0000	0,0001	0,0009	0,0055	0,0197	0,0487	0,0923	0,1417	0,1812	0,1964
	9	0,0000	0,0000	0,0001	0,0012	0,0058	0,0185	0,0442	0,0840	0,1318	0,1746
	10	0,0000	0,0000	0,0000	0,0002	0,0014	0,0056	0,0167	0,0392	0,0755	0,1222
	11	0,0000	0,0000	0,0000	0,0000	0,0002	0,0013	0,0049	0,0142	0,0337	0,0667
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0040	0,0115	0,0278
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0029	0,0085
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
17	0	0,4181	0,1668	0,0631	0,0225	0,0075	0,0023	0,0007	0,0002	0,0000	0,0000
	1	0,3741	0,3150	0,1893	0,0957	0,0426	0,0169	0,0060	0,0019	0,0005	0,0001
	2	0,1575	0,2800	0,2673	0,1914	0,1136	0,0581	0,0260	0,0102	0,0035	0,0010
	3	0,0415	0,1556	0,2359	0,2393	0,1893	0,1245	0,0701	0,0341	0,0144	0,0052
	4	0,0076	0,0605	0,1457	0,2093	0,2209	0,1868	0,1320	0,0796	0,0411	0,0182
	5	0,0010	0,0175	0,0668	0,1361	0,1914	0,2081	0,1849	0,1379	0,0875	0,0472
	6	0,0001	0,0039	0,0236	0,0680	0,1276	0,1784	0,1991	0,1839	0,1432	0,0944
	7	0,0000	0,0007	0,0065	0,0267	0,0668	0,1201	0,1685	0,1927	0,1841	0,1484
	8	0,0000	0,0001	0,0014	0,0084	0,0279	0,0644	0,1134	0,1606	0,1883	0,1855
	9	0,0000	0,0000	0,0003	0,0021	0,0093	0,0276	0,0611	0,1070	0,1540	0,1855
	10	0,0000	0,0000	0,0000	0,0004	0,0025	0,0095	0,0263	0,0571	0,1008	0,1484
	11	0,0000	0,0000	0,0000	0,0001	0,0005	0,0026	0,0090	0,0242	0,0525	0,0944
	12	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0081	0,0215	0,0472
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0021	0,0068	0,0182
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
18	0	0,3972	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
	1	0,3763	0,3002	0,1704	0,0811	0,0338	0,0126	0,0042	0,0012	0,0003	0,0001
	2	0,1683	0,2835	0,2556	0,1723	0,0958	0,0458	0,0190	0,0069	0,0022	0,0006
	3	0,0473	0,1680	0,2406	0,2297	0,1704	0,1046	0,0547	0,0246	0,0095	0,0031
	4	0,0093	0,0700	0,1592	0,2153	0,2130	0,1681	0,1104	0,0614	0,0291	0,0117
	5	0,0014	0,0218	0,0787	0,1507	0,1988	0,2017	0,1664	0,1146	0,0666	0,0327
	6	0,0002	0,0052	0,0301	0,0816	0,1436	0,1873	0,1941	0,1655	0,1181	0,0708
	7	0,0000	0,0010	0,0091	0,0350	0,0820	0,1376	0,1792	0,1892	0,1657	0,1214
	8	0,0000	0,0002	0,0022	0,0120	0,0376	0,0811	0,1327	0,1734	0,1864	0,1669
	9	0,0000	0,0000	0,0004	0,0033	0,0139	0,0386	0,0794	0,1284	0,1694	0,1855
	10	0,0000	0,0000	0,0001	0,0008	0,0042	0,0149	0,0385	0,0771	0,1248	0,1669

n	x	P									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
19	11	0,0000	0,0000	0,0000	0,0001	0,0010	0,0046	0,0151	0,0374	0,0742	0,1214
	12	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0047	0,0145	0,0354	0,0708
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0045	0,0134	0,0327
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0039	0,0117
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0031
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	0	0,3774	0,1351	0,0456	0,0144	0,0042	0,0011	0,0003	0,0001	0,0000	0,0000
	1	0,3774	0,2852	0,1529	0,0685	0,0268	0,0093	0,0029	0,0008	0,0002	0,0000
	2	0,1787	0,2852	0,2428	0,1540	0,0803	0,0358	0,0138	0,0046	0,0013	0,0003
	3	0,0533	0,1796	0,2428	0,2182	0,1517	0,0869	0,0422	0,0175	0,0062	0,0018
	4	0,0112	0,0798	0,1714	0,2182	0,2023	0,1491	0,0909	0,0467	0,0203	0,0074
	5	0,0018	0,0266	0,0907	0,1636	0,2023	0,1916	0,1468	0,0933	0,0497	0,0222
	6	0,0002	0,0069	0,0374	0,0955	0,1574	0,1916	0,1844	0,1451	0,0949	0,0518
	7	0,0000	0,0014	0,0122	0,0443	0,0974	0,1525	0,1844	0,1797	0,1443	0,0961
	8	0,0000	0,0002	0,0032	0,0166	0,0487	0,0981	0,1489	0,1797	0,1771	0,1442
	9	0,0000	0,0000	0,0007	0,0051	0,0198	0,0514	0,0980	0,1464	0,1771	0,1762
	10	0,0000	0,0000	0,0001	0,0013	0,0066	0,0220	0,0528	0,0976	0,1449	0,1762
	11	0,0000	0,0000	0,0000	0,0003	0,0018	0,0077	0,0233	0,0532	0,0970	0,1442
	12	0,0000	0,0000	0,0000	0,0000	0,0004	0,0022	0,0083	0,0237	0,0529	0,0961
	13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0024	0,0085	0,0233	0,0518
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0082	0,0222
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0022	0,0074
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	0	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
	1	0,3774	0,2702	0,1368	0,0576	0,0211	0,0068	0,0020	0,0005	0,0001	0,0000
	2	0,1887	0,2852	0,2293	0,1369	0,0669	0,0278	0,0100	0,0031	0,0008	0,0002
	3	0,0596	0,1901	0,2428	0,2054	0,1339	0,0716	0,0323	0,0123	0,0040	0,0011
	4	0,0133	0,0898	0,1821	0,2182	0,1897	0,1304	0,0738	0,0350	0,0139	0,0046
	5	0,0022	0,0319	0,1028	0,1746	0,2023	0,1789	0,1272	0,0746	0,0365	0,0148
	6	0,0003	0,0089	0,0454	0,1091	0,1686	0,1916	0,1712	0,1244	0,0746	0,0370
	7	0,0000	0,0020	0,0160	0,0545	0,1124	0,1643	0,1844	0,1659	0,1221	0,0739
	8	0,0000	0,0004	0,0046	0,0222	0,0609	0,1144	0,1614	0,1797	0,1623	0,1201
	9	0,0000	0,0001	0,0011	0,0074	0,0271	0,0654	0,1158	0,1597	0,1771	0,1602
	10	0,0000	0,0000	0,0002	0,0020	0,0099	0,0308	0,0686	0,1171	0,1593	0,1762
	11	0,0000	0,0000	0,0000	0,0005	0,0030	0,0120	0,0336	0,0710	0,1185	0,1602
	12	0,0000	0,0000	0,0000	0,0001	0,0008	0,0039	0,0136	0,0355	0,0727	0,1201
	13	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0045	0,0146	0,0366	0,0739
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0049	0,0150	0,0370
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0049	0,0148
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0046
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

ANNEXE 2

Table de la loi de Poisson

	λ									
x	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0000	0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020	0,0031
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	λ									
x	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496	0,1353
1	0,3662	0,3614	0,3543	0,3452	0,3347	0,3230	0,3106	0,2975	0,2842	0,2707
2	0,2014	0,2169	0,2303	0,2417	0,2510	0,2584	0,2640	0,2678	0,2700	0,2707
3	0,0738	0,0867	0,0998	0,1128	0,1255	0,1378	0,1496	0,1607	0,1710	0,1804
4	0,0203	0,0260	0,0324	0,0395	0,0471	0,0551	0,0636	0,0723	0,0812	0,0902
5	0,0045	0,0062	0,0084	0,0111	0,0141	0,0176	0,0216	0,0260	0,0309	0,0361
6	0,0008	0,0012	0,0018	0,0026	0,0035	0,0047	0,0061	0,0078	0,0098	0,0120
7	0,0001	0,0002	0,0003	0,0005	0,0008	0,0011	0,0015	0,0020	0,0027	0,0034
8	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002	0,0003	0,0005	0,0006	0,0009
9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002
	λ									
x	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608	0,0550	0,0498
1	0,2572	0,2438	0,2306	0,2177	0,2052	0,1931	0,1815	0,1703	0,1596	0,1494
2	0,2700	0,2681	0,2652	0,2613	0,2565	0,2510	0,2450	0,2384	0,2314	0,2240
3	0,1890	0,1966	0,2033	0,2090	0,2138	0,2176	0,2205	0,2225	0,2237	0,2240
4	0,0992	0,1082	0,1169	0,1254	0,1336	0,1414	0,1488	0,1557	0,1622	0,1680
5	0,0417	0,0476	0,0538	0,0602	0,0668	0,0735	0,0804	0,0872	0,0940	0,1008
6	0,0146	0,0174	0,0206	0,0241	0,0278	0,0319	0,0362	0,0407	0,0455	0,0504
7	0,0044	0,0055	0,0068	0,0083	0,0099	0,0118	0,0139	0,0163	0,0188	0,0216
8	0,0011	0,0015	0,0019	0,0025	0,0031	0,0038	0,0047	0,0057	0,0068	0,0081
9	0,0003	0,0004	0,0005	0,0007	0,0009	0,0011	0,0014	0,0018	0,0022	0,0027
10	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0004	0,0005	0,0006	0,0008
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002	0,0002
12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	λ									
x	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
0	0,0450	0,0408	0,0369	0,0334	0,0302	0,0273	0,0247	0,0224	0,0202	0,0183
1	0,1397	0,1304	0,1217	0,1135	0,1057	0,0984	0,0915	0,0850	0,0789	0,0733
2	0,2165	0,2087	0,2008	0,1929	0,1850	0,1771	0,1692	0,1615	0,1539	0,1465

x	λ									
x	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
3	0,2237	0,2226	0,2209	0,2186	0,2158	0,2125	0,2087	0,2046	0,2001	0,1954
4	0,1733	0,1781	0,1823	0,1858	0,1888	0,1912	0,1931	0,1944	0,1951	0,1954
5	0,1075	0,1140	0,1203	0,1264	0,1322	0,1377	0,1429	0,1477	0,1522	0,1563
6	0,0555	0,0608	0,0662	0,0716	0,0771	0,0826	0,0881	0,0936	0,0989	0,1042
7	0,0246	0,0278	0,0312	0,0348	0,0385	0,0425	0,0466	0,0508	0,0551	0,0595
8	0,0095	0,0111	0,0129	0,0148	0,0169	0,0191	0,0215	0,0241	0,0269	0,0298
9	0,0033	0,0040	0,0047	0,0056	0,0066	0,0076	0,0089	0,0102	0,0116	0,0132
10	0,0010	0,0013	0,0016	0,0019	0,0023	0,0028	0,0033	0,0039	0,0045	0,0053
11	0,0003	0,0004	0,0005	0,0006	0,0007	0,0009	0,0011	0,0013	0,0016	0,0019
12	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006
13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001
x	λ									
x	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
0	0,0166	0,0150	0,0136	0,0123	0,0111	0,0101	0,0091	0,0082	0,0074	0,0067
1	0,0679	0,0630	0,0583	0,0540	0,0500	0,0462	0,0427	0,0395	0,0365	0,0337
2	0,1393	0,1323	0,1254	0,1188	0,1125	0,1063	0,1005	0,0948	0,0894	0,0842
3	0,1904	0,1852	0,1798	0,1743	0,1687	0,1631	0,1574	0,1517	0,1460	0,1404
4	0,1951	0,1944	0,1933	0,1917	0,1898	0,1875	0,1849	0,1820	0,1789	0,1755
5	0,1600	0,1633	0,1662	0,1687	0,1708	0,1725	0,1738	0,1747	0,1753	0,1755
6	0,1093	0,1143	0,1191	0,1237	0,1281	0,1323	0,1362	0,1398	0,1432	0,1462
7	0,0640	0,0686	0,0732	0,0778	0,0824	0,0869	0,0914	0,0959	0,1002	0,1044
8	0,0328	0,0360	0,0393	0,0428	0,0463	0,0500	0,0537	0,0575	0,0614	0,0653
9	0,0150	0,0168	0,0188	0,0209	0,0232	0,0255	0,0281	0,0307	0,0334	0,0363
10	0,0061	0,0071	0,0081	0,0092	0,0104	0,0118	0,0132	0,0147	0,0164	0,0181
11	0,0023	0,0027	0,0032	0,0037	0,0043	0,0049	0,0056	0,0064	0,0073	0,0082
12	0,0008	0,0009	0,0011	0,0013	0,0016	0,0019	0,0022	0,0026	0,0030	0,0034
13	0,0002	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013
14	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	
15	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002
x	λ									
x	5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	6,0
0	0,0061	0,0055	0,0050	0,0045	0,0041	0,0037	0,0033	0,0030	0,0027	0,0025
1	0,0311	0,0287	0,0265	0,0244	0,0225	0,0207	0,0191	0,0176	0,0162	0,0149
2	0,0793	0,0746	0,0701	0,0659	0,0618	0,0580	0,0544	0,0509	0,0477	0,0446
3	0,1348	0,1293	0,1239	0,1185	0,1133	0,1082	0,1033	0,0985	0,0938	0,0892
4	0,1719	0,1681	0,1641	0,1600	0,1558	0,1515	0,1472	0,1428	0,1383	0,1339
5	0,1753	0,1748	0,1740	0,1728	0,1714	0,1697	0,1678	0,1656	0,1632	0,1606
6	0,1490	0,1515	0,1537	0,1555	0,1571	0,1584	0,1594	0,1601	0,1605	0,1606
7	0,1086	0,1125	0,1163	0,1200	0,1234	0,1267	0,1298	0,1326	0,1353	0,1377
8	0,0692	0,0731	0,0771	0,0810	0,0849	0,0887	0,0925	0,0962	0,0998	0,1033
9	0,0392	0,0423	0,0454	0,0486	0,0519	0,0552	0,0586	0,0620	0,0654	0,0688
10	0,0200	0,0220	0,0241	0,0262	0,0285	0,0309	0,0334	0,0359	0,0386	0,0413

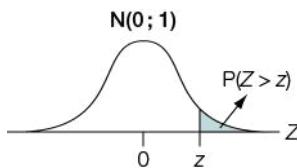
λ										
x	5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	6,0
11	0,0093	0,0104	0,0116	0,0129	0,0143	0,0157	0,0173	0,0190	0,0207	0,0225
12	0,0039	0,0045	0,0051	0,0058	0,0065	0,0073	0,0082	0,0092	0,0102	0,0113
13	0,0015	0,0018	0,0021	0,0024	0,0028	0,0032	0,0036	0,0041	0,0046	0,0052
14	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013	0,0015	0,0017	0,0019	0,0022
15	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009
16	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001
λ										
x	6,1	6,2	6,3	6,4	6,5	6,6	6,7	6,8	6,9	7,0
0	0,0022	0,0020	0,0018	0,0017	0,0015	0,0014	0,0012	0,0011	0,0010	0,0009
1	0,0137	0,0126	0,0116	0,0106	0,0098	0,0090	0,0082	0,0076	0,0070	0,0064
2	0,0417	0,0390	0,0364	0,0340	0,0318	0,0296	0,0276	0,0258	0,0240	0,0223
3	0,0848	0,0806	0,0765	0,0726	0,0688	0,0652	0,0617	0,0584	0,0552	0,0521
4	0,1294	0,1249	0,1205	0,1162	0,1118	0,1076	0,1034	0,0992	0,0952	0,0912
5	0,1579	0,1549	0,1519	0,1487	0,1454	0,1420	0,1385	0,1349	0,1314	0,1277
6	0,1605	0,1601	0,1595	0,1586	0,1575	0,1562	0,1546	0,1529	0,1511	0,1490
7	0,1399	0,1418	0,1435	0,1450	0,1462	0,1472	0,1480	0,1486	0,1489	0,1490
8	0,1066	0,1099	0,1130	0,1160	0,1188	0,1215	0,1240	0,1263	0,1284	0,1304
9	0,0723	0,0757	0,0791	0,0825	0,0858	0,0891	0,0923	0,0954	0,0985	0,1014
10	0,0441	0,0469	0,0498	0,0528	0,0558	0,0588	0,0618	0,0649	0,0679	0,0710
11	0,0244	0,0265	0,0285	0,0307	0,0330	0,0353	0,0377	0,0401	0,0426	0,0452
12	0,0124	0,0137	0,0150	0,0164	0,0179	0,0194	0,0210	0,0227	0,0245	0,0263
13	0,0058	0,0065	0,0073	0,0081	0,0089	0,0099	0,0108	0,0119	0,0130	0,0142
14	0,0025	0,0029	0,0033	0,0037	0,0041	0,0046	0,0052	0,0058	0,0064	0,0071
15	0,0010	0,0012	0,0014	0,0016	0,0018	0,0020	0,0023	0,0026	0,0029	0,0033
16	0,0004	0,0005	0,0005	0,0006	0,0007	0,0008	0,0010	0,0011	0,0013	0,0014
17	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006
18	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001
λ										
x	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
0	0,0008	0,0007	0,0007	0,0006	0,0006	0,0005	0,0005	0,0004	0,0004	0,0003
1	0,0059	0,0054	0,0049	0,0045	0,0041	0,0038	0,0035	0,0032	0,0029	0,0027
2	0,0208	0,0194	0,0180	0,0167	0,0156	0,0145	0,0134	0,0125	0,0116	0,0107
3	0,0492	0,0464	0,0438	0,0413	0,0389	0,0366	0,0345	0,0324	0,0305	0,0286
4	0,0874	0,0836	0,0799	0,0764	0,0729	0,0696	0,0663	0,0632	0,0602	0,0573
5	0,1241	0,1204	0,1167	0,1130	0,1094	0,1057	0,1021	0,0986	0,0951	0,0916
6	0,1468	0,1445	0,1420	0,1394	0,1367	0,1339	0,1311	0,1282	0,1252	0,1221
7	0,1489	0,1486	0,1481	0,1474	0,1465	0,1454	0,1442	0,1428	0,1413	0,1396
8	0,1321	0,1337	0,1351	0,1363	0,1373	0,1381	0,1388	0,1392	0,1395	0,1396
9	0,1042	0,1070	0,1096	0,1121	0,1144	0,1167	0,1187	0,1207	0,1224	0,1241
10	0,0740	0,0770	0,0800	0,0829	0,0858	0,0887	0,0914	0,0941	0,0967	0,0993
11	0,0478	0,0504	0,0531	0,0558	0,0585	0,0613	0,0640	0,0667	0,0695	0,0722
12	0,0283	0,0303	0,0323	0,0344	0,0366	0,0388	0,0411	0,0434	0,0457	0,0481
13	0,0154	0,0168	0,0181	0,0196	0,0211	0,0227	0,0243	0,0260	0,0278	0,0296
14	0,0078	0,0086	0,0095	0,0104	0,0113	0,0123	0,0134	0,0145	0,0157	0,0169
15	0,0037	0,0041	0,0046	0,0051	0,0057	0,0062	0,0069	0,0075	0,0083	0,0090
16	0,0016	0,0019	0,0021	0,0024	0,0026	0,0030	0,0033	0,0037	0,0041	0,0045

x	λ									
x	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
17	0,0007	0,0008	0,0009	0,0010	0,0012	0,0013	0,0015	0,0017	0,0019	0,0021
18	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
19	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0003	0,0004
20	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002
21	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001
x	λ									
x	8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,9	9,0
0	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001
1	0,0025	0,0023	0,0021	0,0019	0,0017	0,0016	0,0014	0,0013	0,0012	0,0011
2	0,0100	0,0092	0,0086	0,0079	0,0074	0,0068	0,0063	0,0058	0,0054	0,0050
3	0,0269	0,0252	0,0237	0,0222	0,0208	0,0195	0,0183	0,0171	0,0160	0,0150
4	0,0544	0,0517	0,0491	0,0466	0,0443	0,0420	0,0398	0,0377	0,0357	0,0337
5	0,0882	0,0849	0,0816	0,0784	0,0752	0,0722	0,0692	0,0663	0,0635	0,0607
6	0,1191	0,1160	0,1128	0,1097	0,1066	0,1034	0,1003	0,0972	0,0941	0,0911
7	0,1378	0,1358	0,1338	0,1317	0,1294	0,1271	0,1247	0,1222	0,1197	0,1171
8	0,1395	0,1392	0,1388	0,1382	0,1375	0,1366	0,1356	0,1344	0,1332	0,1318
9	0,1256	0,1269	0,1280	0,1290	0,1299	0,1306	0,1311	0,1315	0,1317	0,1318
10	0,1017	0,1040	0,1063	0,1084	0,1104	0,1123	0,1140	0,1157	0,1172	0,1186
11	0,0749	0,0776	0,0802	0,0828	0,0853	0,0878	0,0902	0,0925	0,0948	0,0970
12	0,0505	0,0530	0,0555	0,0579	0,0604	0,0629	0,0654	0,0679	0,0703	0,0728
13	0,0315	0,0334	0,0354	0,0374	0,0395	0,0416	0,0438	0,0459	0,0481	0,0504
14	0,0182	0,0196	0,0210	0,0225	0,0240	0,0256	0,0272	0,0289	0,0306	0,0324
15	0,0098	0,0107	0,0116	0,0126	0,0136	0,0147	0,0158	0,0169	0,0182	0,0194
16	0,0050	0,0055	0,0060	0,0066	0,0072	0,0079	0,0086	0,0093	0,0101	0,0109
17	0,0024	0,0026	0,0029	0,0033	0,0036	0,0040	0,0044	0,0048	0,0053	0,0058
18	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019	0,0021	0,0024	0,0026	0,0029
19	0,0005	0,0005	0,0006	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014
20	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004	0,0005	0,0005	0,0006
21	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003
22	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
x	λ									
x	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10,0
0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0000
1	0,0010	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0005	0,0005	0,0005
2	0,0046	0,0043	0,0040	0,0037	0,0034	0,0031	0,0029	0,0027	0,0025	0,0023
3	0,0140	0,0131	0,0123	0,0115	0,0107	0,0100	0,0093	0,0087	0,0081	0,0076
4	0,0319	0,0302	0,0285	0,0269	0,0254	0,0240	0,0226	0,0213	0,0201	0,0189
5	0,0581	0,0555	0,0530	0,0506	0,0483	0,0460	0,0439	0,0418	0,0398	0,0378
6	0,0881	0,0851	0,0822	0,0793	0,0764	0,0736	0,0709	0,0682	0,0656	0,0631
7	0,1145	0,1118	0,1091	0,1064	0,1037	0,1010	0,0982	0,0955	0,0928	0,0901
8	0,1302	0,1286	0,1269	0,1251	0,1232	0,1212	0,1191	0,1170	0,1148	0,1126
9	0,1317	0,1315	0,1311	0,1306	0,1300	0,1293	0,1284	0,1274	0,1263	0,1251
10	0,1198	0,1210	0,1219	0,1228	0,1235	0,1241	0,1245	0,1249	0,1250	0,1251
11	0,0991	0,1012	0,1031	0,1049	0,1067	0,1083	0,1098	0,1112	0,1125	0,1137
12	0,0752	0,0776	0,0799	0,0822	0,0844	0,0866	0,0888	0,0908	0,0928	0,0948
13	0,0526	0,0549	0,0572	0,0594	0,0617	0,0640	0,0662	0,0685	0,0707	0,0729
14	0,0342	0,0361	0,0380	0,0399	0,0419	0,0439	0,0459	0,0479	0,0500	0,0521
15	0,0208	0,0221	0,0235	0,0250	0,0265	0,0281	0,0297	0,0313	0,0330	0,0347
16	0,0118	0,0127	0,0137	0,0147	0,0157	0,0168	0,0180	0,0192	0,0204	0,0217

λ	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10,0
x	11	12	13	14	15	16	17	18	19	20
17	0,0063	0,0069	0,0075	0,0081	0,0088	0,0095	0,0103	0,0111	0,0119	0,0128
18	0,0032	0,0035	0,0039	0,0042	0,0046	0,0051	0,0055	0,0060	0,0065	0,0071
19	0,0015	0,0017	0,0019	0,0021	0,0023	0,0026	0,0028	0,0031	0,0034	0,0037
20	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019
21	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
22	0,0001	0,0001	0,0002	0,0002	0,0002	0,0002	0,0003	0,0003	0,0004	0,0004
23	0,0000	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
24	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001
λ	11	12	13	14	15	16	17	18	19	20
x	0	1	2	3	4	5	6	7	8	9
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0010	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3	0,0037	0,0018	0,0008	0,0004	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000
4	0,0102	0,0053	0,0027	0,0013	0,0006	0,0003	0,0001	0,0001	0,0001	0,0000
5	0,0224	0,0127	0,0070	0,0037	0,0019	0,0010	0,0005	0,0002	0,0001	0,0001
6	0,0411	0,0255	0,0152	0,0087	0,0048	0,0026	0,0014	0,0007	0,0004	0,0002
7	0,0646	0,0437	0,0281	0,0174	0,0104	0,0060	0,0034	0,0019	0,0010	0,0005
8	0,0888	0,0655	0,0457	0,0304	0,0194	0,0120	0,0072	0,0042	0,0024	0,0013
9	0,1085	0,0874	0,0661	0,0473	0,0324	0,0213	0,0135	0,0083	0,0050	0,0029
10	0,1194	0,1048	0,0859	0,0663	0,0486	0,0341	0,0230	0,0150	0,0095	0,0058
11	0,1194	0,1144	0,1015	0,0844	0,0663	0,0496	0,0355	0,0245	0,0164	0,0106
12	0,1094	0,1144	0,1099	0,0984	0,0829	0,0661	0,0504	0,0368	0,0259	0,0176
13	0,0926	0,1056	0,1099	0,1060	0,0956	0,0814	0,0658	0,0509	0,0378	0,0271
14	0,0728	0,0905	0,1021	0,1060	0,1024	0,0930	0,0800	0,0655	0,0514	0,0387
15	0,0534	0,0724	0,0885	0,0989	0,1024	0,0992	0,0906	0,0786	0,0650	0,0516
16	0,0367	0,0543	0,0719	0,0866	0,0960	0,0992	0,0963	0,0884	0,0772	0,0646
17	0,0237	0,0383	0,0550	0,0713	0,0847	0,0934	0,0963	0,0936	0,0863	0,0760
18	0,0145	0,0255	0,0397	0,0554	0,0706	0,0830	0,0909	0,0936	0,0911	0,0844
19	0,0084	0,0161	0,0272	0,0409	0,0557	0,0699	0,0814	0,0887	0,0911	0,0888
20	0,0046	0,0097	0,0177	0,0286	0,0418	0,0559	0,0692	0,0798	0,0866	0,0888
21	0,0024	0,0055	0,0109	0,0191	0,0299	0,0426	0,0560	0,0684	0,0783	0,0846
22	0,0012	0,0030	0,0065	0,0121	0,0204	0,0310	0,0433	0,0560	0,0676	0,0769
23	0,0006	0,0016	0,0037	0,0074	0,0133	0,0216	0,0320	0,0438	0,0559	0,0669
24	0,0003	0,0008	0,0020	0,0043	0,0083	0,0144	0,0226	0,0328	0,0442	0,0557
25	0,0001	0,0004	0,0010	0,0024	0,0050	0,0092	0,0154	0,0237	0,0336	0,0446
26	0,0000	0,0002	0,0005	0,0013	0,0029	0,0057	0,0101	0,0164	0,0246	0,0343
27	0,0000	0,0001	0,0002	0,0007	0,0016	0,0034	0,0063	0,0109	0,0173	0,0254
28	0,0000	0,0000	0,0001	0,0003	0,0009	0,0019	0,0038	0,0070	0,0117	0,0181
29	0,0000	0,0000	0,0001	0,0002	0,0004	0,0011	0,0023	0,0044	0,0077	0,0125
30	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0013	0,0026	0,0049	0,0083
31	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0007	0,0015	0,0030	0,0054
32	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0004	0,0009	0,0018	0,0034
33	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0010	0,0020
34	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0012
35	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0007
36	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004
37	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
38	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
39	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

ANNEXE 3

Table de la loi normale centrée réduite



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007

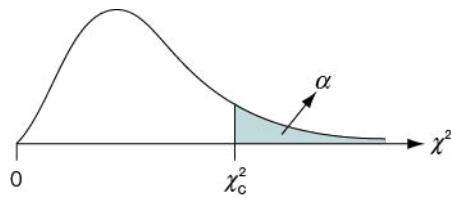
ANNEXE 4

Table de la loi de Student

<i>dl</i>	$\alpha = P(T > t_{\alpha; dl})$				
	0,10	0,05	0,025	0,01	0,005
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
35	1,306	1,690	2,030	2,438	2,724
40	1,303	1,684	2,021	2,423	2,704
50	1,299	1,676	2,009	2,403	2,678
60	1,296	1,671	2,000	2,390	2,660
100	1,290	1,660	1,984	2,364	2,626
120	1,289	1,658	1,980	2,358	2,617
∞	1,282	1,645	1,960	2,326	2,576

ANNEXE 5

Table de la loi du khi-deux



df	α			
	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

RÉPONSES AUX EXERCICES

Chapitre 1

Exercices de compréhension 1.1

1. a) Type : qualitative nominale. Échelle nominale.
b) Type : quantitative discrète. Échelle d'intervalle.
c) Type : quantitative continue. Échelle de rapport.
d) Type : qualitative ordinale. Échelle ordinale.
e) Type : quantitative discrète. Échelle ordinale.
2. a) Faux b) Vrai c) Faux d) Faux

Exercices 1.1

1. i) a) Population : L'ensemble des citoyens de la ville de Québec.
b) Échantillon : Les 200 citoyens soumis au sondage.
c) Unité statistique : Un citoyen.
d) Variable : La chaîne de télévision favorite.
e) Catégories : {Radio-Canada, TVA, Télé-Québec, V, RDI, RDS, VRACK, autres}.
f) Type de variable : Variable qualitative nominale.
- ii) a) Population : Toutes les années comprises entre 2000 et 2010.
b) Échantillon : Il n'y a pas d'échantillon prélevé, l'étude portant sur chacune de ces années.
c) Unité statistique : Une année.
d) Variable : Le taux de chômage.
e) Valeurs : Des pourcentages entre 0 % et 100 % théoriquement, mais, dans la réalité, on peut s'attendre à ce qu'ils soient inférieurs à 15 %.
f) Type de variable : Variable quantitative continue.
- iii) a) Population : Tous les ménages du quartier étudié.
b) Échantillon : Les 380 ménages du quartier soumis à l'enquête.
c) Unité statistique : Un ménage.
d) Variable : Le nombre d'enfants par ménage.
e) Valeurs : Un nombre entier entre 0 et disons 15, les familles de plus de 15 enfants étant bien sûr assez rares de nos jours.
f) Type de variable : Variable quantitative discrète.
- iv) a) Population : L'ensemble de tous les habitants du Québec en 2011.
b) Échantillon : Il n'y a pas d'échantillon, puisque c'est un recensement.
c) Unité statistique : Une personne habitant au Québec.
d) Variable : La langue maternelle.
e) Catégories : Le français seulement, l'anglais seulement, ni le français ni l'anglais, plusieurs langues maternelles.
f) Type de variable : Variable qualitative nominale.

2. a) Quantitative continue. d) Quantitative continue.
b) Qualitative nominale. e) Qualitative nominale.
c) Quantitative discrète. f) Qualitative ordinale.

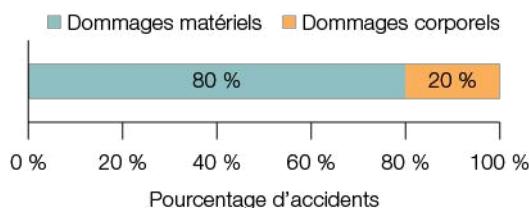
3. a) Qualitative nominale ; échelle nominale.
b) Quantitative discrète ; échelle ordinale.
c) Quantitative discrète ; échelle de rapport.
d) Quantitative continue ; échelle ordinale.
e) Quantitative continue ; échelle de rapport.
f) Qualitative ordinale ; échelle ordinale.
g) Quantitative discrète ; échelle d'intervalle.

4. Heure du lever du soleil : quantitative continue avec une échelle d'intervalle. En effet, 0 h est fixé par convention et ne signifie pas qu'il y a absence de temps ; on peut comparer les heures en mesurant l'écart entre les valeurs (10 h représente 2 h de plus que 8 h), mais on ne peut établir un rapport (10 h/8 h n'a pas de sens).

Vitesse des vents : quantitative continue avec une échelle de rapport, car toutes les opérations mathématiques sont possibles avec cette échelle.

Exercice de compréhension 1.2

- a) Tableau 1 : Répartition des accidents selon le nombre de victimes.
Tableau 2 : Répartition des victimes selon la gravité des blessures.
- b) i) Faux. 20 % des accidents ont fait au moins une victime (ou 16 % ont fait une seule victime).
ii) Faux. 40 % des victimes ont eu des blessures graves ou mortelles.
iii) Faux. On doit construire un diagramme en bâtons.
iv) Vrai
- c) Titre : Répartition des accidents selon la nature des dommages.



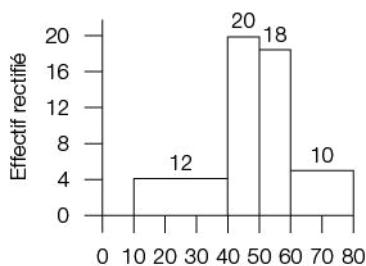
Exercice de compréhension 1.3

1. Nombre de classes ≈ 7
2. $E = 937 - 252 = 685 \$$
3. – Amplitude calculée $= \frac{685}{7} = 97,86 \$$
– Amplitude choisie : 100 \$
4. Première classe : $250 \$ \leq X < 350 \$$

Exercices de compréhension 1.4

1.

Amplitude	Classe	Effectif	Effectif rectifié
30	[10; 40[12	4
10	[40; 50[20	20
10	[50; 60[18	18
20	[60; 80[10	5
	Total	60	



2.

Classe	Pourcentage
[10; 20[14,3 % (1/7)
[20; 30[42,9 % (3/7)
[30; 50[28,6 % (2/7)
[50; 70[14,3 % (1/7)
Total	100,0 %

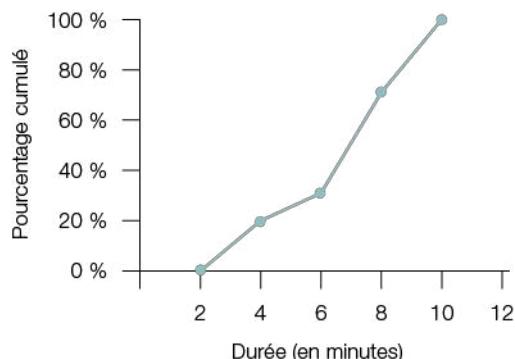
3. $\frac{\text{Aire de la portion en bleu}}{\text{Aire totale du polygone}} \times 100 \% = \frac{2}{8} \times 100 \% = 25 \%$

Exercice de compréhension 1.5

Répartition des appels téléphoniques de l'échantillon selon la durée

Durée (en min)	Pourcentage des appels	Pourcentage cumulé
[2; 4[20 %	20 %
[4; 6[10 %	30 %
[6; 8[40 %	70 %
[8; 10[30 %	100 %
Total	100 %	

Répartition cumulative des appels téléphoniques de l'échantillon selon la durée



Exercices 1.2

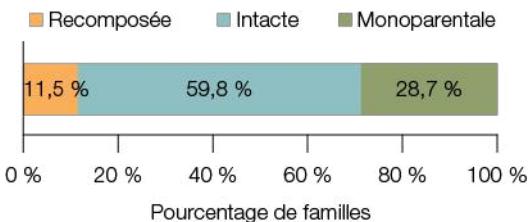
1. a) iii); d'après le titre, les unités statistiques sont les mariages et les décès, et non les Québécois.
 b) En se basant sur les données de 2011, on peut dire que les saisons ont une influence sur les mariages au Québec : il y a beaucoup plus de mariages en été (51 % des mariages) et au printemps (23 %) qu'en hiver (9 %). Par contre, les saisons n'ont pas d'influence sur les décès : bien que l'on ne trouve pas exactement 25 % des décès chaque saison, les écarts sont trop faibles pour être significatifs.

2. a) **Répartition des familles avec enfants à la maison selon le type, Québec, 2011**

Type de famille	Nombre	Pourcentage
Recomposée	146 144	11,5 %
Intacte	761 581	59,8 %
Monoparentale	365 515	28,7 %
Total	1 273 240	100,0 %

Note: On peut se permettre de ne pas répéter la source des données sous un graphique ou un tableau quand celle-ci a déjà été citée dans l'énoncé du problème.

- b) **Répartition des familles avec enfants à la maison selon le type, Québec, 2011**



- c) i) Faux. Il faut dire : 11,5 % des familles avec enfants à la maison...
 ii) Vrai

3. a) **Répartition des longs métrages produits selon les marchés, Québec, 2012**

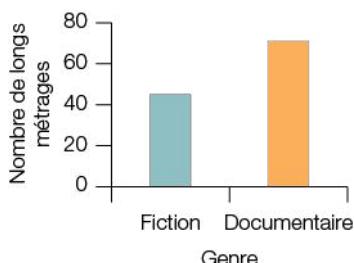
Marché	Nombre	Pourcentage
Cinéma	30	25,9 %
Télévision	37	31,9 %
Autres	49	42,2 %
Total	116	100,0 %

Analyse des données

En 2012, près du tiers des longs métrages québécois ciblaient le marché de la télévision et le quart ciblaient celui du cinéma.

- b) Puisque l'on demande d'utiliser les effectifs, il faut construire un diagramme à rectangles (verticaux ou horizontaux), car les diagrammes circulaires ou linéaires requièrent l'utilisation des fréquences relatives.

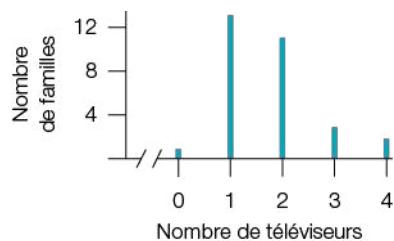
Répartition des longs métrages produits selon le genre, Québec, 2012



4. a) Nombre de téléviseurs ; cinq valeurs : 0, 1, 2, 3, 4.
 b) Quantitative discrète.
 c) Total = 30 données.
 d) Série statistique ordonnée :

0 1 1 1 1 1 1 1 1 1 1
 1 1 2 2 2 2 2 2 2 2 2
 2 3 3 3 4 4

Répartition des familles selon leur nombre de téléviseurs



5. a) Femmes : 54,3 % ; Hommes : 45,7 %.
 Moins de 60 000 \$: 22,3 % ; 60 000 \$ et plus : 77,7 %.

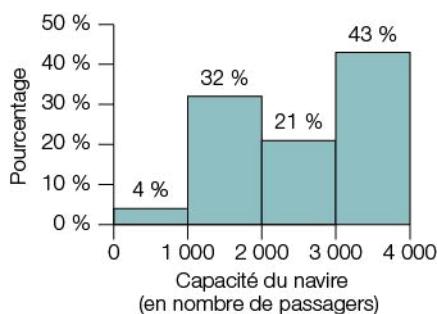
Répartition des croisiéristes de l'échantillon selon l'âge, Québec, 2013

Âge (en ans)	Nombre	Pourcentage
Moins de 45	124	5,3 %
45 ≤ $X < 55$	254	10,9 %
55 ≤ $X < 65$	692	29,7 %
65 et plus	1 260	54,1 %
Total	2 330	100,0 %

Analyse des données

Plus de la moitié des croisiéristes (54 %) ont 65 ans et plus et seulement 5 % ont moins de 45 ans.

Répartition des croisiéristes selon la capacité du navire, Québec, 2013

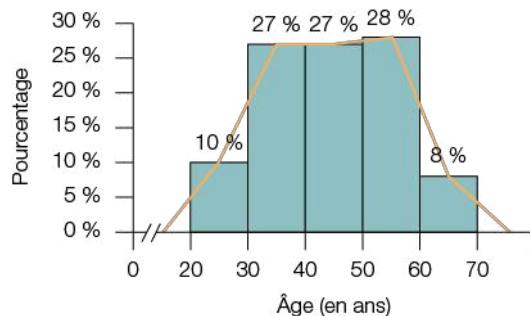


- ii) Faux. Le titre indique que les unités statistiques étudiées sont des croisiéristes, et non des navires de croisière. Il faut donc dire que 64 % des croisiéristes voyagent sur des navires de 2 000 passagers et plus.

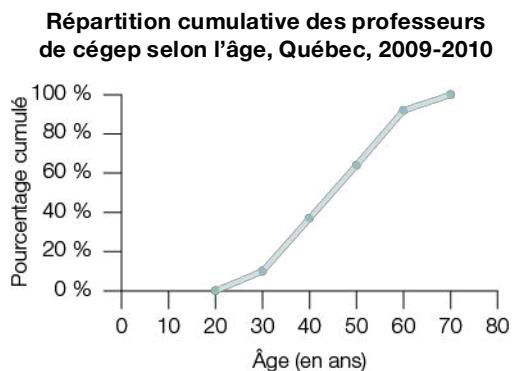
- d) i) – Le titre est erroné. Il faudrait écrire : Répartition des croisiéristes de l'échantillon ayant déjà visité le Québec selon le nombre de visites au Québec avant la croisière, Québec, 2013.
 – On n'a pas fait le bon graphique. Pour une variable quantitative discrète, il faut tracer un diagramme en bâtons, pas un diagramme à rectangles verticaux.
 ii) 350 croisiéristes ($15\% \times 2\,330$)
 iii) 20 % des croisiéristes de l'échantillon ayant déjà visité le Québec avant la croisière y sont venus 3 fois.

6. a) 1. La table de Sturges suggère approximativement 8 classes.
 $E = X_{\max} - X_{\min} = 11,6 - 0,1 = 11,5$
 3. – Amplitude calculée = $11,5/8 \approx 1,44$
 – Amplitude choisie = 1,5, car il est plus agréable de travailler avec des multiples de 5 (ou, dans ce cas-ci, des multiples de 0,5).
 4. Première classe : $0 \leq X < 1,5$
 b) 1. La table de Sturges suggère approximativement 7 classes.
 $E = X_{\max} - X_{\min} = 206 - 142 = 64$
 3. – Amplitude calculée = $64/7 \approx 9,14$
 – Amplitude choisie = 10 : cela facilitera grandement la lecture du tableau.
 4. Première classe : $140 \leq X < 150$
7. a) En 2009-2010, 37 % des professeurs de cégep ont moins de 40 ans, 27 % ont entre 40 et 50 ans et 36 % ont plus de 50 ans. Comme il est à prévoir que la plupart de ces derniers prendront leur retraite dans les 10 prochaines années, il faudra donc être en mesure de remplacer près de 36 % du corps professoral d'ici 2020.

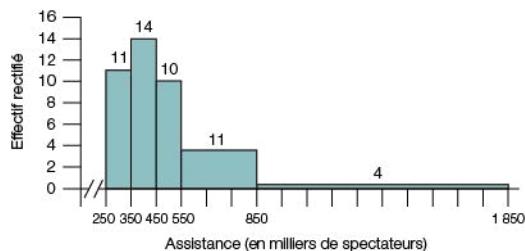
Répartition des professeurs de cégep selon l'âge, Québec, 2009-2010



- c) Points de l'ogive : (20 ; 0 %), (30 ; 10 %),
(40 ; 37 %), (50 ; 64 %), (60 ; 92 %),
(70 ; 100 %)



8. a) **Répartition des 50 films les plus populaires selon l'assistance, Québec, 2009 à 2011**

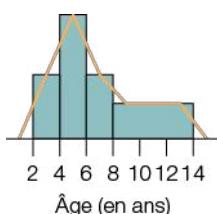


- b) Diagramme linéaire : 8.
Analyse des données : 4; 8; 30.

9. a) **Répartition des enfants selon l'âge**

Âge (en ans)	Pourcentage d'enfants
$2 \leq X < 4$	18,2 %
$4 \leq X < 6$	36,4 %
$6 \leq X < 8$	18,2 %
$8 \leq X < 14$	27,3 %
Total	100,1 %

- b) **Répartition des enfants selon l'âge**



10. La surface sous le polygone de fréquences située à droite de 35 ans occupe une portion plus grande de la surface totale sous le polygone de l'année 2011 que sous celui de l'année 1991.

11. a) 1. La table de Sturges suggère de construire approximativement 7 classes avec 48 données.
2. $E = X_{\max} - X_{\min} = 78,6 - 23,1 = 55,5 \%$
3. – Amplitude calculée = $55,5/7 \approx 7,9 \%$
– Amplitude choisie = 10 %
4. Première classe : $20 \% \leq X < 30 \%$

- b) **Répartition des 48 collèges publics selon le pourcentage d'étudiants en formation technique, Québec, 2010**

Pourcentage d'étudiants en formation technique	Nombre de collèges	Pourcentage de collèges
$20 \% \leq X < 30 \%$	1	2,1 %
$30 \% \leq X < 40 \%$	1	2,1 %
$40 \% \leq X < 50 \%$	7	14,6 %
$50 \% \leq X < 60 \%$	15	31,3 %
$60 \% \leq X < 70 \%$	17	35,4 %
$70 \% \leq X < 80 \%$	7	14,6 %
Total	48	100,1 %

Source : Ministère de l'Éducation, du Loisir et du Sport. Réseaux, DSID Portail informationnel, système Socrate, données au 25 décembre 2012.

- c) *Analyse des données*

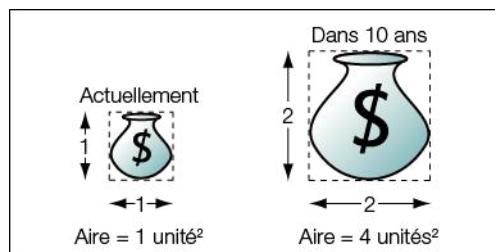
En 2010, 19 % des collèges comptent moins de 50 % de leurs étudiants en formation technique, 31 % en comptent de 50 à 60 % et la moitié des collèges ont de 60 à 80 % de leur clientèle en formation technique.

12. a) Non, la surface de l'icône représentant les frais de scolarité dans 10 ans est environ quatre fois plus grande que la surface de celle qui représente les frais de scolarité actuellement (voir le graphique 1 ci-dessous). Comme le pourcentage double, la surface de l'icône doit doubler.

Une bonne représentation de la situation est donnée par le graphique 2.

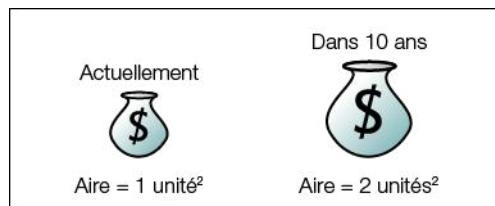
Graphique 1

Mauvaise représentation



Graphique 2

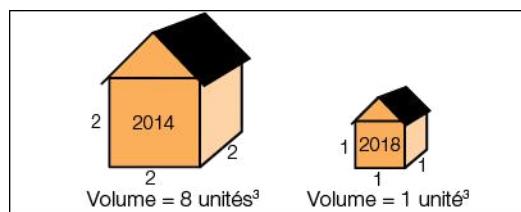
Bonne représentation



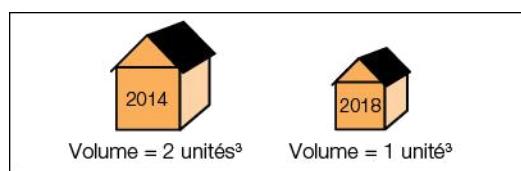
- b) Le graphique 2. Ici, c'est le volume qui doit être proportionnel à l'effectif. Puisqu'on estime que le nombre de maisons neuves en 2018 sera égal à la moitié de ce qu'il était en 2014, le volume de la maison de 2018 doit être égal à la moitié du volume

de celle de 2014; or, il est bien évident que pour le graphique 1, la maison de 2014 peut contenir plus de deux fois la maison de 2018.

Graphique 1
Mauvaise représentation



Graphique 2
Bonne représentation



Exercice de compréhension 1.6

Centre de classe: 2,5 ; 7,5 ; 15 ; 25 ; 35.

a) Près de 50 % (49,5 %).

$$\text{b) Moyenne } (\mu) = 2,5 \times 23,4 \% + 7,5 \times 27,1 \% + 15 \times 25,4 \% + 25 \times 12,1 \% + 35 \times 12,0 \% \\ = 13,7 \text{ heures par semaine}$$

Exercice de compréhension 1.7

En ce qui concerne l'expérience, les danseurs ne se divisent pas en trois groupes égaux. En conséquence, on doit pondérer les revenus moyens par la proportion de danseurs que l'on retrouve à chacun des trois niveaux d'expérience.

Revenu moyen

$$= 18\ 514 \$ \times \left(\frac{124}{650} \right) + 25\ 323 \$ \times \left(\frac{290}{650} \right) + 35\ 809 \$ \times \left(\frac{236}{650} \right) \\ = 18\ 514 \$ \times 0,191 + 25\ 323 \$ \times 0,446 + 35\ 809 \$ \times 0,363 \\ = 27\ 829,90 \$$$

Le revenu moyen d'un danseur professionnel était de 27 830 \$ au Québec en 2010, soit 6 170 \$ de moins que le revenu moyen des Québécois.

Exercices de compréhension 1.8

1. Faux. Au moins 50 % des étudiants ont une note de 68 ou moins.

$$2. \text{ a) Médiane} = \frac{2+3}{2} = 2,5 \text{ jours}$$

Interprétation

50 % des employés qui ont pris un congé de maladie dans la dernière année se sont absents 2 jours ou moins (ou 3 jours ou plus).

b) Médiane = 2 repas

Interprétation

Au moins 50 % des répondants ont pris 2 repas ou moins à la cafétéria au cours de la dernière semaine.

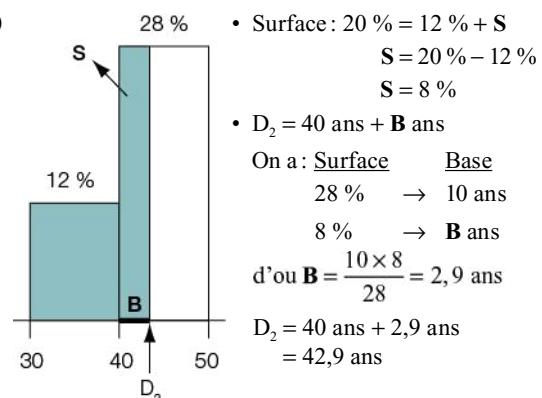
3. Mode : Anglais

Interprétation

En 2012, une majorité (70 %) des 50 albums les plus vendus au Québec sont enregistrés en anglais.

Exercices de compréhension 1.9

1. a)



b) Faux

2. Au 2^e quintile (V_2).

3. 80^e (C_{80})

Exercices 1.3

1. b) et c).

2. a) La moyenne se situe entre 1 et 2, mais plus près de 2 ; $\mu \approx 1,8$.

b) La moyenne se situe entre 8 et 10 ; $\mu \approx 9$.

c) $\mu = 0$

3. a) Moyenne = 11,7 calendriers par personne. Ces 7 personnes ont vendu en moyenne 11,7 calendriers en une journée.

Mode = 6 calendriers. Une pluralité de personnes (28,6 %) ont vendu 6 calendriers en une journée. Médiane = 8 calendriers. Au moins 50 % des personnes ont vendu 8 calendriers ou moins en une journée.

La médiane est la meilleure mesure de tendance centrale pour représenter cette série, car la moyenne est grandement influencée par une donnée beaucoup plus grande que les autres et l'effectif du mode (2) ne se démarque pas assez pour être significatif.

b) Moyenne = 751 spectateurs. Il y a eu en moyenne 751 spectateurs par représentation.

Mode : aucun.

Médiane = 757 spectateurs. La moitié des représentations ont été jouées devant moins de 757 spectateurs.

La moyenne et la médiane sont toutes deux acceptables pour représenter ces données, chacune comportant une interprétation intéressante des données.

4. a) Le mode est la seule mesure de tendance centrale possible.
 Mode : Un livre tous les deux ans.
Interprétation
 Une majorité d'écrivains professionnels (52 %) publient un livre tous les deux ans.

b) Moyenne pondérée

$$= 10 \times \frac{680}{1\,510} + 11,8 \times \frac{830}{1\,510} = 11 \text{ livres}$$

c) • $\mu = 2,2$ genres. Les écrivains professionnels ont publié dans 2,2 genres littéraires en moyenne.
 • $Mo = 1$ genre. Une pluralité d'écrivains professionnels (34 %) ont publié dans un seul genre littéraire.
 • Médiane = 2 genres. Au moins 50 % des écrivains professionnels ont publié dans 2 genres littéraires ou moins.

5. a) Seulement 63 des 1 510 auteurs professionnels.
 b) Classe modale : moins de 5 000 \$. Une majorité d'écrivains professionnels (65 %) ont tiré moins de 5 000 \$ de leur création littéraire.
 c) Médiane = 3 852 \$. On peut estimer que 50 % des écrivains professionnels ont tiré moins de 3 852 \$ de leur création littéraire.
 d) $\mu = 9\,525 \$$. Les écrivains professionnels ont tiré en moyenne 9 525 \$ de leur création littéraire. L'écart entre la moyenne (9 525 \$) et la médiane (3 852 \$) est attribuable au fait que certains auteurs ont un revenu nettement plus élevé que les autres.
 e) La classe modale, car le pourcentage d'auteurs dans cette classe de revenu se démarque nettement de celui des autres classes.

6. a) 35,9 %
 b) Moyenne = 4 854 \$. En 2011-2012, le montant moyen accordé aux cégepiens de la formation technique par le Programme de prêts et bourses est de 4 854 \$.
 c) Classe modale : Moins de 2 000 \$. Une pluralité de bénéficiaires de la formation technique (26 %) ont reçu moins de 2 000 \$ du Programme de prêts et bourses en 2011-2012.
 d) Médiane $\approx 4\,407 \$$. On peut estimer que 50 % des bénéficiaires de la formation technique ont reçu moins de 4 407 \$ du Programme de prêts et bourses en 2011-2012.
 e) $Q_1 \approx 1\,901 \$$. On peut estimer que 25 % des bénéficiaires de la formation technique ont reçu moins de 1 901 \$ du Programme de prêts et bourses en 2011-2012.

7. a) Médiane $\approx 29,6$ ans. On peut estimer que 50 % des immigrants accueillis au Québec en 2012 ont moins de 29,6 ans, ce qui est bien inférieur à l'âge médian de 41,5 ans de la population du Québec.
 b) Classe modale : [30 ans ; 45 ans]. Une pluralité (39 %) d'immigrants accueillis par le Québec en 2012 ont entre 30 et 45 ans.
 c) Moyenne = 28,2 ans. On peut estimer que les immigrants accueillis au Québec en 2012 ont en moyenne 28,2 ans.

- d) $V_2 \approx 24,5$ ans. On peut estimer que 40 % des immigrants accueillis par le Québec en 2012 ont moins de 24,5 ans.

e) Au 9^e décile (D_9). On peut estimer que 90 % des immigrants accueillis au Québec en 2012 ont moins de 45 ans.

Moyenne pondérée

$$= 60 \times \left(\frac{3}{10} \right) + 70 \times \left(\frac{2}{10} \right) + 65 \times \left(\frac{2}{10} \right) + 80 \times \left(\frac{3}{10} \right) = 69$$

1. a) Quantitative discrète.
 b) Les ménages québécois se divisent en trois groupes : le premier tiers des ménages compte 1 seule personne, le deuxième tiers en compte 2 et le dernier tiers compte 3 personnes ou plus.
 c) $\mu = 2,3$ personnes. En moyenne, on compte 2,3 personnes par ménage au Québec en 2011.
 d) $Me = 2$ personnes. Au moins 50 % des ménages québécois comptent 2 personnes ou moins en 2011.

2. a) Oui : $\mu = 66,3 \times 0,50 + 60,2 \times 0,50 = 63,3$ kg.
 b) Non, car on ne connaît pas le poids de chacun des sous-groupes dans le nouveau groupe.
 c) Oui : $\mu = 66,3 \times \frac{10}{50} + 60,2 \times \frac{40}{50} = 61,4$ kg

3. a) 30,1 ; 1,9.
 b) 30,2 ; 2,1.
 c) $D_1 \approx 22,7$ ans. On peut estimer que 10 % des femmes qui ont donné naissance à un enfant en 2011 avaient moins de 22,7 ans.
 d) On peut estimer que 80 % des femmes qui ont donné naissance à un enfant en 2011 avaient moins de 34,6 ans. On peut en déduire que 20 % des femmes avaient plus de 34,6 ans, ce qui est intéressant à souligner considérant qu'en 1991 ce pourcentage était de 8 %.

4. a) $\mu \approx 31,5$ min. En 2010, les travailleurs de la région de Montréal mettent en moyenne 31,5 minutes pour se rendre au travail. C'est 12,5 minutes de plus que les travailleurs des villes de moins de 250 000 habitants.
 b) 27 %. (Les travailleurs qui prennent plus de 45 minutes pour se rendre au travail prendront plus de 90 minutes pour l'aller-retour entre la maison et le travail.)
 c) $Me \approx 31,7$ min. On peut estimer que 50 % des travailleurs de la région de Montréal mettent moins de 31,7 minutes pour se rendre au travail en 2010.
 d) $D_1 \approx 7,5$ min. On peut estimer que seulement 10 % des travailleurs de la région de Montréal mettent moins de 7,5 minutes pour se rendre au travail en 2010.
 e) $Q_1 \approx 17,8$ min. On peut estimer que 25 % des travailleurs de la région de Montréal mettent moins de 17,8 minutes pour se rendre au travail en 2010.

Exercices de compréhension 1.10

1. a) 1 b) 2 c) 1
2. a) A b) B

- c) Faux. La plupart des données d'une distribution dont la moyenne est 70 et dont l'écart type est 10 sont comprises entre 60 et 80.

d) $\sigma = 0$

3. $\bar{x} = 26$ ans et $\sigma = 6,6$ ans

4. a) $\sigma^2 = 21,3$ et $\sigma = 4,6$ b) $s^2 = 24$ et $s = 4,9$

Exercices de compréhension 1.11

1. $\bar{x} = 27,4$; $\sigma = 3,8$; $s = 4,4$.

2. $\bar{x} = 8,3$; $\sigma = 4,3$; $s = 4,4$.

3. $\bar{x} = 5,1$; $\sigma = 1,5$; s = aucune valeur ou erreur.

Exercices 1.4

1. a) Histogramme 1: $\mu \approx 11$; histogramme 2: $\mu \approx 7$.

b) Histogramme 1.

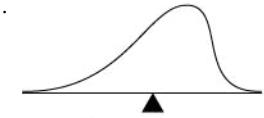
c) 33,3 % : surface du rectangle (3 unités) \div surface de l'histogramme (9 unités).

2. a) Nombre de buts; quantitative discrète.

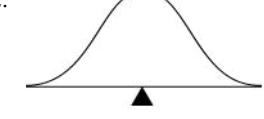
b) Mode = 1 but et $M_e = 2$ buts. En saison régulière 2013-2014, les Canadiens de Montréal ont marqué un but pour une pluralité (27 %) des matchs disputés, et pour au moins 50 % d'entre eux, ils ont marqué 2 buts ou moins.

c) $\mu = 2,5$ buts et $\sigma = 1,7$ but. En 2013-2014, les Canadiens de Montréal ont marqué en moyenne 2,5 buts par match en saison régulière. Au cours de la plupart des matchs, ils ont marqué de 1 à 4 buts (entiers compris dans l'intervalle [0,8 ; 4,2]).

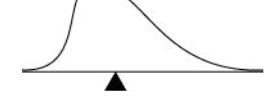
3. a) 1.



2.



3.

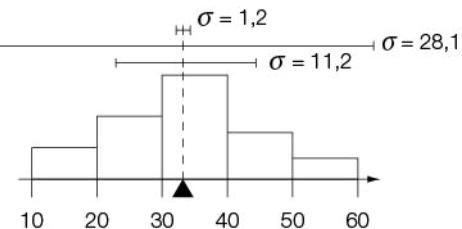


b) L'affirmation ii) est vraie.

4. a) La série A.

b) Il y a des Québécois qui ont un revenu personnel nettement plus élevé que les autres, ce qui fait augmenter la moyenne et la rend moins représentative des données. Pour les revenus des contribuables, la médiane est toujours la meilleure mesure de tendance centrale à utiliser.

c)

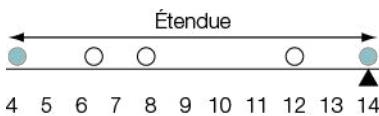


- 1,2 est trop petit pour représenter la moyenne des écarts. De plus, il n'y a pas environ les deux tiers de la surface de l'histogramme dans l'intervalle $[\mu - \sigma; \mu + \sigma]$;

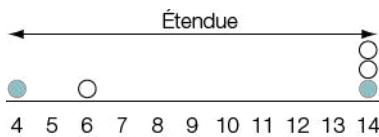
- 28,1 est trop grand pour représenter la moyenne des écarts : il est plus grand que le plus grand des écarts;

- l'écart type de cette distribution est 11,2. Il est plus plausible que cet écart corresponde à la moyenne des écarts. De plus, environ les deux tiers de la surface de l'histogramme semblent compris dans l'intervalle $[\mu - \sigma; \mu + \sigma]$.

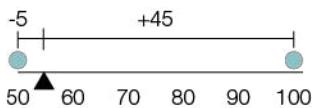
5. a) La moyenne μ ne peut pas être égale à 14. En effet, le plus grand résultat, étant donné l'étendue, est 14, et les trois autres sont compris entre 4 et 14 inclusivement; donc 14 ne peut pas être le centre d'équilibre de la série.



b) Il est plausible que la médiane soit égale à 14. Exemple : 4, 6, 14, 14, 14.



c) La moyenne ne peut pas être égale à 55. En effet, pour que le centre d'équilibre de la série soit 55, il faudrait ajouter trois données à gauche de la moyenne, de telle façon que la somme des écarts situés à gauche de la moyenne soit égale, mais de signe négatif, à la somme des écarts situés à droite de la moyenne, soit 45, ce qui est mathématiquement impossible. Même en choisissant les trois plus petites valeurs possibles, on obtient la série : 50, 50, 50, 50 et 100 avec une moyenne de 60.



6. a) Étendue $E = 88$.

b) $\bar{x} = 324$ et $s = 34$. De 2006 à 2011, on a une moyenne annuelle de 324 accidents avec dommages corporels impliquant une motoneige au Québec. Pour la plupart des années considérées, on dénombre entre 290 et 358 accidents de ce type par année.

c) $M_e = 319,5$ (il ne faut pas oublier de mettre les données en ordre). Pour 50 % des années considérées, le nombre annuel d'accidents avec dommages corporels impliquant une motoneige est de 319 ou moins (ou est inférieur à 319,5).

7. a) $\mu = 44,7$ ans et $\sigma = 11,3$ ans. En 2009-2010, les professeurs de cégep ont en moyenne 44,7 ans. La plupart d'entre eux ont entre 33,4 ans et 56,0 ans.

- b) Non, car le coefficient de variation (25,2 %) est supérieur à 15 %.
- c) $M_e = 44,8$ ans. On peut estimer que 50 % des professeurs de cégep ont moins de 44,8 ans en 2009-2010.
8. a) Écart type de 2 < écart type de 3 < écart type de 1
b) Écart type de 2 < écart type de 1 < écart type de 3
9. a) Dans le graphique 2, les courbes ont la même moyenne.
b) Dans le graphique 1, les courbes ont le même écart type.
10. a) La série B. Dans la série A, il y a des données qui sont deux fois, et même cinq fois, plus grandes que les autres, ce qui est loin d'être le cas dans la série B.
b) Le coefficient de variation de la série A est de 46,7 % ($\mu = 6$ et $\sigma = 2,8$), alors que celui de la série B est de 2,6 % ($\mu = 106$ et $\sigma = 2,8$) ; cette série est donc effectivement beaucoup plus homogène que l'autre.
11. a) [45 ans; 55 ans]. En 2011, une pluralité de propriétaires d'une motocyclette (37 %) ont entre 45 et 55 ans au Québec.
b) $\mu \approx 47,5$ ans et $\sigma \approx 10,5$ ans. En 2011, l'âge moyen des propriétaires d'une motocyclette est de 47,5 ans au Québec. La plupart d'entre eux ont entre 37 et 58 ans.
c) Approximatives, car les données sont groupées en classes.
d) 48,8 ; 31,7.

Exercices de compréhension 1.12

1. a) Cote z de A = -3 ; cote z de B = 0 ; cote z de C = 1,5.
b) Cote z de A = -1 ; cote z de B = 0 ; cote z de C = 0,5.
2. a) Faux b) Faux c) Vrai d) Vrai
3. Très bonne.
4. a)
-
- b) Écart entre A et $\mu = 2 \times 10 = 20$
Écart entre B et $\mu = -1 \times 10 = -10$
Écart entre C et $\mu = -1,5 \times 10 = -15$
- c) $A = 50 + 20 = 70$
 $B = 50 - 10 = 40$
 $C = 50 - 15 = 35$

5. Cote z de Mia = $\frac{85 - 52}{13} = 2,5$

Cote z de Thomas = $\frac{25 - 12}{6} = 2,2$

Cote z de Lucie = $\frac{75 - 47}{10} = 2,8$

Lucie devrait être nommée meilleure vendeuse du mois.

Exercices 1.5

1. a) Usine A : $\bar{x} = 89,6$ kg/cm² et $s = 7,6$ kg/cm².
Usine B : $\bar{x} = 98$ kg/cm² et $s = 12,2$ kg/cm².

- b) Le contrat sera accordé à l'usine A. La production de l'usine A, pour laquelle $CV = 8,5\%$, est plus homogène que celle de l'usine B, pour laquelle $CV = 12,4\%$.

2. a) i)

3. a) 1. Cote $z = 2$.

2. Cote $z = 0$.

- b) 1. 20 points.

2. 0 point.

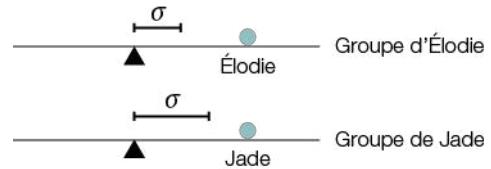
- c) 1. 85 points.

2. 65 points.

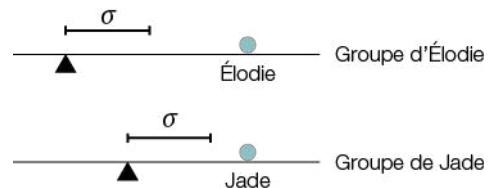
3. 55 points.

4. 50 points.

4. a) Le groupe auquel Élodie appartient. La distance entre la moyenne et la note étant la même dans les deux cas, pour que la cote z d'Élodie soit plus grande, il faut que l'écart type de son groupe soit compris un plus grand nombre de fois dans cette distance que l'écart type du groupe de Jade ; par conséquent, l'écart type du groupe d'Élodie doit être plus petit.



- b) Le groupe auquel Élodie appartient. Étant donné qu'Élodie a une cote z plus élevée que Jade, sa note est nécessairement plus éloignée de la moyenne du groupe que ne l'est la note de Jade. Comme les deux étudiantes ont la même note, la moyenne du groupe d'Élodie est nécessairement plus faible.

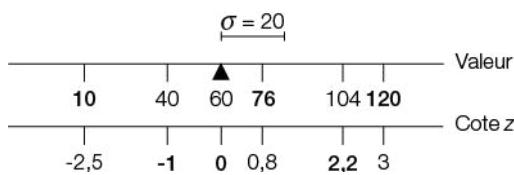


5. a) $CV_Q = 9,5\%$ et $CV_F = 14,1\%$; les salaires des enseignants québécois sont plus homogènes que ceux des enseignants français.

- b) Moyenne pondérée = 73,6 points

6. Au plus 16 000 personnes : celles dont la cote z est inférieure à -2,5 et celles dont la cote z est supérieure à 2,5.

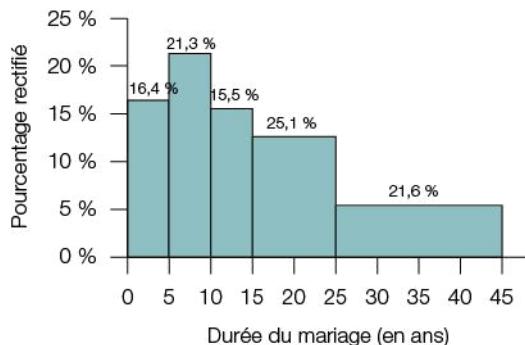
- 1.
7. a) i) Gestion de production.
ii) Production artistique.
iii) Conception de jeux.
- b) 33 ; 41.
8. Le commerçant est justifié d'attribuer la baisse de ses recettes aux travaux, car une cote z de -5 est exceptionnelle.
9. a) La plupart des étudiants ont une note qui se situe à ± 10 points de la moyenne, soit entre 50 et 70 points.
- b) La note de Lucie se situe à 1,5 écart type au-dessus de la moyenne.
- c) Lucie a 15 points de plus que la moyenne:
 $1,5 \times 10 = 15$.
- d) Lucie a obtenu une note de 75 points:
 $60 + 15 = 75$ points.
- 10.



Exercices récapitulatifs

1. a) i) $\mu = 35,5$ ans et $\sigma = 10,5$ ans. On peut estimer que les propriétaires d'une PME québécoise avaient en moyenne 35,5 ans lorsqu'ils ont démarré leur entreprise; la plupart d'entre eux avaient entre 25 et 46 ans.
ii) La distribution n'est pas homogène, car le $CV = 29,6\% > 15\%$.
iii) Cote $z = 2,1$. Par rapport à la moyenne d'âge des entrepreneurs au moment où ils lancent leur entreprise, l'âge de cette personne est plus élevé de 2,1 écarts types, ce qui est assez rare: très peu d'entrepreneurs sont aussi âgés lorsqu'ils lancent leur entreprise.
- b) i) On peut estimer que 50 % des nouveaux propriétaires d'une PME ont investi moins de 59 453 \$ pour lancer leur entreprise.
ii) La majorité des nouveaux propriétaires d'une PME (59 %) n'ont créé aucun emploi, à part le leur, lorsqu'ils ont lancé leur entreprise.
c) i) 32 ; 20. ii) 60 ; 35.
2. a) i) Faux. Selon le titre du tableau, le total correspond au nombre de divorces en 2008, et non au nombre de mariages. Il faut donc écrire: «Dans 37,7 % des divorces survenus en 2008, le couple était marié depuis moins de 10 ans.»
ii) Vrai
b) Les pourcentages de la 4^e et de la 5^e classe doivent être rectifiés afin de respecter le principe de proportionnalité. L'amplitude de base choisie est de cinq ans.

Répartition des divorces selon la durée du mariage, Québec, 2008



- c) $Q_1 = 7$ ans. On peut estimer que, dans 25 % des divorces survenus en 2008, le couple était marié depuis moins de 7 ans.
3. a) Q1 : Variable qualitative nominale et échelle nominale.
Q2 : Variable quantitative discrète et échelle de rapport.
Q3 : Variable qualitative ordinaire et échelle ordinale.
- b) Port du Québec : 42 %; port hors Québec : 58 %.
- c) i) Un diagramme en bâtons.
ii) $Me = 1$ nuit. Pour au moins 50 % des répondants qui ont séjourné au Québec avant de monter à bord du navire, la durée du séjour a été d'une nuit.
iii) $\bar{x} = 1,7$ nuit et $s = 0,9$ nuit. La durée moyenne du séjour des répondants qui ont séjourné au Québec avant de monter à bord du navire a été de 1,7 nuit. Pour la plupart, le séjour a duré une ou deux nuits.
4. a) 1. La table de Sturges suggère approximativement 10 classes.
2. $E = X_{\max} - X_{\min} = 984 - 54 = 930$
3. – Amplitude calculée = $930/10 = 93$
– Amplitude choisie = 100
4. Première classe : $50 \leq X < 150$
b) Moyenne pondérée = 7 793 \$, au dollar près.
5. a) Parce que le principe de proportionnalité entre volume et effectif n'est pas respecté. En effet, visuellement, on a l'impression que le cylindre du *Grand Journal* peut contenir au moins trois fois celui du *Petit Journal*, ce qui nous amène à conclure qu'il y a trois fois plus de détenteurs de placements qui lisent *Le Grand Journal* que *Le Petit Journal*. Or, ce n'est pas du tout le cas: les effectifs indiquent que ce rapport est plutôt de 1,1 fois plus ($151\ 300 \div 141\ 000$). En fait, la hauteur du cylindre pour *Le Petit Journal* devrait correspondre à 93 % ($141\ 000 \div 151\ 300$) de la hauteur du cylindre du *Grand Journal*.

b) Catégorie «Total»

À 61 % ($124\ 900 \div 205\ 000$) de la hauteur du cylindre du *Grand Journal*. Avec la bonne hauteur pour le cylindre du *Petit Journal*, on aurait visuellement beaucoup plus avantage *Le Grand Journal* qu'avec ce qui est présenté dans le dépliant actuel.

Catégorie «Hommes»

À 58 % de la hauteur du cylindre du *Grand Journal*. Là encore, le cylindre du *Petit Journal* est beaucoup trop haut, ce qui désavantage visuellement le *Grand Journal*.

Catégorie «Femmes»

À 65 % de la hauteur du cylindre du *Grand Journal*.

Chapitre 2

Exercice de compréhension 2.1

	V	V'	Total
A	48 %	40 %	88 %
A'	7 %	5 %	12 %
Total	55 %	45 %	100 %

a) $P(A' \cap V') = 5 \%$

b) $P(A' \cup V') = 12 \% + 45 \% - 5 \% = 52 \%$

Exercices 2.1

1. a) $S = \{0, 1, 2, 3, \dots, 15\}$

b) $S = [0 \text{ min}; 5 \text{ min}]$

c) $S = \{0, 1, 2, \dots, 20\}$

d) $S = \{\text{janvier, février, mars, ..., décembre}\}$

2. a) $S = \{\text{PPP, PPF, PFP, PFF, FPP, FPF, FFP, FFF}\}$

b) i) $A = \{\text{PPF, PFP, FPP}\}$ et $P(A) = 3/8$

ii) $B = \{\text{FFF}\}$ et $P(B) = 1/8$

iii) $C = \{\text{PPF, PFP, FPP, PPP}\}$ et $P(C) = 4/8$

iv) $D = \emptyset$ et $P(D) = 0$

v) $E = S$ et $P(E) = 1$

3. a) $D_1 \cap R_1$: «La première carte pigée est une dame rouge»

b) $R_1 \cap R_2$: «Les deux cartes pigées sont rouges»

c) $D_1 \cup D_2$: «La première ou la deuxième carte pigée est une dame» ou «Au moins une des deux cartes pigées est une dame»

d) $R_1 \cap R'_2$: «La première carte pigée est rouge et la deuxième est noire»

4. a) $P(A) = 6/10$

d) $P(A' \cap B') = 3/10$

b) $P(A \cap B) = 2/10$

e) $P(A \cap B') = 4/10$

c) $P(A \cup B) = 7/10$

5. M: «le nouveau-né pèse moins de 2 500 g»

G: «le nouveau-né est un garçon»

a) $P(M) = 5,7 \%$

d) $P(G \cup M') = 54,1 \%$

b) $P(G' \cap M) = 3,0 \%$

e) Non, $P(G) = 51,1 \%$

c) $P(G' \cap M') = 45,9 \%$

6. I: «le ménage est branché à Internet»

E: «le ménage a des enfants»

a) $P(I') = 18,5 \%$

b) $P(E') = 75,2 \%$

c) $P(I \cap E') = 58,1 \%$

d) $P(I' \cup E) = 41,9 \%$

7. a) $S = \{\text{l'ensemble des 365 jours de l'année}\}$.

Soit J: «La personne est née en janvier»; alors, $P(J) = 31/365 = 8,5 \%$.

b) On suppose *a priori* que les éléments de S sont équiprobables.

c) $P(J) = 7\ 100/88\ 700 = 8,0 \%$. L'écart de 0,5 point de pourcentage par rapport au résultat obtenu en a) s'explique par le fait que les éléments de S ne sont pas équiprobables. Il y a plus de naissances en juillet, août, septembre et octobre qu'en décembre, janvier et février. De nos jours, les méthodes contraceptives permettent aux gens qui le désirent de planifier la naissance de leurs enfants.

d) $P(\text{Décéder en mars}) = 8,9 \%$

$P(\text{Décéder en septembre}) = 7,6 \%$

Statistiquement, il y a un peu plus de risques de décéder en mars.

e) 46,8 %

8. a) $P(H) = 27,4 \%$

b) $P(-15\ 000) = 24,6 \%$

c) $P(F \cap -15\ 000) = 20,5 \%$

d) $P(H \cap -15\ 000) = 4,2 \%$

e) $P(H \cup 15\ 000 \text{ ou plus}) = 79,5 \%$

Exercice de compréhension 2.2

	Moins de 5 000 \$ (A)	5 000 \$ et plus (A')	Total
Moins de 45 ans (B)	422	288	710
45 ans et plus (B')	552	238	790
Total	974	526	1 500

a) $P(A' \cap B') = \frac{238}{790} = 30,1 \%$

b) $P(A \cap B) = \frac{422}{1\ 500} = 28,1 \%$

c) i) $P(A) = \frac{974}{1\ 500} = 64,9 \%$

ii) $P(A | B) = \frac{422}{710} = 59,4 \%$

Non. La probabilité est plus basse chez les écrivains de moins de 45 ans.

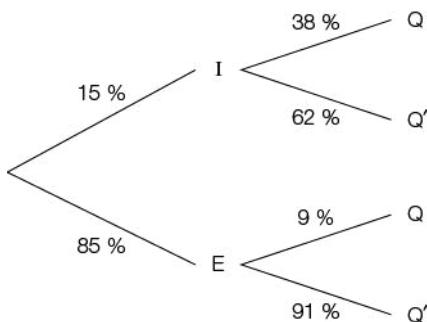
Exercice de compréhension 2.3

a) i) Faux ii) Vrai iii) Vrai

b) 29 ; 4 ; 3.

c) Plus le revenu du ménage est élevé, plus la probabilité de consommer de l'eau embouteillée à la maison augmente.

Exercice de compréhension 2.4



- $P(Q) = P(I \cap Q) + P(E \cap Q)$
 $= P(I) P(Q | I) + P(E) P(Q | E)$
 $= (15 \% \times 38 \%) + (85 \% \times 9 \%)$
 $= 5,7 \% + 7,65 \%$
 $= 13,4 \%$
- $P(Q' | E) = 91 \%$
- $P(E \cap Q') = P(E) P(Q' | E) = 85 \% \times 91 \% = 77,4 \%$
- $P(I | Q) = \frac{P(I \cap Q)}{P(Q)} = \frac{5,7 \%}{13,4 \%} = 42,5 \%$
- $P(Q) = 13,4 \% \neq P(Q | I) = 38 \%.$ Comme $P(Q) \neq P(Q | I)$, les événements Q et I sont dépendants.

Globalement, il y a 13 % de chances qu'un travailleur consacre plus de 40 heures par semaine à son travail. Or, si ce travailleur est un travailleur indépendant, cette probabilité augmente à 38 %.

Exercices 2.2

- a) i) $2/3$ iii) $4/7$ v) $6/7$
ii) $2/6$ iv) $1/4$ vi) $2/7$
b) i) Non, car $A \cap B \neq \emptyset$.
ii) Non. On a $P(A) = 60 \%$ et $P(A | B) = 66,7 \%$, donc $P(A) \neq P(A | B)$.

- B : détenir un baccalauréat
M : détenir une maîtrise
D : détenir un doctorat
a) $P(F) = 59,5 \%$ d) $P(H | D \cup M) = 47,2 \%$
b) $P(H \cap M) = 10,6 \%$ e) $P(D) \neq P(D | H) \neq P(D | F)$
c) $P(B | F) = 76,6 \%$
Donc, la probabilité de détenir un doctorat dépend du sexe du diplômé.

Interprétation

Globalement, il y a 3,5 % de chances que le diplôme obtenu soit un doctorat. Or, si l'on tient compte du sexe, cette probabilité augmente à 4,7 % si le diplômé est un homme et elle diminue à 2,8 % si c'est une femme. Il y a plus de chances qu'un homme diplômé détienne un doctorat.

- a) 1/3 (Seuls les nombres 3 et 6 sont divisibles par 3.)
b) Les événements A et B sont indépendants puisque $P(A) = P(A | B)$.

Interprétation

Le fait de savoir que le nombre obtenu est pair ne modifie pas la probabilité d'obtenir un nombre divisible par 3 ; elle est toujours de 1/3.

- a) $P(V | F) = 4/12 = 33,3 \%$
b) 100 %
c) Non. Les événements V et F sont dépendants, car $P(V) \neq (V | F)$; on a aussi $P(F) \neq P(F | V)$.

Interprétation

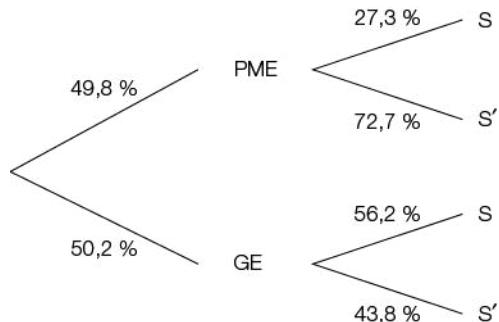
Les chances d'obtenir un valet, soit 7,7 %, augmentent à 33,3 % si la carte pigée est une figure.

- a) $P(H_c \cap F_c) = 65,7 \%$
b) $P(H_c) = 74 \%$
c) $P(H_c \cup F_c) = 85 \%$
d) $P(F_v | H_v) = 29,2 \%$
e) $P(H_v | F_v) = 33,3 \%$
f) $P(H_d) \neq P(H_d | F_c) \neq P(H_d | F_v) \neq P(H_d | F_d)$. Donc, la probabilité que le marié soit divorcé dépend de l'état matrimonial de la mariée.

Interprétation

Si l'on ne tient pas compte de l'état matrimonial de la mariée, la probabilité que le marié soit divorcé est de 23,6 %. Or, si l'on tient compte de l'état matrimonial de la mariée, cette probabilité diminue à 13,6 % si la mariée est célibataire alors qu'elle augmente à 47,6 % si elle est veuve, et à 57,5 % si elle est divorcée. Un divorcé a donc beaucoup plus de chances d'épouser une femme divorcée qu'une femme célibataire.

- a)



- $P(S' | PME) = 72,7 \%$
- $P(S' \cap GE) = P(GE \cap S') = 22 \%$
- $P(S) = P(PME \cap S) + P(GE \cap S) = 41,8 \%$
- La probabilité diminue de 14,5 points de pourcentage, car $P(S) = 41,8 \%$ alors que $P(S | PME) = 27,3 \%$. Les événements S et PME sont dépendants.
- $P(S') = 100 \% - P(S) = 58,2 \%$
- $P(GE | S) = \frac{P(GE \cap S)}{P(S)} = 67,5 \%$

- a) $P(R) = P(A \cap R) + P(B \cap R) + P(C \cap R) = 4,3 \%$
b) $P(C | R) = \frac{P(C \cap R)}{P(R)} = 27,9 \%$

8. P : musique sur support physique
 N : musique sur support numérique
 F : Musique en français
 a) $P(P \cap F) = 28\%$
 b) $P(F') = 61,8\%$
 c) $P(N | F) = \frac{P(N \cap F)}{P(F)} = 26,7\%$

9. Soit les événements :

W : «possède un site Web»

A : «moins de 20 employés»

B : «de 20 à 99 employés»

C : «de 100 à 499 employés»

a) i) $P(A \cap W) = 32,3\%$

ii) $P(A | W') = \frac{P(A \cap W')}{P(W')} = \frac{18,7\%}{30,2\%} = 61,9\%$

b) Oui, car on a $P(W) \neq P(W | A) \neq P(W | B) \neq P(W | C)$.

Interprétation

Globalement, la probabilité qu'une PME ait un site Web est de 70 %. Or, si l'on tient compte de la taille de l'entreprise, cette probabilité diminue à 63 % s'il y a moins de 20 employés et, à l'inverse, elle augmente à 81 % et à 97 % s'il y a de 20 à 99 employés et de 100 à 499 employés, respectivement. Une tendance se dégage : plus l'entreprise compte d'employés, plus la probabilité qu'elle ait un site Web est élevée.

c) i) $P(W_1 \cap W_2 \cap W_3) = 69,8\% \times 69,8\% \times 69,8\% = 34\%$

ii) $P(\text{au moins un } W \text{ en 3 piges}) = P(\text{tous les résultats possibles}) - P(\text{aucun } W \text{ en 3 piges}) = P(S) - P(W'_1 \cap W'_2 \cap W'_3) = 100\% - 2,8\% = 97,2\%$

10. R : la personne est en retard

V : la personne prend sa voiture

T : la personne prend le train

M : la personne prend le métro

a) $P(R) = P(R \cap V) + P(R \cap T) + P(R \cap M) = 3,8\%$

b) $P(V | R) = \frac{P(V \cap R)}{P(R)} = 52,6\%$

11. PME : petite ou moyenne entreprise

GE : grande entreprise

S : secteur des services

B : secteur des biens

a) $P(\text{PME} \cap S) = 35,5\%$ b) $P(\text{GE} \cap B) = 12,5\%$

c) $P(\text{PME} | S) = 47,4\%$ d) $P(\text{GE} | B) = 49,9\%$

e) Non. À 1 % près, $P(S) \approx P(S | \text{PME}) \approx P(S | \text{GE})$.

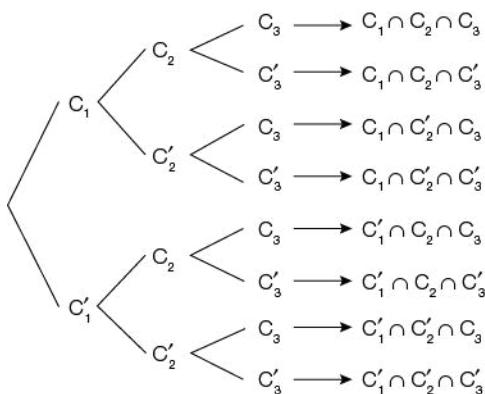
Globalement, la probabilité qu'une entreprise soit dans le secteur des services est de 75 %. Or, cette probabilité est presque la même, à 1 point de pourcentage près¹,

1. Nous présenterons au chapitre 6 une règle permettant de décider à partir de quelle valeur un écart peut être jugé suffisamment grand pour conclure à la dépendance de deux variables. Pour le moment, on peut considérer qu'un écart de 1 point de pourcentage n'est pas assez grand.

que l'entreprise soit une PME (74 %) ou une grande entreprise (76 %).

f) i) $P(\text{PME}_1 \cap \text{PME}_2) = 48\% \times 48\% = 23\%$
 ii) $P(\text{au moins une PME}) = P(\text{tous les cas possibles}) - P(\text{aucune PME}) = P(S) - P(\text{PME}'_1 \cap \text{PME}'_2) = 100\% - (52\% \times 52\%) = 73\%$

12. a) 1^{er} tirage 2^e tirage 3^e tirage S



b) L'événement A : «piger trois cartes de cœur» se produit seulement si l'événement C se produit à chacun des trois tirages. Donc,

$$\begin{aligned} P(A) &= P(C_1 \cap C_2 \cap C_3) \\ &= P(C_1) P(C_2 | C_1) P(C_3 | C_1 \cap C_2) \\ &= \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} = 1,3\% \end{aligned}$$

c) L'événement B : «ne piger aucun cœur» se produit seulement si l'événement C' se produit à chacun des trois tirages.

$$\begin{aligned} P(B) &= P(C'_1 \cap C'_2 \cap C'_3) \\ &= P(C'_1) P(C'_2 | C'_1) P(C'_3 | C'_1 \cap C'_2) \\ &= \frac{39}{52} \times \frac{38}{51} \times \frac{37}{50} = 41,4\% \end{aligned}$$

d) L'événement contraire de B est B' : «piger au moins un cœur».

$$P(B') = P(S) - P(B) = 100\% - 41,4\% = 58,6\%$$

Exercices de compréhension 2.5

1. a) $4! = 4 \times 3 \times 2 \times 1 = 24$
 b) $4 \times 3 = 12$

2. $26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17\ 576\ 000$

3. Soit A : «piger 2 ouvriers et 1 technicien».

$$\begin{aligned} P(A) &= \frac{n(A)}{n(S)} = \frac{\binom{6}{2} \binom{4}{1}}{\binom{10}{3}} = \frac{\frac{6!}{2! 4!} \times \frac{4!}{1! 3!}}{\frac{10!}{3! 7!}} = \frac{15 \times 4}{120} \\ &= \frac{60}{120} = 50\% \end{aligned}$$

Exercices 2.3

1. a) 5 040 b) 10 302 c) 15 d) 70

2. 72 jours

3. a) $10^4 = 10\ 000$

4. a) $9 \times 8 \times 7 = 504$
b) $5 \times 8 \times 7 = 280$

5. a) $26^2 \times 10^3 = 676\ 000$

6. a) $\binom{30}{3} = \frac{30!}{3! 27!} = 4\ 060$

b) $\binom{10}{3} = 120$

7. a) $\binom{20}{3} = 1\ 140$

b) $\binom{6}{3} = 20; \quad \frac{20}{1\ 140} = 1,8\%$

c) $\binom{10}{2} \binom{6}{1} = 270; \quad \frac{270}{1\ 140} = 23,7\%$

d) $\binom{10}{2} \binom{10}{1} = 450; \quad \frac{450}{1\ 140} = 39,5\%$

e) $\frac{240}{1\ 140} = 21,1\%$

8. $\frac{\binom{26}{3} \binom{24}{3}}{\binom{50}{6}} = \frac{2\ 600 \times 2\ 024}{15\ 890\ 700} = 33,1\%$

9. a) $9 \times 10^5 = 900\ 000$

b) 1/900 000

c) Le premier chiffre ne peut pas être 0 et il ne peut pas non plus être 1, car dans ce cas vous auriez gagné le gros lot : il reste donc 8 choix possibles pour ce premier chiffre. Les autres chiffres doivent être identiques à ceux du billet, d'où on tire $8 \times 1 \times 1 \times 1 \times 1 = 8$. Les chances de gagner sont de 8 sur 900 000.

d) Le premier chiffre ne peut pas être 0, donc il y a 9 choix possibles pour ce premier chiffre. Le deuxième chiffre ne peut pas être 2, sinon vous auriez gagné 5 000 \$: il reste donc 9 choix possibles pour ce deuxième chiffre. Les autres chiffres doivent être identiques à ceux du billet, d'où on tire $9 \times 9 \times 1 \times 1 \times 1 = 81$. Par exemple, 135 488, 245 488, 335 488, etc.

Les chances de gagner sont de 81 sur 900 000.

Exercices récapitulatifs

1. G: «gagner le match»

E: «jouer à l'extérieur»

D: «jouer à domicile»

	G	G'	Total
E	23	18	41
D	23	18	41
Total	46	36	82

a) $P(D) = 41/82 = 50\%$

b) $P(D | G') = 18/36 = 50\%$

c) $P(G | D) = 23/41 = 56,1\%$

d) $P(G \cap D) = 23/82 = 28\%$

e) $P(G' \cap E') = 18/82 = 22\%$

f) $P(G \cup E) = (46 + 41 - 23)/82 = 78\%$

g) Non, car $P(G) = P(G | E) = P(G | D)$.

Globalement, la probabilité que l'équipe gagne le match est de 56,1 %. Or, cette probabilité est la même que la partie soit disputée à l'extérieur (56,1 %) ou à domicile (56,1 %).

2. V: «prélever une voyelle»

C: «prélever une consonne»

a) i) $P(V_1 \cap V_2 \cap V_3) = (6/26)(5/25)(4/24) = 0,8\%$

ii) $P(V_1 \cap V_2 \cap C_3) = (6/26)(5/25)(20/24) = 3,8\%$

iii) $P(\text{au moins une voyelle})$

= $P(\text{tous les cas}) - P(\text{ne piger aucune voyelle})$

= $P(S) - P(C_1 \cap C_2 \cap C_3)$

= $100\% - (20/26)(19/25)(18/24)$

= $100\% - 43,8\%$

= $56,2\%$

b) i) $P(V_1 \cap V_2 \cap V_3) = (6/26)(6/26)(6/26) = 1,2\%$

ii) $P(V_1 \cap V_2 \cap C_3) = (6/26)(6/26)(20/26) = 4,1\%$

c) $P(V_4 \cap V_5 \cap C_6) = (4/23)(3/22)(19/21) = 2,1\%$

d) Pour ce type de pige, l'ordre des lettres n'a pas d'importance.

Soit A: «piger 2 voyelles et 1 consonne»

$$P(A) = \frac{\binom{6}{2} \binom{20}{1}}{\binom{26}{3}} = \frac{\frac{6!}{2! 4!} \times \frac{20!}{1! 19!}}{\frac{26!}{3! 23!}} = \frac{15 \times 20}{2\ 600} = 11,5\%$$

3. S: «n'avoir aucun problème de santé de longue durée»

M: «avoir un médecin de famille»

	M	M'	Total
S	35,3 %	16,4 %	51,7 %
S'	43,4 %	4,9 %	48,3 %
Total	78,7 %	21,3 %	100,0 %

a) $P(M') = 21,3\%$

b) $P(S \cap M') = 16,4\%$

c) $P(M | S') = P(M \cap S')/P(S') = 43,4\% / 48,3\% = 89,9\%$

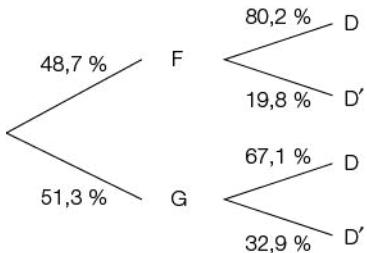
d) $P(M' | S) = P(M' \cap S)/P(S) = 16,4\% / 51,7\% = 31,7\%$

e) Oui, car on a $P(M) \neq P(M | A) \neq P(M | B) \neq P(M | C) \neq P(M | D) \neq P(M | E)$

Globalement, il y a 79 % de chances qu'un Québécois de 15 ans et plus ait un médecin de famille. Or, si l'on tient compte de l'âge, cette probabilité est plus basse chez les plus jeunes, soit 69 % et 71 % pour les deux premières classes d'âge et, à l'inverse, elle est plus élevée chez les plus vieux, soit 86 %, 93 % et 96 % pour les trois dernières classes d'âge.

Une tendance se dégage : plus un Québécois est âgé, plus il y a de chances qu'il ait un médecin de famille.

4. Soit D : «obtenir un diplôme d'études secondaires avant l'âge de 20 ans».



- a) $P(D) = P(D \cap H) + P(D \cap F) = 73,5\%$
 b) i) $P(G | D) = 46,8\%$
 ii) $P(F | D) = 53,1\%$
 c) i) $P(D | F) = P(D) = 73,5\%$
 ii) $P(D | G) = P(D) = 73,5\%$
5. a) 28 %
 b) $(89\% \times 23\%) + (10\% \times 24\%) + (1\% \times 28\%) = 23,2\%$
 c) $\frac{10\% \times 24\%}{23,2\%} = 10,3\%$

d)

Situation matrimoniale			
Nombre de mariages	Encore mariés	Ne sont plus mariés	Total
Mariés une fois	68,5 %	20,5 %	89 %
Mariés deux fois	7,6 %	2,4 %	10 %
Mariés trois fois	0,7 %	0,3 %	1 %
Total	76,8 %	23,2 %	100 %

e)

Situation matrimoniale			
Nombre de mariages	Encore mariés	Ne sont plus mariés	Total
Mariés une fois	77 %	23 %	100 %
Mariés deux fois	76 %	24 %	100 %
Mariés trois fois	72 %	28 %	100 %
Total	76,8 %	23,2 %	100 %

6. a) $\binom{16}{5} = \frac{16!}{5! 11!} = 4\,368$
 b) $\binom{10}{3} \binom{6}{2} + 4\,368 = 41,2\%$
7. a) $\frac{1}{26 \times 25 \times 24 \times 10 \times 9} = 0,000\,07\%$
 b) $\frac{1}{6 \times 5 \times 4 \times 1 \times 9} = 0,09\%$
 c) $\frac{1}{3 \times 2 \times 1 \times 10 \times 9} = 0,19\%$

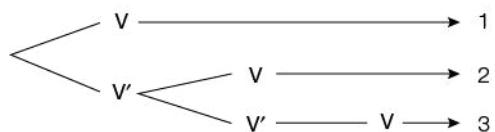
Chapitre 3

Exercice de compréhension 3.1

Distribution de probabilité de X : «nombre de tirages nécessaires pour obtenir une boule verte»

x	1	2	3	Total
f(x)	60 %	30 %	10 %	100 %

1^{er} tirage 2^e tirage 3^e tirage X



$$f(1) = P(X=1) = P(V) = \frac{3}{5} = 60\%$$

$$f(2) = P(X=2) = P(V' \cap V) = \frac{2}{5} \times \frac{3}{4} = 30\%$$

$$f(3) = P(X=3) = P(V' \cap V' \cap V) = \frac{2}{5} \times \frac{1}{4} \times \frac{3}{3} = 10\%$$

Exercice de compréhension 3.2

$$E(X) = 1,5 \text{ tirage}$$

$$\sigma = 0,67 \text{ tirage}$$

Interprétation : 1,5 ; 0,7 ; 1 ; 2.

Exercice de compréhension 3.3

Distribution de probabilité de X : «gain net du joueur»

x	2 \$	-1 \$	-2 \$	Total
f(x)	10 %	30 %	60 %	100 %

$X = \text{gain} - \text{frais de participation}$

$$E(X) = -1,30 \text{ \$}$$

Interprétation : 1,30 \\$; 130 \\$.

Exercices 3.1

1. a) Distribution de probabilité de X : «nombre de femmes piégées»

x	0	1	Total
f(x)	60 %	40 %	100 %

- b) Distribution de probabilité de X : «nombre de femmes piégées»

x	0	1	2	Total
f(x)	30 %	60 %	10 %	100 %

- c) Distribution de probabilité de X : «nombre de femmes piégées»

x	0	1	2	Total
f(x)	36 %	48 %	16 %	100 %

2. Distribution de probabilité de X : «nombre de tirages nécessaires pour obtenir une femme»

x	1	2	3	4	Total
$f(x)$	40 %	30 %	20 %	10 %	100 %

3. a) X : «nombre de jeunes qui ont un téléphone intelligent dans un échantillon de 3 jeunes». Les valeurs de X sont 0, 1, 2, 3.

b) **Distribution de probabilité de X : «nombre de jeunes qui ont un téléphone intelligent dans un échantillon de 3 jeunes»**

x	0	1	2	3	Total
$f(x)$	3,3 %	20,9 %	44,4 %	31,4 %	100,0 %

c) $f(2) = 44,4 \%$

d) $f(3) = 31,4 \%$

e) $E(X) = 2$ jeunes

Interprétation

Si l'on étudie un grand nombre d'échantillons de 3 jeunes, on peut espérer trouver en moyenne 2 jeunes qui ont un téléphone intelligent.

4. $E(X) = 2$ tirages et $\sigma = 1$ tirage. Si on réalise l'expérience aléatoire un très grand nombre de fois, il faudra en moyenne 2 tirages pour obtenir la première femme, l'écart type étant de 1 tirage. Donc, lors de la plupart des expériences, il faut faire entre 1 et 3 tirages avant de pêcher une femme.
5. a) Les chances de vendre au moins 25 abonnements par semaine sont de 80 % pour la femme et de 71 % pour l'homme.
- b) i) $E(X) = 28$ abonnements et $\sigma = 4,5$ abonnements.
Sur un grand nombre de semaines, la femme peut s'attendre à vendre en moyenne 28 abonnements par semaine avec un écart type de 4,5 abonnements. Donc, la plupart du temps, la femme vend entre 24 et 32 abonnements par semaine.
- ii) $E(X) = 27$ abonnements et $\sigma = 5,1$ abonnements.
Sur un grand nombre de semaines, l'homme peut s'attendre à vendre en moyenne 27 abonnements par semaine avec un écart type de 5,1 abonnements. Donc, la plupart du temps, l'homme vend entre 22 et 32 abonnements par semaine.
- c) Sur un grand nombre de semaines :
- i) 368 \$ par semaine (soit $6 \times 28 + 200$).
 - ii) 362 \$ par semaine (soit $6 \times 27 + 200$).
6. a) $S = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$
- b) $X = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$
- c) L'événement antécédent de 7 est $A = \{(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)\}$; donc, $f(7) = P(X = 7) = 6/36$.

d) Distribution de probabilité de X : «total des points obtenus»

x	2	3	4	5	6	7	8	9	10	11	12	Total
$f(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{36}{36}$

e) $E(X) = 252/36 = 7$ points. Si on répète l'expérience aléatoire un grand nombre de fois, on s'attend à ce que le total des points soit en moyenne de 7.

f) $Y = 1, 2, 3, 4, 5, 6$

g) **Distribution de probabilité de Y : «le plus élevé des deux nombres obtenus ou le nombre de points par dé si l'il s'agit d'une paire»**

y	1	2	3	4	5	6	Total
$f(y)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	$\frac{36}{36}$

7. Distribution de probabilité de X : «gain net d'un joueur»

x	2 \$	1 \$	-3 \$	Total
$f(x)$	25 %	25 %	50 %	100 %

$E(X) = -0,75$ \$. Si l'on joue plusieurs parties, on perd en moyenne 0,75 \$ par partie. Par exemple, si l'on joue 100 parties, on perd en moyenne 75 \$.

8. Distribution de probabilité de X : «gain net d'un joueur»

x	3 \$	1 \$	-k \$	Total
$f(x)$	25 %	50 %	25 %	100 %

$E(X) = 0$ \$ si $k = 5$ \$. Si un joueur n'obtient aucune face, il devrait débourser 5 \$.

Exercices 3.2

1. a) X : «nombre d'enfants de rang 2 parmi 10 nouveau-nés» suit une $B(10; 0,36)$.
- b) X : «poids à la naissance» ne suit pas une loi binomiale, car on ne peut pas définir un succès ou un échec à chaque épreuve.
- c) X : «nombre de fumeurs parmi 8 personnes» ne suit pas une loi binomiale, car les épreuves ne sont pas indépendantes. (Le tirage se fait sans remise dans une petite population : $N < 20n$.)
- d) X : «nombre de fumeurs parmi 8 personnes» suit une $B(8; 0,4)$.
- e) X : «nombre de cégépiens parmi 30 qui ont échoué 2 cours ou plus» suit une $B(30; 1/3)$. (Comme la taille de la population est grande par rapport à celle de l'échantillon, un tirage sans remise peut être traité comme un tirage avec remise : statistiquement, la différence entre les deux types de tirages dans le calcul d'une probabilité est négligeable.)
- f) X : «nombre d'enfants dans la famille» ne suit pas une loi binomiale, car on ne peut pas définir un succès ou un échec à chaque épreuve.
- g) X : «nombre de tirages nécessaires pour obtenir un cœur» ne suit pas une loi binomiale, car même si les épreuves sont indépendantes et même si l'on

peut définir, à chaque épreuve, un succès : « piger un cœur », ou un échec : « ne pas piquer un cœur », on ne connaît pas le nombre n d'épreuves de l'expérience aléatoire; ainsi, X ne peut pas correspondre au nombre de succès en n épreuves indépendantes.

2. a) 0, 1, 2, 3, 4, 5

b) $\binom{5}{4} = 5$ façons

c) $A = \{\text{SSSSE, SSSS, SSESS, SESSS, ESSSS}\}$

d) $(0,75)^4(0,25) = 7,9\%$

e) $f(4) = P(X = 4) = P(A) = 5(0,75)^4(0,25) = 39,6\%$

3. a) X suit une $B(4 ; 0,21)$. (On peut considérer que les épreuves sont indépendantes, car dans une grande population ($N > 20n$), un tirage sans remise peut être considéré comme un tirage avec remise.)

Distribution de probabilité de X

x	0	1	2	3	4	Total
$f(x)$	39,0 %	41,4 %	16,5 %	2,9 %	0,2 %	100,0 %

- b) i) $f(2) = 16,5\%$
 ii) $P(X < 3) = 96,9\%$
 iii) $P(X \geq 2) = 19,6\%$
 iv) $P(X \leq 3) = 99,8\%$

4. a) X suit une $B(10 ; 0,12)$; $P(X = 0) = 27,9\%$.
 b) $P(X \geq 1) = 1 - 0,279 = 0,721 = 72,1\%$

5. X suit une $B(5 ; 2/6)$; $P(X = 4) = 4,1\%$.

6. X suit une $B(50 ; 0,245)$;

$$P(X = 8) = \binom{50}{8}(0,245)^8(0,755)^{42} = 5,2\%$$

Exercices 3.3

1. a) X suit une $B(10 ; 0,1)$. $E(X) = 1$ et $\sigma = 0,9$. Sur un grand nombre d'échantillons de 10 clients, une personne en moyenne aura rempli le questionnaire; l'écart type est de 0,9 personne. Donc, dans la plupart des échantillons, il y aura soit 0, 1 ou 2 clients qui auront rempli le questionnaire.
 b) $P(X > 2) = 1 - P(X \leq 2) = 0,0702 \approx 7\%$
2. a) X suit une $B(10 ; 0,3)$; $P(X \geq 5) = 0,1502 \approx 15\%$.
 b) X suit une $B(1\ 000 ; 0,3)$. $E(X) = 300$ et $\sigma = 14,5$. Sur un grand nombre d'échantillons de 1 000 Québécois, on s'attend à trouver en moyenne 300 Québécois par échantillon favorables à l'exploitation du gaz de schiste, l'écart type étant de 14,5 personnes. Donc, la plupart des échantillons devraient comprendre entre 286 et 314 personnes favorables à l'exploitation du gaz de schiste.
 c) i) Oui, car la valeur $X = 250$ correspond à une cote z de -3,45, ce qui est exceptionnel. Avec un tel résultat, on peut affirmer sans grand risque que le pourcentage de Québécois favorables à l'exploitation du gaz de schiste n'est pas de 30% : il est plus faible. Il devrait se situer autour de 25% (soit 250/1 000).
 ii) Pour une valeur de $X = 315$, la cote z est de 1,03. Il n'y a pas lieu de douter de l'hypothèse voulant

que 30 % des Québécois soient favorables à l'exploitation du gaz de schiste.

- d) i) X suit une $B(10 ; 0,3)$. $E(X) = 3$ et $\sigma = 1,4$. On ne peut pas rejeter l'hypothèse, car, pour $X = 2$, la cote z est de -0,7, ce qui n'a rien d'exceptionnel. Selon la table binomiale, la probabilité d'obtenir 2 succès en 10 épreuves est de 23,4 %.
 ii) Pour une $B(1\ 000 ; 0,3)$, on a $E(X) = 300$ et $\sigma = 14,5$; la valeur $X = 200$ correspond à une cote z de -6,9, ce qui conduit à rejeter l'hypothèse voulant que 30 % des Québécois soient favorables à l'exploitation du gaz de schiste.

À noter

En multipliant par 100 la taille de l'échantillon (de 10 à 1 000) et la valeur de X (de 2 à 200), on passe du non-rejet au rejet de l'hypothèse. Un échantillon de grande taille permet toujours une analyse plus précise de la population.

3. a) Succès : la PME cesse ses activités la 1^{re} année.
 On a une $B(6 ; 0,2)$.

- i) $P(X = 2) = 0,2458 = 24,6\%$
 ii) $P(X > 2) = 1 - P(X \leq 2) = 0,0989 = 9,9\%$
 iii) $P(X = 0) = 0,2621 = 26,2\%$

- b) 20 PME

4. a) Succès : le jeune a un baladeur numérique.
 X suit une $B(12 ; 0,35)$.

- i) $f(11) = 0,01\%$
 ii) $f(8) = 0,0199 = 2\%$

- b) $E(X) = 4,2$ et $\sigma = 1,7$. Si l'on prélevait un grand nombre d'échantillons de 12 jeunes de 18 à 24 ans, on compterait en moyenne 4,2 jeunes par échantillon ayant un baladeur numérique, avec un écart type de 1,7. Donc, la plupart des échantillons comprendraient de 3 à 6 jeunes ayant un baladeur numérique.

5. Succès : avoir une bonne réponse.
 X suit une $B(10 ; 0,2)$.

- a) $f(0) = 10,7\%$
 b) $f(10) = 0\%$
 c) $f(6) = 0,0055 = 0,6\%$
 d) $P(X \geq 6) = 0,0064 = 0,6\%$
 e) $P(X < 6) = 99,4\%$

6. $P(X > 3) = 12,1\%$

7. $P(X \geq 9) = 30,4\%$

8. a) Soit Y le nombre d'échecs.
 $P(Y = 2) = 31,2\%$
 b) $P(X \geq 7) = P(Y \leq 1) = 36,7\%$

9. a) Soit Y le nombre d'échecs. Y suit une $B(10 ; 0,30)$.
 $P(X = 8) = P(Y = 2) = 23,4\%$
 b) $P(X < 8) = P(Y > 2) = 61,7\%$

Exercices 3.4

1. a) X suit une $Po(4,5)$; $f(0) = 1,1\%$.
 b) X suit une $Po(1,5)$; $P(X \geq 1) = 1 - 0,2231 = 77,7\%$.
 c) Une panne avec une probabilité de 33,5 %.

2. a) $\sigma = 1,5$ visiteur. La variation du nombre de visiteurs autour de la moyenne est de $\pm 1,5$ personne : la plupart du temps, entre 12 h et 21 h, de 1 à 4 personnes par minute visitent le site Web.

b) X suit une $Po(2,4)$; $P(X > 9) = 0,02\%$.

c) X suit une $Po(2,4)$; $P(X > 3) = 1 - P(X \leq 3) = 22,1\%$.

d) X suit une $Po(1,2)$; $f(0) = 30,1\%$.

3. a) X suit une $B(40 ; 0,01)$; donc, $E(X) = np = 0,4$ et $\sigma = \sqrt{npq} = 0,6$. Si le contrat est respecté, la plupart des échantillons contiendront soit 0 ou 1 pièce défectueuse.

b) i) Oui : pour $X \geq 3$, on a une cote $z \geq 4,3$.
ii) $B(40 ; 0,01) \approx Po(0,4)$, $P(X \geq 3) = 0,8\%$ de chances. Avec une probabilité aussi faible, il n'y a pratiquement aucun risque de se tromper en affirmant que le contrat n'a pas été respecté.

4. X suit une $Po(5)$; $P(5 \leq X \leq 10) = 54,6\%$. (On comprend pourquoi cette île est un paradis pour les chasseurs.)

5. a) X suit une $Po(1)$; $P(X = 0) = 36,8\%$.
b) X suit une $Po(0,5)$; $P(X < 3) = 98,6\%$.

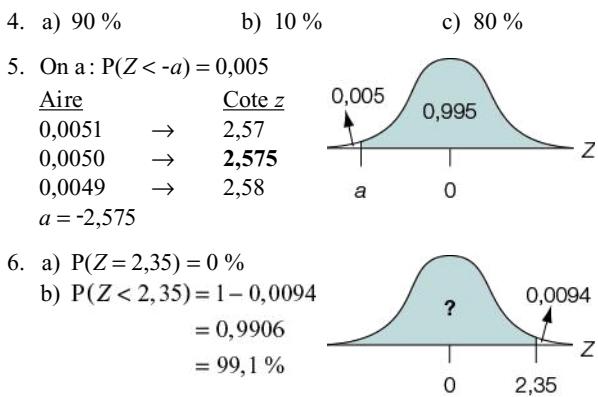
6. $B(100 ; 0,015) \approx Po(1,5)$

a) $P(X = 2) = 25,1\%$
b) $P(X \leq 2) = 80,9\%$
c) $P(X \geq 2) = 44,2\%$

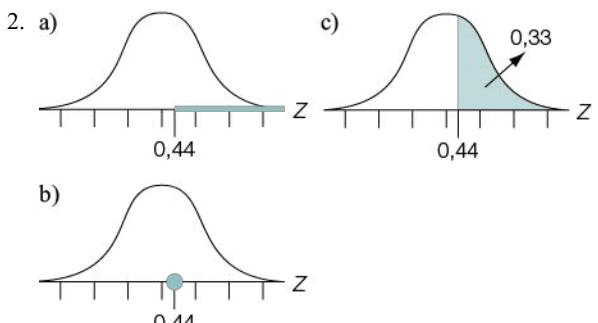
7. $B(150 ; 0,02) \approx Po(3)$

a) $P(X = 3) = 22,4\%$
b) $P(X < 5) = 81,5\%$

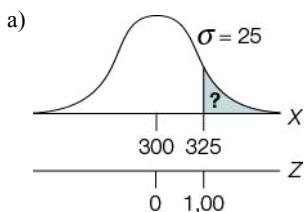
8. $B(120 ; 0,04) \approx Po(4,8)$; $P(X \geq 4) = 70,6\%$



Exercices 3.5

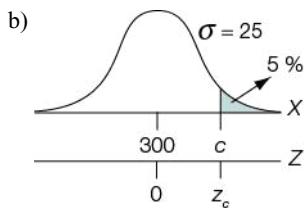


Exercice de compréhension 3.5



$$\begin{aligned} P(X > 325) &= P\left(Z > \frac{325 - 300}{25}\right) \\ &= P(Z > 1,00) \\ &= 0,1587 \\ &= 15,9 \% \end{aligned}$$

Il y a près de 16 % de risques de ne pas pouvoir répondre à la demande.



- On cherche c , telle que $P(X > c) = 5\%$.
 - En cote z , on a $P(Z > z_c) = 0,05$. Selon la table, $z_c = 1,645$.
 - La valeur c se situe à 1,645 écart type à droite de la moyenne.
 $c = 300 + 1,645 \times 25$
 $c = 341$
- On doit commander 341 douzaines d'œufs tous les deux jours.

Exercice de compréhension 3.6

- a) $P(X_B = 5) \approx P(4,5 < X_N < 5,5)$
b) $P(12 < X_B \leq 18) \approx P(12,5 < X_N < 18,5)$

Exercice de compréhension 3.7

X : «nombre de jeunes parmi 100 qui jouent à des jeux en ligne».

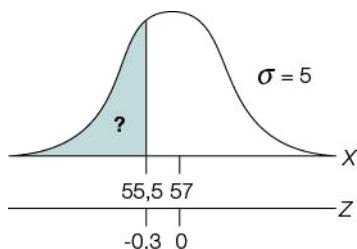
X suit une $B(100 ; 0,57)$ où $X = 0, 1, 2, \dots, 100$.

$P(X \leq 55) = ?$

- Vérification des critères pour utiliser la loi normale :
On a : $np = 100 \times 0,57 = 57 \geq 5$
 $nq = 100 \times 0,43 = 43 \geq 5$
Donc, $B(100 ; 0,57) \approx N(\mu ; \sigma^2)$ où $\mu = np = 57$
 $\sigma = \sqrt{npq} = \sqrt{100 \times 0,57 \times 0,43} = 4,95 \approx 5$

- Correction de continuité :
 $P(X_B \leq 55) \approx P(X_N \leq 55,5)$
- Approximation avec la $N(57 ; 5^2)$:

$$\begin{aligned} P(X_N \leq 55,5) &= P\left(Z \leq \frac{55,5 - 57}{5}\right) \\ &= P(Z \leq -0,3) \\ &= 0,3821 = 38,2 \% \end{aligned}$$



Exercices 3.6

1. a) $N(3412,5 ; 556^2)$, car $\mu = 3412,5$ g et $\sigma = 556$ g.
b) i) 5,3 % (soit $2383 \div 45313$)
ii) $P(X < 2500) \approx P(Z < -1,64) \approx 0,0505 \approx 5,1 \%$
c) $P(2500 < X < 4000) \approx P(-1,64 < Z < 1,06) \approx 1 - 0,0505 - 0,1446 = 0,8049 = 80,5 \%$
Le tableau de distribution donne 82,6 %.
2. a) Presque 48 ans (47,9 ans).
b) 64,4 % ($1 - 0,3336 - 0,0228$)
c) Entre 21 et 63 ans (entre $\mu - 3\sigma$ et $\mu + 3\sigma$).
3. a) 40,4 %
b) Environ 2,3 %.
c) 113 (112,6)
d) 2 personnes sur 100 (soit $0,0192 \times 100$).
4. a) $P(X_N > 12,5)$
b) $P(7,5 < X_N < 9,5)$
5. X suit une $B(400 ; 0,61)$.
 $np = 244 \geq 5$ et $nq = 156 \geq 5$.
 $B(400 ; 0,61) \approx N(244 ; 9,8^2)$
 $P(X_B > 250) \approx P(X_N > 250,5) = 25,5 \%$

6. X suit une $B(100 ; 0,5)$.
 $np = 50 \geq 5$ et $nq = 50 \geq 5$.
 $B(100 ; 0,5) \approx N(50 ; 5^2)$
 - a) $P(49,5 \leq X_N \leq 60,5) = 52,2 \%$
 - b) $P(49,5 \leq X_N \leq 60,5) = 52,2 \%$
 - c) $P(53,5 \leq X_N \leq 54,5) = 5,8 \%$
 - d) $P(X_N > 62,5) = 0,6 \%$
7. a) X suit une $N(60 ; 9,5^2)$; $P(X > 72) = 10,4 \%$.
b) $P(46 < X < 74) = 85,8 \%$
c) On a $z_k = -2,575$. Le fabricant remplacera les piles qui durent moins de 3 ans (35,5 mois).
8. X suit une $B(500 ; 0,55)$.
 $np = 275 \geq 5$ et $nq = 225 \geq 5$.
 $B(500 ; 0,55) \approx N(275 ; 11,1^2)$
 $P(X_B \geq 260) \approx P(X_N \geq 259,5) = 91,9 \%$

Exercices récapitulatifs

1. a) i) La fonction de probabilité.
ii) La table binomiale.
iii) La loi $Po(2,4)$.
iv) La loi $N(7,5 ; 2,7^2)$.
v) La fonction de probabilité. (On ne peut pas utiliser la loi normale, car $nq < 5$.)
- b) $np = 12 \geq 5$ et $nq = 28 \geq 5$.
 $B(40 ; 0,3) \approx N(12 ; 2,9^2)$
 $P(X_B \geq 10) \approx P(X_N \geq 9,5) = 1 - 0,1949 = 80,5 \%$

2. On pose X : «gain net d'un joueur».

Distribution de probabilité de X :
«gain net d'un joueur»

x	8 \$	5 \$	2 \$	-40 \$	Total
$f(x)$	1/2	1/4	1/8	1/8	8/8

Oui, car l'espérance de gain net est de 0,50 \$ par partie pour le joueur. Si l'on joue souvent, on gagne en moyenne 0,50 \$ par partie.

3. a) X suit une $N(30; 0,1^2)$;

$$P(30 - E < X < 30 + E) = P(29,8 \text{ cm} < X < 30,2 \text{ cm}) = 95,4 \%$$

b) 4,6 %

c) On cherche E tel que $P(30 - E < X < 30 + E) = 99 \%$.

De la représentation graphique de la situation, on déduit :

$$E = 2,575 \times 0,1 = 0,26 \text{ cm.}$$

Si le client tolère un écart de 0,26 cm par rapport à la moyenne, l'entreprise vendra 99 % de sa production.

4. a) X suit une $B(8; 0,6)$;

$$P(X = 5) = 27,9 \%$$
 (avec la fonction de probabilité).

b) Non. Pour $E(X) = 4,8$ et $\sigma = 1,4$, la cote z de 7 est 1,57.

Ce n'est pas une cote z exceptionnelle.

c) X suit une $B(8; 0,3)$;

$$P(X > 5) = 1,1 \%$$
 (avec la table binomiale).

5. a) X suit la loi $N(30,2; 5,3^2)$.

b) Pourcentage estimé : 14,9 %.

Pourcentage réel : 14,8 %.

Écart : 0,1 point de pourcentage.

c) 10 % des mères avaient plus de 37 ans $(30,2 + 1,28 \times 5,3)$.

6. a) i) On a $1625 > 20 \times 10 = 200$. Un tirage sans remise peut être considéré comme un tirage avec remise quand la population est grande par rapport à l'échantillon, concrètement quand $N > 20n$.

ii) X suit une $B(10; 0,9)$.

$$f(8) = \binom{10}{8} (0,9)^8 (0,1)^2 = 45 (0,9)^8 (0,1)^2 = 19,4 \%$$

Note: On peut aussi utiliser la table binomiale $B(10; 0,1)$ pour Y : «nombre d'échecs en 10 tirages»:

$$P(X = 8) = P(Y = 2) = f(2) = 19,4 \%$$

$$\text{iii)} \quad \binom{10}{8} = 45 \text{ façons}; (0,9)^8 (0,1)^2 = 0,0043 = 0,4 \%$$

- b) X suit une $B(10\ 000; 0,000\ 03) \approx Po(0,3)$;

$$P(X > 2) = 0,0039 = 0,4 \%$$

7. a) X : «nombre de résidents en faveur du projet sur 100 répondants».

X suit une $B(100; 0,5)$.

$$P(X \geq 60) = ?$$

$$\bullet np = 50 \geq 5 \text{ et } nq = 50 \geq 5$$

$$B(100; 0,5) \approx N(50; 5^2)$$

$$\bullet P(X_B \geq 60) \approx P(X_N \geq 59,5)$$

$$\bullet P(X_N \geq 59,5) = P(Z \geq 1,9) = 2,9 \%$$

- b) Oui, il y a seulement 2,9 % de chances d'obtenir un échantillon comptant au moins 60 résidents favorables au projet dans le cas d'une répartition égale des opinions (50:50); il est fort probable que la répartition réelle soit différente de la répartition présumée par le conseiller.

8. a) X suit une $Po(0,2)$; $P(X = 0) = 81,9 \%$.

- b) X suit une $Po(2)$; $P(X > 3) = 14,3 \%$.

Chapitre 4

Exercices 4.1

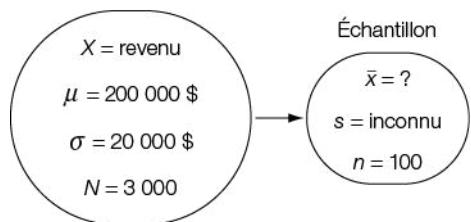
1. a) On sélectionne chaque 10^e individu (60/6) de la liste en commençant par le 3^e. Les individus portant les numéros suivants forment donc l'échantillon : 3, 13, 23, 33, 43, 53.
- b) Les individus portant les numéros suivants formeront l'échantillon : 8, 18, 28, 38, 48, 58.

2. a) L'échantillonnage par grappes.
 b) L'échantillonnage aléatoire simple.
 c) L'échantillonnage à l'aveuglette ou accidentel.
 d) L'échantillonnage de volontaires.
 e) L'échantillonnage systématique.
 f) L'échantillonnage par quotas.
 g) L'échantillonnage stratifié.
 h) Les échantillons décrits en a), b), e) et g) sont aléatoires.

Exercices de compréhension 4.1

1. a) \bar{x}_4
 b) $\bar{x}_2, \bar{x}_3, \bar{x}_4$ et \bar{x}_5
 c) \bar{x}_1 et \bar{x}_6

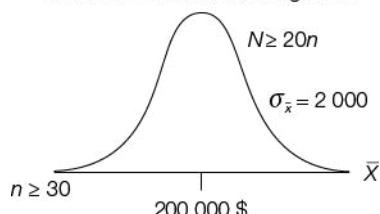
2. a) Population



- b) $n \geq 30; 200\ 000 \text{ $}; \sigma_{\bar{x}}$; Non, car $N \geq 20n$.

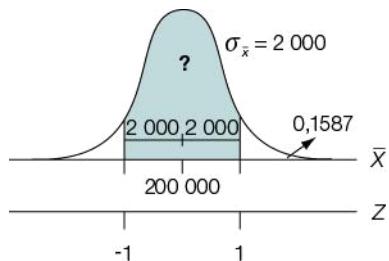
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20\ 000}{\sqrt{100}} = 2\ 000 \text{ $}$$

Distribution d'échantillonnage de \bar{X}



c) 193 800 \$ et 206 500 \$.

d)



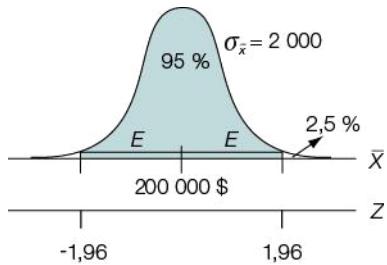
$$P(|\bar{x} - \mu| \leq 2 000) = ?$$

Pour un écart de 2 000 \$, on a :

$$z = \frac{E}{\sigma_{\bar{x}}} = \frac{2 000}{2 000} = 1$$

$$\begin{aligned} P(|\bar{x} - \mu| \leq 2 000) &= P(-1 < z < 1) \\ &= 1 - 2 \times 0,1587 \\ &= 68,3 \% \end{aligned}$$

e) 3 920 \$



On cherche E tel que :

$$P(|\bar{x} - \mu| \leq E \$) = 95 \%$$

$$P(Z > 1,96) = 0,025$$

$$E = z\sigma_{\bar{x}} = 1,96 \times 2 000$$

$$E = 3 920 \$$$

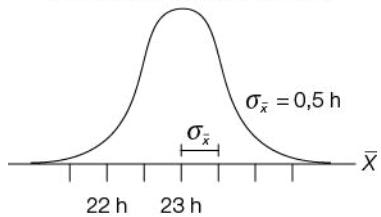
Exercices 4.2

- Si l'on désigne par E l'écart maximal recherché, alors $E = 1,96 \times 1,6 = 3,1$. Pour 95 % des échantillons possibles, l'écart entre \bar{x} et μ est d'au plus 3,1 ans.
- i) 8 étudiants. Toutes les moyennes qui s'écartent de plus de 3,1 ans de μ (dont la valeur est 44,5 ans) ne seront pas dans cette zone, soit les moyennes plus petites que 41,4 ans et plus grandes que 47,6 ans. On compte 8 échantillons sur 145 dans cette situation.
ii) 5,5 % des 145 échantillons étudiés.
Note : Si l'on avait prélevé tous les échantillons possibles, il y en aurait eu 5 %.
- 69,7 % ($101 / 145$) des étudiants ont obtenu un échantillon dont la moyenne est comprise entre 42,9 ans et 46,1 ans.
Note : Si l'on avait prélevé tous les échantillons possibles, il y en aurait eu 68,3 %.

- Paramètres pour la population : $\mu = 0,90$ cm et $\sigma = 0,06$ cm.
Statistiques de l'échantillon : $\bar{x} = 0,88$ cm et $s = 0,075$ cm.
Pour la distribution d'échantillonnage de \bar{X} : $\mu_{\bar{x}} = 0,90$ cm et $\sigma_{\bar{x}} = 0,01$ cm.
- i) Entre 0,72 cm et 1,08 cm : $(0,90 - 3 \times 0,06) \leq \bar{x} \leq (0,90 + 3 \times 0,06)$
ii) Entre 0,87 cm et 0,93 cm : $(0,90 - 3 \times 0,01) \leq \bar{x} \leq (0,90 + 3 \times 0,01)$
Oui.

- i) $\mu = 23$ h, $\bar{x} = 22$ h et $\mu_{\bar{x}} = \mu = 23$ h.
ii) $\sigma = 3$ h, $s = 2,5$ h et
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0,5$$
 h (car $N \geq 20n$).

iii) Distribution d'échantillonnage de \bar{X}

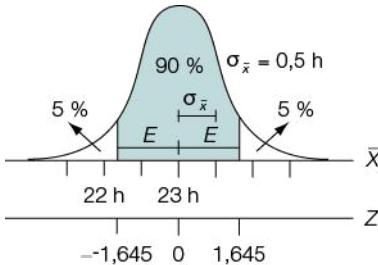


$$b) \bar{x}_{\min} = 23 - 3 \times 0,05 = 21,5 \text{ h}$$

c) Soit E l'écart maximal cherché pour cette zone.

$$\text{On a } E = 1,645 \times 0,5 = 0,8 \text{ h.}$$

Pour 90 % des échantillons possibles, l'écart entre \bar{x} et μ est d'au plus 0,8 h.



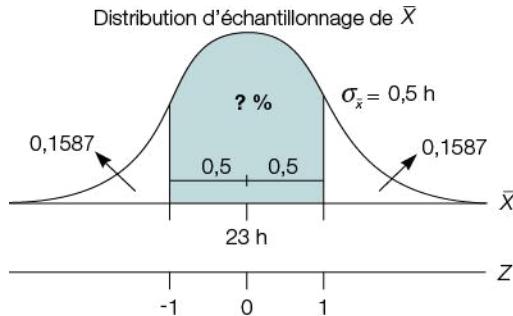
- i) 1 h ; non, car l'écart est plus grand que 0,8 h [voir la représentation graphique en c)].

- ii) 22,8 h ou 23,3 h (ou tout autre nombre compris entre 22,2 h et 23,8 h).

- 68,3 %. Pour toute courbe normale, on sait qu'il y a 68,3 % des données qui se situent à au plus 1 écart type (0,5 h dans ce cas-ci) de la moyenne.

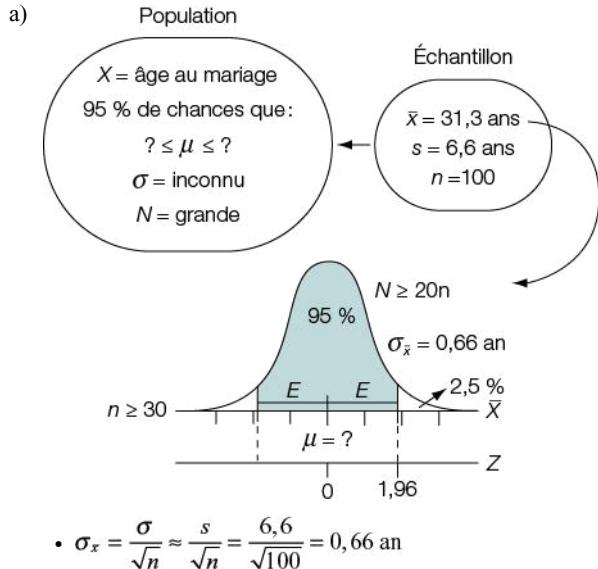
On peut obtenir ce pourcentage comme suit : $Z = 0,5 / 0,5 = 1$

$$\begin{aligned} P(|\bar{x} - \mu| \leq 0,5 \text{ h}) &= P(-1 \leq z \leq 1) \\ &= 1 - 2 \times 0,1587 \\ &= 0,6826 = 68,3 \% \end{aligned}$$



1. a) 80,6 % de chances. b) 71,6 % de chances.
2. a) $n \geq 30$; 8 457 \$; 149,75 \$ (il faut utiliser le facteur de correction). b) 90,5 % c) 192 \$(191,68 \$)
3. a) Le poids du colis. Le kilogramme. b) 68,3 %, car les moyennes \bar{x} comprises entre 3,2 kg et 3,6 kg ont une cote z comprise entre -1 et 1. c) $\sigma_{\bar{x}} = 0,2$ kg, car de 3,4 kg à 3,6 kg il y a 1 écart type, puisque la cote z de 3,6 est 1. d) $\mu = 3,4$ kg (le centre de la courbe normale, car $\mu_{\bar{x}} = \mu$). e) $\sigma = 1,6$ kg, car $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \Rightarrow 0,2 = \frac{\sigma}{\sqrt{64}}$ f) 3. [3,2 kg ; 3,4 kg], car l'aire sous la courbe au-dessus de cet intervalle est la plus grande.
4. a) $\mu = 6$ h et $\sigma = 2,4$ h. b) 4,33 h; 5,67 h; 6,67 h. c) Distribution d'échantillonnage de \bar{X} . d) $\mu_{\bar{x}} = 6$ h et $\sigma_{\bar{x}} = 1$ h. e) Oui; on a $\mu_{\bar{x}} = \mu = 6$ h. f) Oui; en remplaçant σ , N et n par leurs valeurs dans l'égalité, on trouve $\sigma_{\bar{x}} = 1$ h.

Exercice de compréhension 4.2



- Marge d'erreur: $E = z\sigma_{\bar{x}} = 1,96 \times 0,66 = 1,3 \text{ an}$
- Intervalle de confiance: $\mu = 31,3 \pm 1,3 \text{ an}$
 $\mu \in [30,0 \text{ ans}; 32,6 \text{ ans}]$
Interprétation de l'intervalle de confiance
 95 ; 30 ; 32,6.

- b) i) Méthodologie: 100 ; 1,3 ; 19.
 ii) Ponctuelle; oui; sur la marge d'erreur de 1,3 an, considérée comme petite par rapport à 31,3 ans.
- c) i) Faux (d'au plus 1,3 an) ii) Faux

Exercice de compréhension 4.3

- a) $E = 30 \text{ ml}$ $\sigma = 80 \text{ ml}$
 Niveau de confiance de 99 % $\Rightarrow z = 2,575$

$$E = z\sigma_{\bar{x}} \Leftrightarrow E = z \frac{\sigma}{\sqrt{n}}$$

$$30 = 2,575 \times \frac{80}{\sqrt{n}}$$

$$\sqrt{n} = \frac{2,575 \times 80}{30} = 6,87$$

$$n = (6,87)^2 = 47,2$$

Un minimum de 48 contenants.

- b) i) 150 ii) 1 ; 2.

Exercices 4.3

1. a) $\bar{x} = 49,7 \text{ g}$ et $E = z\sigma_{\bar{x}} = 0,1 \text{ g}$.

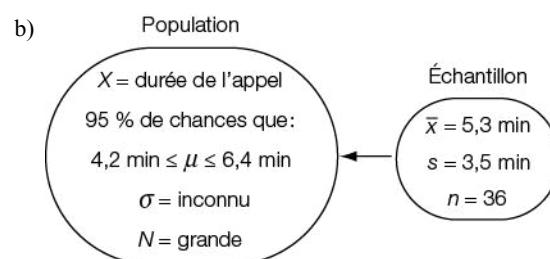
Intervalle de confiance: $\mu \in [49,6 \text{ g}; 49,8 \text{ g}]$.

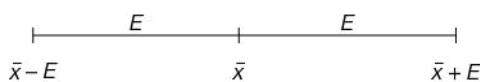
Interprétation

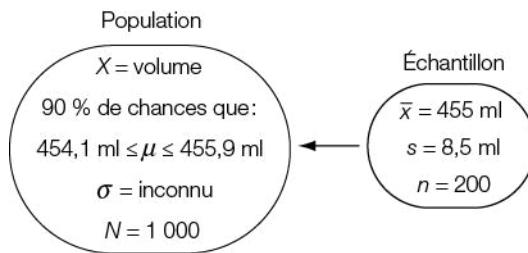
Il y a 95 % de chances que le poids moyen de l'ensemble des contenants remplis par la machine se situe entre 49,6 g et 49,8 g.

- b) Le risque d'erreur est de 5 %.
Interprétation
 Il y a 5 % de chances que le véritable poids moyen de tous les contenants remplis par la machine soit plus petit que 49,6 g ou plus grand que 49,8 g.
- c) La marge d'erreur est $E = 0,1 \text{ g}$.
Interprétation
 Il y a 95 % de chances d'avoir un écart d'au plus 0,1 g entre le poids moyen des 100 contenants de l'échantillon et le poids moyen de tous les contenants produits.

2. a) i) $z = 1,28$ ii) $z = 1,81$ iii) $z = 2,17$
 b) i) Le niveau de confiance de 80 %.
 ii) Le niveau de confiance de 80 %.
3. a) Pour $n = 36$, $\bar{x} = 5,3 \text{ min}$ et $s = 3,5 \text{ min}$, on trouve l'intervalle [4,2 min ; 6,4 min].



- c) 5,3. Méthodologie : 36; 1,1; 19.
4. Pour $n = 125$, $\bar{x} = 337 \text{ ml}$ et $\sigma = 3 \text{ ml}$, on trouve l'intervalle $[336,5 \text{ ml}; 337,5 \text{ ml}]$.
5. a) i) Faux. C'est la moyenne μ de la production qui a 95 % de chances de se situer quelque part entre les bornes de l'intervalle construit.
Les chances que la moyenne \bar{x} de l'échantillon se situe dans l'intervalle de confiance ne sont pas de 95 %, mais de 100 %. En effet, \bar{x} , soit 337 ml, est nécessairement au centre de l'intervalle de confiance puisque les bornes de cet intervalle sont $\bar{x} - E$ et $\bar{x} + E$.
- 
- ii) Faux. L'affirmation est imprécise : qu'est-ce qui est dans l'intervalle construit ? Vous ? μ ? \bar{x} ?
iii) Vrai iv) Vrai
- b) i) Faux
ii) Faux. Il y a 0 % de risques que \bar{x} ne soit pas dans l'intervalle [voir 5a) i)].
iii) Vrai
iv) L'affirmation est imprécise : de quelle moyenne s'agit-il ?
 - Si c'est de la moyenne \bar{x} de l'échantillon, c'est faux : il y a 0 % de risques.
 - Si c'est de la moyenne μ de la population, c'est vrai.
6. a) $E = t\sigma_{\bar{x}}$ où t suit la loi de Student.
b) i) $t = 2,977$ ii) $t = 2,052$
c) Le niveau de confiance de 95 % avec $n = 28$.
7. $t = 2,831$; $\mu \in [44,4 \text{ min}; 51,6 \text{ min}]$.
8. a) Attention ! Il faut utiliser le facteur de correction, car $N < 20n$.



- b) 10 %; 0,9 ml.
c) i) Une estimation ponctuelle.
ii) Une estimation par intervalle de confiance.
d) $E = 3 \times 0,54 = 1,6 \text{ ml}$
L'emploi de cette marge d'erreur pour estimer μ augmenterait les chances que l'intervalle de confiance contienne μ à presque 100 % (exactement 99,7 %), mais il diminuerait la précision de l'estimation en accroissant la marge d'erreur de 0,9 ml à 1,6 ml. On fait donc le choix de courir un certain risque que l'intervalle construit ne contienne pas μ pour augmenter la précision de l'estimation.

9. a) Pour $n = 500$, $\bar{x} = 13,9 \text{ kg}$ et $s = 1,4 \text{ kg}$, on a l'intervalle $[13,8 \text{ kg}; 14,0 \text{ kg}]$.
b) Oui, car la marge d'erreur n'est pas très grande (elle est inférieure à 0,1 kg) par rapport à 13,9 kg.
10. a) Plus grande. b) Plus petite.
11. a) $\mu = 34 \text{ minutes}$; c'est une estimation très acceptable, car la marge d'erreur n'est que de 1,3 minute.

- b) $\mu \in [32,7 \text{ min}; 35,3 \text{ min}]$
Interprétation
Il y a 95 % de chances que le temps moyen que les Québécois de moins de 35 ans passent chaque jour sur Internet à s'informer de l'actualité se situe en réalité entre 32,7 minutes et 35,3 minutes.

12. $\bar{x} = 5,62$, $s = 0,13$, $t = 2,947$;
 $\mu \in [5,53 \text{ L}/100 \text{ km}; 5,71 \text{ L}/100 \text{ km}]$.
13. Au moins 74 sacs.
14. Au moins 189 familles.

15. Au moins 295 (294,1) appels. (On utilise $s = 3,5$ comme estimateur de σ .)

Exercice de compréhension 4.4

$$\text{a) } \hat{p} = \frac{320+256}{800} = 72\%; 90\%.$$

$$\bullet \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{72 \times 28}{800}} = 1,6\%$$

- Marge d'erreur:
 $E = z\sigma_{\hat{p}} = 1,645 \times 1,6\% = 2,6\%$
- Intervalle de confiance:
 $p = 72 \pm 2,6\%$
 $p \in [69,4\%; 74,6\%]$

- b) 72 %. Méthodologie : 800; 2,6; 18. À la radio : 72 %.
c) La radio.

Exercice de compréhension 4.5

- a) $E = 3\%$; niveau de confiance de 95 %, d'où $z = 1,96$.

$$E = z\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$3 = 1,96\sqrt{\frac{50 \times 50}{n}}$$

$$3^2 = 1,96^2 \left(\frac{50 \times 50}{n} \right)$$

$$n = \frac{1,96^2 \times 50 \times 50}{3^2} = 1\,067,1$$

Il faut un minimum de 1 068 personnes.

$$\text{b) } E = z\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$3 = 1,96\sqrt{\frac{20 \times 80}{n}}$$

$$3^2 = 1,96^2 \left(\frac{20 \times 80}{n} \right)$$

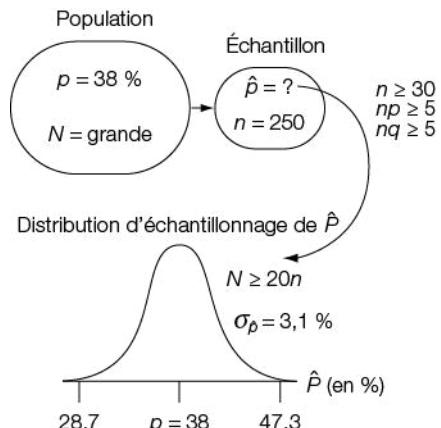
$$n = \frac{1,96^2 \times 20 \times 80}{3^2} = 683$$

Il faut un minimum de 683 personnes.

Exercices 4.4

- $$1. \text{ a) } \hat{p}_{\min} = 38\% - 3 \times 3,1\% = 28,7\%$$

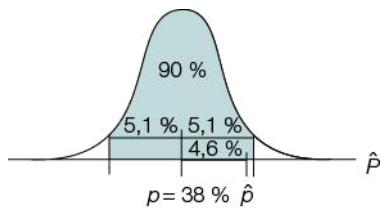
$$\hat{p}_{\max} = 38\% + 3 \times 3,1\% = 47,3\%$$



- b) i) $E = 1,28 \times 3,1 = 4\%$

ii) Non, l'écart entre \hat{p} et p est de 4,6 %
(soit $42,6 - 38,0$): il est supérieur à l'écart de 4 % calculé en i).

iii) Oui, l'écart de 4,6 % entre le pourcentage \hat{p} et p est inférieur à $E = 5,1\%$ ($1,645 \times 3,1$). La position approximative de \hat{p} est indiquée ci-dessous.



2. a) Non. Comme $p = 2\%$, on a $np = 100 \times 0,02 = 2 < 5$: une des conditions permettant d'affirmer que la distribution d'échantillonnage de \hat{P} suit une normale n'est pas respectée.

- b) Soit la variable aléatoire X : «nombre de pièces défectueuses parmi 100». Pour un pourcentage de pièces défectueuses $\hat{p} \geq 3\%$, il faut que $X \geq 3$. X suit une $B(100; 0,02)$, où $n > 20$, $p = 0,02 < 0,1$ et $np = 2 \leq 5$. On peut donc utiliser la loi de Poisson où $\lambda = 2$ pour effectuer les calculs (on peut aussi utiliser la fonction de probabilité d'une binomiale).

$$\begin{aligned} P(\hat{p} \geq 3\%) &= P(X \geq 3) = 1 - 0,677 = 0,323 \\ &= 32,3\% \text{ pour une Po}(2) \end{aligned}$$

3. a) 78,9 % b) 3,1 %

4. a) On a $N = 4\ 536$; le pourcentage de femmes dans l'usine est $p = 3\ 280 \div 4\ 536 = 72,3\%$.

Taille de l'échantillon	Écart type σ_p	Marge d'erreur E
$n = 100$	4,5 %	8,8 %
$n = 150$	3,7 %	7,3 %
$n = 200$	3,2 %	6,3 %

- b) Quand la taille de l'échantillon augmente, la marge d'erreur de l'estimation diminue.

5. a) 2,7 % (avec $\hat{p} = 82 \%$, on a $E = 1,96 \times 1,39 \%$.)

- b) $p \in [79,3\% ; 84,7\%]$

Interprétation

Il y a 95 % de chances que le pourcentage réel d'internautes qui utilisent les réseaux sociaux se situe entre 79,3 % et 84,7 % en 2013.

- c) i) La marge d'erreur sera plus grande.
 ii) Le risque d'erreur sera plus petit.
 iii) L'intervalle de confiance sera plus grand.

d) La marge d'erreur est de 9,5 %. Non, la marge d'erreur est trop grande.

6. a) $E = 0,5 \%$, $p \in [1\% ; 2\%]$

Interprétation

Il y a 90 % de chances que le pourcentage réel d'enfants de 5 à 9 ans allergiques aux arachides se situe entre 1 % et 2 % au Québec.

- b) Aucun risque (0 %). Le pourcentage \hat{p} de l'échantillon est toujours au centre de l'intervalle de confiance puisque, pour construire l'intervalle, on prend $\hat{p} \pm E$, soit $1,5\% \pm 0,5\%$.

7. a) 3,1 %. Comme le texte de l'article donne le pourcentage pour plusieurs sources de divertissement, la méthodologie ne peut pas énumérer toutes les marges d'erreur associées à chacun de ces pourcentages. On y présente donc la marge d'erreur maximale, soit celle que l'on obtient en posant $\hat{p} = 50\%$, ce qui donne une marge d'erreur d'au plus 3,1 % pour un échantillon de 1 000 répondants.

- b) La marge d'erreur pour cette question est 2,9 %, ce qui est compatible avec la marge d'erreur d'au plus 3,1 % de la méthodologie.

- $$8. \text{ a) i) } 97\% \quad \text{ii) } p \in [95\%; 99\%]$$

- b) 5 %

- c) Il y a 95 % de chances que l'écart soit d'au plus 2 % entre le pourcentage de 97 % trouvé dans l'échantillon et le pourcentage réel de diplômés qui occupent un emploi à temps plein 2 ans après l'obtention de leur bac en administration.

- d) i) Pour $E = 2\%$ et $\hat{p} = 97\%$, on obtient
 $n = 280$ personnes.

- ii) $n = 2\,401$. En posant $\hat{p} = 50\%$ dans le calcul de la taille de l'échantillon, on s'assure que les estimations faites avec les différents pourcentages échantillonaux auront une marge d'erreur d'au plus 2 %.

- e) i) Augmenter la taille de l'échantillon.
ii) Diminuer la taille de l'échantillon.

9. a) Environ 385 personnes. Comme on ignore la valeur de \hat{p} , on utilise $\hat{p} = 50\%$ pour calculer n .

- b) $\hat{p} = 160 / 385 = 41,6\%$; on obtient l'intervalle $[36,7\%; 46,5\%]$.

10. a) On a $\hat{p} = 56\%$ et $E = 3,9\%$:
l'intervalle de confiance est $[52,1\%; 59,9\%]$.
b) Environ 2 367 personnes.

Exercices récapitulatifs

1. a) Les acteurs âgés de 25 à 35 ans.
b) C'est très peu probable. En effet, si on néglige les valeurs ayant moins de 0,3 % de chances d'être obtenues, la plus petite moyenne possible pour un échantillon de taille 60 est:

$$\bar{x}_{\min} = \mu - 3\sigma_{\bar{x}} = 38,7 - 3 \times \frac{11,6}{\sqrt{60}} = 34,2 \text{ ans}$$

c) $2,2 (1,44 \times 1,5)$
d) L'écart sera d'au plus 5,4 %. Attention ! On a une petite population, car $N < 20n$.

$$E = 1,645 \times \sqrt{\frac{46,6 \times 53,4}{200}} \sqrt{\frac{1400 - 200}{1400 - 1}}$$

2. a) 56,7 %; la marge d'erreur est de 5,6 %.

- b) Entre 43,5 ans et 55,5 ans; on a $\bar{x} = 49,5$ ans,
 $s = 15,5$ ans, $t = 2,052$ et $E = 6$ ans.

- c) $\bar{x} = 46,25$ \$, $s = 31,08$ \$ et $E = 3,52$ \$.

$$\mu \in [42,73 \$; 49,77 \$]$$

Il y a 95 % de chances que le montant moyen des achats, pour l'ensemble des clients, se situe entre 42,73 \$ et 49,77 \$.

- d) 15,5 ; 24,5. (On a $\hat{p} = 20\%$ et $E = 4,5\%$.)
e) Un sondage effectué par un centre de jardinage auprès de sa clientèle a donné les résultats suivants : 56,7 % des clients sont des femmes et l'âge moyen de la clientèle est de 49,5 ans ; on estime qu'à chacune de leur visite, les clients achètent en moyenne pour 46,25 \$ et ils utilisent la carte de débit pour régler leurs achats dans une proportion de 20 %.

Méthodologie

Ce sondage a été effectué auprès d'un échantillon aléatoire de 300 clients. Pour un échantillon de cette taille, la marge d'erreur est la suivante pour chacune des variables étudiées, 19 fois sur 20 : 5,6 % pour le sexe ; 6 ans pour l'âge ; 3,52 \$ pour le montant des achats ; 4,5 % pour le mode de paiement.

Chapitre 5

Exercice de compréhension 5.1

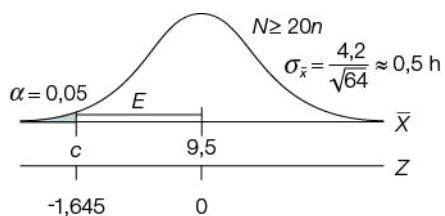
- a) 1. $\begin{cases} H_0: \mu = 9,5 \text{ h} \\ H_1: \mu < 9,5 \text{ h} \\ \alpha = 0,05 \end{cases}$

On a: $n = 64$

$$\bar{x} = 8,3 \text{ h}$$

$$s = 4,2 \text{ h}$$

2. $n \geq 30$



Point critique

- $E = z\sigma_{\bar{x}} = 1,645 \times 0,5 = 0,8 \text{ h}$
- $c = 9,5 - 0,8 = 8,7 \text{ h}$

Règle de décision

Rejeter H_0 si la moyenne \bar{x} de l'échantillon est inférieure à 8,7 h.

Décision et conclusion

Comme $\bar{x} = 8,3 \text{ h} < 8,7 \text{ h}$, on rejette H_0 . Les adolescents de 2012 consacrent en moyenne moins d'heures par semaine à l'écoute de la radio que ceux de 2005.

- b) À au plus 5 %
c) H_0
d) H_1
e) H_0

Exercices 5.1

- 93,3 est situé à plus de 3 écarts types de μ . C'est donc une valeur qui a très peu de chances d'être obtenue comme moyenne d'échantillon quand la moyenne de la population est 100. Il faut rejeter H_0 .
• 101,3 est situé à moins de 1 écart type de μ . L'écart entre \bar{x} et μ est sûrement imputable au hasard de l'échantillonnage. On ne rejette pas H_0 .
• 104,6 est situé dans une zone de la courbe d'échantillonnage de \bar{X} où il n'est pas facile de décider si l'écart entre \bar{x} et μ est trop grand pour être attribuable au hasard. On a besoin d'une règle pour prendre une décision. Si $\bar{x} = 104,6$, la construction d'un test d'hypothèse s'avère nécessaire.
2. a) L'hypothèse H_0 .
b) La probabilité de rejeter l'hypothèse H_0 alors que cette hypothèse est vraie.
c) L'affirmation est fausse. Un test d'hypothèse ne prouve jamais que l'hypothèse H_0 est vraie ; seule une analyse complète de la population permettrait de le montrer. Dans un test d'hypothèse, lorsqu'on décide de ne pas rejeter l'hypothèse nulle, cela signifie qu'il n'y a aucune évidence statistique justifiant son rejet.
3. a) $H_0: \mu = 500 \text{ ml}$
 $H_1: \mu \neq 500 \text{ ml}$
b) $H_0: \mu = 2,7 \text{ nuitées}$
 $H_1: \mu > 2,7 \text{ nuitées}$
c) $H_0: \mu = 35 \text{ mm}$
 $H_1: \mu \neq 35 \text{ mm}$
d) $H_0: \mu = 6,25 \text{ cm}$
 $H_1: \mu < 6,25 \text{ cm}$
e) $H_0: \mu = 4,4 \text{ jours par année}$
 $H_1: \mu < 4,4 \text{ jours par année}$

4. a) $H_0: \mu = 5 \text{ kg}$

$H_1: \mu \neq 5 \text{ kg}$

$c_1 = 4,95 \text{ kg}$ et $c_2 = 5,05 \text{ kg}$

Règle de décision

Rejeter H_0 si la moyenne de l'échantillon prélevé est supérieure à 5,05 kg ou inférieure à 4,95 kg.

b) Il y en a deux : le lundi à 10 h et le mardi à 16 h.

c) Statistiquement, un seuil de signification de 0,05 veut dire qu'il y a 5 % de risques de rejeter l'hypothèse H_0 alors que celle-ci est vraie. Dans le contexte, le seuil de 0,05 signifie qu'il y a 5 % de risques de conclure que la machine est déréglée alors qu'en fait, il n'en est rien.

5. $H_0: \mu = 50 \text{ kg/cm}^2$

$H_1: \mu > 50 \text{ kg/cm}^2$

$\sigma_{\bar{x}} = 0,38 \text{ kg/cm}$

Règle de décision

Rejeter H_0 si la moyenne \bar{x} de l'échantillon prélevé est supérieure à 50,9 kg/cm^2 .

Décision et conclusion

Puisque $\bar{x} = 54,5 \text{ kg/cm}^2 > 50,9 \text{ kg/cm}^2$, on rejette H_0 . Oui, on peut dire qu'il y a augmentation significative de la résistance moyenne à la rupture.

L'écart de 4,5 kg/cm^2 entre \bar{x} et μ est statistiquement significatif : il y a tout lieu de croire que cette augmentation de la résistance est attribuable à l'introduction du nouvel alliage.

6. La moyenne de l'échantillon est $\bar{x} = 6,0 \text{ min}$ et l'écart type de la population est $\sigma = 3,2 \text{ min}$.

Note : On a $n < 30$ mais, comme la distribution du temps de service suit un modèle normal, la distribution d'échantillonnage de \bar{X} suit un modèle normal.

$H_0: \mu = 8,3 \text{ min}$

$H_1: \mu < 8,3 \text{ min}$

$\sigma_{\bar{x}} = 0,64 \text{ min}$

Règle de décision

Rejeter H_0 si la moyenne \bar{x} de l'échantillon prélevé est inférieure à 7,2 min.

Décision et conclusion

Comme $\bar{x} = 6 \text{ min} < 7,2 \text{ min}$, on rejette H_0 .

L'informatisation a permis d'accélérer le service à la clientèle.

7. $H_0: \mu = 23,3 \text{ mois}$

$H_1: \mu > 23,3 \text{ mois}$

$\sigma_{\bar{x}} = 3,9 \text{ mois}$ (Attention ! Il faut utiliser le facteur de correction, car $N < 20n$.)

Règle de décision

Rejeter H_0 si la moyenne \bar{x} de l'échantillon est supérieure à 29,7 mois.

Décision et conclusion

Comme $\bar{x} = 30,8 \text{ mois} > 29,7 \text{ mois}$, on rejette H_0 .

La moyenne d'âge des enfants au moment de leur adoption en 2011 est plus élevée que la moyenne d'âge de 23,3 mois observée dans les années 1990.

8. $H_0: \mu = 10 \text{ min}$

$H_1: \mu > 10 \text{ min}$

$\bar{x} = 11 \text{ min}; s = 0,97 \text{ min}; \sigma_{\bar{x}} = 0,2 \text{ min}$.

On utilise la loi de Student avec $dl = 24$ et $t_c = 1,711$.

Règle de décision

Rejeter H_0 si la moyenne \bar{x} de l'échantillon est supérieure à 10,3 min.

Décision et conclusion

Comme $\bar{x} = 11 \text{ min} > 10,3$, on rejette H_0 .

L'écart observé entre \bar{x} et μ est statistiquement significatif. Le temps moyen requis pour remplir le formulaire est supérieur à 10 minutes.

9. $H_0: \mu = 5 \text{ min}$

$H_1: \mu > 5 \text{ min}$

$\bar{x} = 5,3 \text{ min}; s = 1,9 \text{ min}; \sigma_{\bar{x}} = 0,4 \text{ min}$.

On utilise la loi de Student avec $dl = 19$ et $t_c = 1,729$.

Règle de décision

Rejeter H_0 si la moyenne \bar{x} de l'échantillon est supérieure à 5,7 min.

Décision et conclusion

Comme $\bar{x} = 5,3 \text{ min} < 5,7 \text{ min}$, on ne rejette pas H_0 . Les données échantillonnelles ne permettent pas de douter du respect de la norme de 5 minutes comme temps moyen d'intervention pour la première équipe de pompiers.

10. a) $H_0: \mu = 1\ 000 \text{ h}$

$H_1: \mu > 1\ 000 \text{ h}$

$c = 1\ 039,1 \text{ h}$

On rejette H_0 . La durée de vie moyenne est supérieure à 1 000 h.

b) Il y a moins de 1 % de chances d'obtenir une moyenne échantillonnable de 1 050 heures si la moyenne de la population est de 1 000 heures. La probabilité est tellement faible [en fait, $P(\bar{x} > 1\ 050) = P(z > 2,98) = 0,14 \text{ \%}$] qu'on décide de rejeter H_0 en espérant qu'une telle situation ne se soit pas produite.

Les risques de se tromper en décidant de rejeter H_0 sont d'au plus 1 % (exactement 0,14 %).

c) Non, on ne rejette pas H_0 .

Non, cela signifie simplement que l'écart entre \bar{x} et μ n'est pas assez significatif statistiquement pour permettre de rejeter H_0 . Il faudrait étudier la totalité de la population pour déterminer la véritable moyenne. Mais une telle étude entraînerait inévitablement une destruction complète de la production.

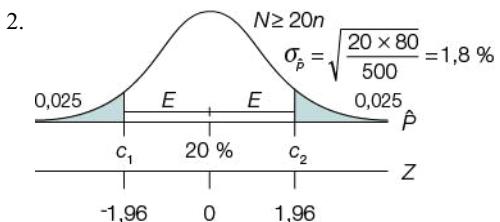
Exercices de compréhension 5.2

1. a) 1. $| H_0: p = 20 \text{ \%}$

$| H_1: p \neq 20 \text{ \%}$

$| \alpha = 0,05$

On a $n = 500$.



3. Points critiques

$$E = z\sigma_{\hat{p}} = 1,96 \times 1,8 \% = 3,5 \%$$

$$c_1 = 20 \% - 3,5 \% = 16,5 \%$$

$$c_2 = 20 \% + 3,5 \% = 23,5 \%$$

4. Règle de décision

Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est inférieur à 16,5 % ou supérieur à 23,5 %.

- b) Pourcentage \hat{p} : n° 1 : 21,0 % et n° 2 : 20,4 %.
On rejette H_0 pour les échantillons suivants:
n° 4: $\hat{p} = 16 \% < 16,5 \%$; il n'y a pas assez de bonbons rouges.
n° 7: $\hat{p} = 24 \% > 23,5 \%$; il y a trop de bonbons rouges.
- c) 5 %

2. a) i) 25 %
ii) $H_0: p = 25 \%$; $H_1: p > 25 \%$.
iii) 4 %

b) $\sigma_{\hat{p}} = \sqrt{\frac{25 \times 75}{100}} = 4,3 \%$

- c) Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est supérieur à 32,6 %.

- d) Exemples de valeurs qui permettraient:
– de rejeter H_0 : $\hat{p} = 33,4 \%$ et $\hat{p} = 34,1 \%$.
– de ne pas rejeter H_0 : $\hat{p} = 29,5 \%$ et $\hat{p} = 31,2 \%$.

Exercices 5.2

1. a) $H_0: p = 80 \%$
 $H_1: p < 80 \%$

Règle de décision

Rejeter H_0 si le pourcentage de l'échantillon est inférieur à 77,2 %.

Décision et conclusion

Comme $\hat{p} = 78,6 \% > 77,2 \%$, on ne rejette pas H_0 . Les données échantillonnelles ne permettent pas de rejeter l'affirmation du sociologue.

- b) Non, mais on n'a pas pu prouver statistiquement qu'il avait tort; on doit donc lui accorder le bénéfice du doute.

2. a) $H_0: p = 76 \%$
 $H_1: p \neq 76 \%$
b) H_0
c) Au plus 3 % (soit $2 \times 1,5 \%$)
d) C'est le pourcentage de la population.
e) $\sigma_{\hat{p}} = \sqrt{\frac{76 \times 24}{1300}} = 1,2 \%$
f) Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est inférieur à 73,4 % ou supérieur à 78,6 %.
g) 72 % (plus petit que 73,4 %); 79,5 % (plus grand que 78,6 %).

3. $H_0: p = 49,4 \%$

$H_1: p > 49,4 \%$

$$\sigma_{\hat{p}} = \sqrt{\frac{49,4 \times 50,6}{1003}} = 1,6 \%$$

Règle de décision

Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est supérieur à 52 %.

Décision et conclusion

Comme $\hat{p} = 54,8 \% > 52 \%$, on rejette H_0 .

Trois semaines après le référendum, le pourcentage de Québécois en accord avec la souveraineté du Québec a augmenté: il est probable qu'il se situe autour de 55 %.

4. $H_0: p = 65 \%$

$H_1: p > 65 \%$

$$\sigma_{\hat{p}} = \sqrt{\frac{65 \times 35}{200}} \times \sqrt{\frac{1500 - 200}{1500 - 1}} = 3,1 \%$$

Règle de décision

Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est supérieur à 70,1 %.

Décision et conclusion

Comme $\hat{p} = 68 \% < 70,1 \%$, on ne rejette pas H_0 .

La campagne de sensibilisation ne semble pas avoir fait augmenter le pourcentage de jeunes conscients des dangers de la divulgation de renseignements personnels sur Internet. L'écart observé entre p et \hat{p} n'est pas significatif: il est probablement attribuable au hasard de l'échantillonnage.

5. a) $H_0: p = 80 \%$

$H_1: p \neq 80 \%$

$n = 1200$

$$\sigma_{\hat{p}} = \sqrt{\frac{80 \times 20}{1200}} = 1,2 \%$$

Règle de décision

Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est inférieur à 76,9 % ou supérieur à 83,1 %.

- b) Les taux de survie sont les suivants:

Abitibi: 76 %	Saguenay: 85 %
Côte-Nord: 78 %	Gaspésie: 82 %

Conclusion

Le taux de survie des plants est différent de 80 % en Abitibi et au Saguenay. Il semble inférieur au taux prévu en Abitibi, et supérieur au Saguenay.

6. $H_0: p = 14,6 \%$

$H_1: p < 14,6 \%$

$$\sigma_{\hat{p}} = \sqrt{\frac{14,6 \times 85,4}{150}} = 2,9 \%$$

Règle de décision

Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est inférieur à 9,8 %.

Décision et conclusion

Comme $\hat{p} = 9,3 \% < 9,8 \%$, on rejette H_0 .

Le pourcentage de Canadiens âgés de 16 à 25 ans qui se situent à un très faible niveau de compétence en compréhension de texte est inférieur à 14,6 %.

Exercices 5.3

1. $H_0: \mu_{UL} - \mu_M = 0$

$H_1: \mu_{UL} - \mu_M < 0$

$\sigma_{\bar{x}_M} = 1,1 \text{ an}$ et $\sigma_{\bar{x}_{UL}} = 1,0 \text{ an}$

$\sigma_{\bar{x}_{UL} - \bar{x}_M} = \sqrt{1,1^2 + 1,0^2} = 1,5 \text{ an}$

Règle de décision

Rejeter H_0 si $\bar{x}_{UL} - \bar{x}_M$ est inférieur à -3,5 ans.

Décision et conclusion

Comme $\bar{x}_{UL} - \bar{x}_M = -5,1 < -3,5$, on rejette H_0 .

Au moment de la rupture du couple, les personnes vivant en union libre sont en moyenne plus jeunes que celles qui sont mariées.

2. $H_0: \mu_F - \mu_H = 0$

$H_1: \mu_F - \mu_H > 0$

$\sigma_{\bar{x}_F - \bar{x}_H} = \sqrt{0,08^2 + 0,08^2} = 0,11 \text{ h}$

Règle de décision

Rejeter H_0 si $\bar{x}_F - \bar{x}_H$ est supérieur à 0,18 h.

Décision et conclusion

Comme $\bar{x}_F - \bar{x}_H = 0,20 > 0,18$, on rejette H_0 .

Les données échantillonnelles permettent d'affirmer que les femmes dorment en moyenne plus longtemps que les hommes.

3. $H_0: \mu_F - \mu_G = 0$

$H_1: \mu_F - \mu_G > 0$

$\sigma_{\bar{x}_F - \bar{x}_G} = \sqrt{0,4^2 + 0,4^2} = 0,6$

Règle de décision

Rejeter H_0 si $\bar{x}_F - \bar{x}_G$ est supérieur à 1,0.

Décision et conclusion

Comme $\bar{x}_F - \bar{x}_G = 2,7 > 1,0$, on rejette H_0 .

Les filles arrivent à l'école mieux préparées que les garçons.

4. $H_0: \mu_d = 0$

$H_1: \mu_d > 0$

$\bar{d} = 4,0$ et $s_d = 3,6$

$\sigma_{\bar{d}} = 1,5$ et $t_c = 2,015$

Règle de décision

Rejeter H_0 si la moyenne \bar{d} des différences observées au test de mémoire dans l'échantillon est supérieure à 3.

Décision et conclusion

Comme $\bar{d} = 4,0 > 3$, on rejette H_0 .

Le programme d'activités augmente la mémoire des personnes âgées.

5. $H_0: \mu_d = 0$

$H_1: \mu_d < 0$

$\bar{d} = -7$ et $s_d = 3$

$\sigma_{\bar{d}} = 1,2$ et $t_c = 3,365$

Règle de décision

Rejeter H_0 si la moyenne \bar{d} des différences observées au test de niveau de stress dans l'échantillon est inférieure à -4.

Décision et conclusion

Comme $\bar{d} = -7 < -4$, on rejette H_0 .

Les ateliers font diminuer le niveau de stress des étudiants.

6. $H_0: \mu_d = 0$

$H_1: \mu_d < 0$

$\bar{d} = -5 \text{ jours}$ et $s_d = 1,7 \text{ jour}$

$\sigma_{\bar{d}} = 0,8 \text{ jour}$ et $t_c = 3,747$

Règle de décision

Rejeter H_0 si la moyenne \bar{d} des différences du nombre moyen de jours de germination observées dans l'échantillon est inférieure à -3 jours.

Décision et conclusion

Comme $\bar{d} = -5 \text{ jours} < -3 \text{ jours}$, on rejette H_0 .

Les graines enrobées germent plus rapidement que les graines non enrobées. Le produit est efficace.

7. $H_0: p_G - p_F = 0$

$H_1: p_G - p_F > 0$

$\hat{p} = 27,6 \%$

$\sigma_{\hat{p}_G - \hat{p}_F} = \sqrt{1,6^2 + 1,6^2} = 2,3 \%$

Règle de décision

Rejeter H_0 si la différence $\hat{p}_G - \hat{p}_F$ est supérieure à 5,4 %.

Décision et conclusion

Comme $\hat{p}_G - \hat{p}_F = 5,6 \% > 5,4 \%$, on rejette H_0 .

La proportion d'enfants de 4 ans qui écoutent fréquemment la télévision pendant les repas est plus élevée chez les garçons que chez les filles.

8. $H_0: p_1 - p_2 = 0$

$H_1: p_1 - p_2 > 0$

$\hat{p}_1 = 81 \%$ et $\hat{p}_2 = 73 \%$

$\hat{p} = \frac{243 + 365}{800} = 76 \%$

$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{2,5^2 + 1,9^2} = 3,1 \%$

Règle de décision

Rejeter H_0 si $\hat{p}_1 - \hat{p}_2$ est supérieur à 5,1 %.

Décision et conclusion

Comme $\hat{p}_1 - \hat{p}_2 = 8 \% > 5,1 \%$, on rejette H_0 .

Le pourcentage d'élèves qui réussissent l'épreuve obligatoire de mathématique est plus élevé chez les élèves qui lisent une heure et plus par semaine pour le plaisir que chez ceux qui lisent moins.

9. a) $H_0: p_H - p_F = 0$

$H_1: p_H - p_F < 0$

$\hat{p}_H = 37 \%$ et $\hat{p}_F = 46 \%$

$\hat{p} = \frac{222 + 184}{1000} = 40,6 \%$

$\sigma_{\hat{p}_H - \hat{p}_F} = \sqrt{2,0^2 + 2,5^2} = 3,2 \%$

Règle de décision

Rejeter H_0 si $\hat{p}_H - \hat{p}_F$ est inférieur à -7,5 %.

Décision et conclusion

Comme $\hat{p}_H - \hat{p}_F = -9\% < -7,5\%$, on rejette H_0 .
Le taux d'emploi des étudiants est inférieur à celui des étudiantes.

b) $H_0: \mu_H - \mu_F = 0$

$H_1: \mu_H - \mu_F > 0$

$$\sigma_{\bar{x}_H - \bar{x}_F} = \sqrt{0,17^2 + 0,18^2} = 0,25 \text{ h}$$

Règle de décision

Rejeter H_0 si $\bar{x}_H - \bar{x}_F$ est supérieur à 0,58 h.

Décision et conclusion

Comme $\bar{x}_H - \bar{x}_F = 1,1 \text{ h} > 0,58 \text{ h}$, on rejette H_0 .
La moyenne d'heures de travail par semaine des étudiants est supérieure à celle des étudiantes.

Exercices récapitulatifs

1. a) 58 %

b) $H_0: p = 50\%$

$H_1: p > 50\%$

$\alpha = 0,05$

On a $n = 2\,425$ et $\hat{p} = 54\%$.

(Ne pas oublier de représenter graphiquement le test.)

$$\sigma_{\hat{p}} = \sqrt{\frac{50 \times 50}{2\,425}} = 1\%$$

Règle de décision

Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est supérieur à 51,6 %.

Décision et conclusion

Comme $\hat{p} = 54\% > 51,6\%$, on rejette H_0 .
En 2013, plus de 50 % des amateurs de jeux vidéo sont des hommes. Le pourcentage se situe fort probablement autour de 54 %.

c) i) $H_0: \mu = 33 \text{ ans}$

$H_1: \mu < 33 \text{ ans}$

$\alpha = 0,05$

On a $n = 2\,425$; $\bar{x} = 31 \text{ ans}$; $s = 12,5 \text{ ans}$.

(Ne pas oublier de représenter graphiquement le test.)

$$\sigma_{\bar{x}} = \frac{12,5}{\sqrt{2\,425}} = 0,3 \text{ an}$$

Règle de décision

Rejeter H_0 si la moyenne \bar{x} de l'échantillon est inférieure à 32,5 ans.

Décision et conclusion

Comme $\bar{x} = 31 \text{ ans} < 32,5 \text{ ans}$, on rejette H_0 .
En 2013, la moyenne d'âge des joueurs de jeux vidéo est inférieure à celle de 2011. On peut l'estimer ponctuellement à 31 ans.

ii) $H_0: p = 34\%$

$H_1: p < 34\%$

$\alpha = 0,05$

On a $n = 2\,425$ et $\hat{p} = 33\%$.

(Ne pas oublier de représenter graphiquement le test.)

$$\sigma_{\hat{p}} = \sqrt{\frac{34 \times 66}{2\,425}} = 1\%$$

Règle de décision

Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est inférieur à 32,4 %.

Décision et conclusion

Comme $\hat{p} = 33\% > 32,4\%$, on ne rejette pas H_0 .

Les données de l'échantillon ne permettent pas de conclure que le pourcentage de joueurs qui utilisent plus souvent une console de jeu qu'un autre type de plate-forme en 2013 est plus faible qu'en 2010. L'écart observé entre le pourcentage de 33 % de l'échantillon et celui de 34 % de l'étude de 2010 n'est pas significatif statistiquement : il est attribuable au hasard de l'échantillonnage.

iii) $H_0: p = 45\%$

$H_1: p \neq 45\%$

$\alpha = 0,05$

On a $n = 2\,425$ et $\hat{p} = 46\%$.

(Ne pas oublier de représenter graphiquement le test.)

$$\sigma_{\hat{p}} = \sqrt{\frac{45 \times 55}{2\,425}} = 1\%$$

Règle de décision

Rejeter H_0 si le pourcentage \hat{p} de l'échantillon est inférieur à 43 % ou supérieur à 47 %.

Décision et conclusion

Comme $43\% < \hat{p} < 47\%$ (car $\hat{p} = 46\%$), on ne rejette pas H_0 .

Statistiquement, rien ne permet de penser que le pourcentage d'amateurs de jeux vidéo qui y jouent plusieurs jours par semaine est différent de 45 %.

2. a) $H_0: p_{2014} - p_{2013} = 0$

$H_1: p_{2014} - p_{2013} < 0$

$\alpha = 0,05$

$\hat{p}_{2014} = 25,6\%$ et $\hat{p}_{2013} = 27\%$

$$\hat{p} = \frac{216 + 230}{1\,700} = 26,2\%$$

$$\sigma_{\hat{p}_{2014} - \hat{p}_{2013}} = \sqrt{1,6^2 + 1,5^2} = 2,2\%$$

Règle de décision

Rejeter H_0 si $\hat{p}_{2014} - \hat{p}_{2013}$ est inférieur à -3,6 %.

Décision et conclusion

Comme $\hat{P}_{2014} - \hat{P}_{2013} = -1,4\% > -3,6\%$, on ne rejette pas H_0 .

Rien ne permet de conclure que le pourcentage de cyberacheteurs en janvier 2014 est plus faible qu'en janvier 2013, la différence échantillonale de 1,4 % est attribuable au hasard de l'échantillonnage.

b) $H_0 : \mu_{2014} - \mu_{2013} = 0$

$H_1 : \mu_{2014} - \mu_{2013} < 0$

$\alpha = 0,05$

$$\sigma_{\bar{x}_{2014} - \bar{x}_{2013}} = \sqrt{3,2^2 + 3,2^2} = 4,53 \text{ \$}$$

Règle de décision

Rejeter H_0 si $\bar{x}_{2014} - \bar{x}_{2013}$ est inférieur à -7,45 \$.

Décision et conclusion

Comme $\bar{x}_{2014} - \bar{x}_{2013} = -77 \text{ \$} < -7,45 \text{ \$}$, on rejette H_0 .

Le montant moyen des achats des cyberacheteurs en janvier 2014 est inférieur à celui de janvier 2013.

3. $H_0 : \mu_d = 0$

$H_1 : \mu_d < 0$

$\alpha = 0,05$

$\bar{d} = -16$ accidents et $s_d = 18,7$ accidents

$\sigma_{\bar{d}} = 7,1$ accidents et $t_c = 1,943$

Règle de décision

Rejeter H_0 si la moyenne \bar{d} des différences du nombre d'accidents observées dans l'échantillon est inférieure à -13,8 accidents.

Décision et conclusion

Comme $\bar{d} = -16$ accidents < -13,8 accidents, on rejette H_0 . Les radars photo font diminuer le nombre d'accidents.

Chapitre 6

Exercice de compréhension 6.1

a) 1. H_0 : Les consommateurs ne seront pas influencés par la couleur de la bière : ils se répartiront proportionnellement entre les deux couleurs.

H_1 : Les consommateurs seront influencés par la couleur de la bière : ils ne se répartiront pas proportionnellement entre les deux couleurs.

Couleurs	Blonde	Rousse	Total
O	530	270	800
% T	3/5 (60 %)	2/5 (40 %)	5/5 (100 %)
T	480	320	800

La condition d'application est respectée, car on a $T \geq 5$.

3. $\chi^2 = 13,0$

Règle de décision

Pour $dl = 1$ et $\alpha = 0,01$, on a $\chi^2_c = 6,63$.

Rejeter H_0 si $\chi^2 > 6,63$.

Décision et conclusion

Comme $\chi^2 = 13,0 > 6,63$, on rejette H_0 .

Les consommateurs seront influencés par la couleur de la bière. L'analyse des écarts montre que, toutes proportions gardées, les bières blondes se vendront vraisemblablement plus que les bières rouges.

b) 1. H_0 : Les consommateurs se répartiront uniformément entre les trois bières blondes.

H_1 : Les consommateurs ne se répartiront pas uniformément entre les trois bières blondes.

Sortes	B1	B2	B3	Total
O	156	162	212	530
% T	1/3	1/3	1/3	3/3
T	176,7	176,7	176,7	530

La condition d'application est respectée, car on a $T \geq 5$.

3. $\chi^2 = 10,7$

Règle de décision

Pour $dl = 2$ et $\alpha = 0,01$, on a $\chi^2_c = 9,21$.

Rejeter H_0 si $\chi^2 > 9,21$.

Décision et conclusion

Comme $\chi^2 = 10,7 > 9,21$, on rejette H_0 .

Les consommateurs ne se répartiront pas uniformément entre les trois bières blondes.

L'analyse des écarts montre que la bière B3 se vendra vraisemblablement plus que les deux autres.

Exercice de compréhension 6.2

1. H_0 : suit une loi normale.

H_1 : ne suit pas une loi normale.

2. 90 et plus.

Graphique : cote z de 90 = 1,22 ; $P(X > 90) = 0,112$; $P(80 < X < 90) = 0,2520$.

Tableau : [80 ; 90[: % T = 25,2 % ; T = 25,2.
90 et plus : % T = 11,1 % ; T = 11,1.

Oui, car tous les $T \geq 5$.

4. $dl = 5 - 2 - 1 = 2$; $\chi^2_c = 9,21$.

Rejeter H_0 si $\chi^2 > \chi^2_c = 9,21$.

5. Comme $\chi^2 = 17,8 > 9,21$, on rejette H_0 .

La distribution des notes à l'examen d'anglais de l'ensemble des élèves de la population ne suit pas un modèle normal.

Exercices 6.1

1. a) On ne peut pas identifier la roulette truquée, car on ne connaît pas la répartition de la surface selon les couleurs.

b) i) La roulette 1.

ii) Sur la distribution attendue de 100 lancers d'une roulette non truquée. En effet, si les résultats s'éloignent beaucoup de la répartition « 60 % rouge, 30 % verte et 10 % noire » (soit la

probabilité d'obtenir chacune des couleurs), on ne peut pas les attribuer entièrement au hasard; donc la roulette est truquée.

2. a) On ne peut pas déterminer les effectifs théoriques sous cette hypothèse nulle : y aura-t-il 60 % de faces ? 80 % ? 15 % ? Comme on ne sait pas comment la pièce a été truquée, on ne peut pas prévoir son comportement.
 - b) Sous cette hypothèse nulle, il est facile de calculer les effectifs théoriques puisque, si la pièce n'est pas truquée, théoriquement, on obtiendra face à 50 % des lancers. Les effectifs théoriques sont : face : 500 ; pile : 500.
 - c) La situation 2, car il est difficile de décider si les écarts entre les résultats obtenus et ceux attendus, soit 500, sont trop grands pour être attribués au hasard. Par contre, pour la situation 1, les écarts sont sans doute attribuables au hasard, alors que pour la situation 3, avec 72 % de faces, la pièce est vraisemblablement truquée.
3. a) H_0 : Les préférences des clients se répartissent uniformément entre les trois saveurs de crème glacée.
b) Vanille : 15 ; Chocolat : 15 ; Citron : 15.

4. H_0 : Les naissances se répartissent uniformément entre les quatre phases de la lune.

H_1 : Les naissances ne se répartissent pas uniformément entre les quatre phases de la lune.

Règle de décision

Rejeter H_0 si $\chi^2 > 7,81$.

Décision et conclusion

Comme $\chi^2 = 3,73 < 7,81$, on ne rejette pas H_0 .

Les données échantillonnelles ne permettent pas d'affirmer que la lune exerce une influence sur la répartition des naissances. Au seuil de signification de 5%, l'écart entre les effectifs observés et théoriques n'est statistiquement pas significatif; il doit être attribué au hasard.

5. a) H_0 : Les traversées se répartissent uniformément entre les 5 périodes de la journée.
 H_1 : Les traversées ne se répartissent pas uniformément entre les 5 périodes de la journée.

$$T = 21$$

Règle de décision

Rejeter H_0 si $\chi^2 > 13,3$.

Décision et conclusion

Comme $\chi^2 = 17,8 > 13,3$, on rejette H_0 .

Les traversées ne se répartissent pas uniformément. L'analyse des écarts montre que, toutes proportions gardées, il y a vraisemblablement plus de traversées pour les 2 périodes de la journée comprises entre 10 h et 18 h. (Dans l'étude, on observe que c'est aussi dans cet intervalle de temps que le trafic est le plus intense.)

- b) H_0 : Les traversées se répartissent proportionnellement entre les périodes de jour et de nuit, soit 3/5 et 2/5.

H_1 : Les traversées ne se répartissent pas proportionnellement entre les périodes de jour et de nuit.

Périodes de jour : $O : 78$; $T : 63$.

Périodes de nuit : $O : 27$; $T : 42$.

Règle de décision

Rejeter H_0 si $\chi^2 > 6,63$.

Décision et conclusion

Comme $\chi^2 = 8,9 > 6,63$, on rejette H_0 .

Les traversées ne se répartissent pas proportionnellement entre les périodes de jour et de nuit. L'analyse des écarts montre qu'il y a vraisemblablement moins de traversées de la route par un caribou la nuit.

- c) H_0 : Les traversées se répartissent uniformément entre les 3 périodes de jour.

H_1 : Les traversées ne se répartissent pas uniformément entre les 3 périodes de jour.

$O : 16$	35	27
$T : 26$	26	26

Règle de décision

Rejeter H_0 si $\chi^2 > 9,21$.

Décision et conclusion

Comme $\chi^2 = 7 < 9,21$, on ne rejette pas H_0 .

Il est vraisemblable que les traversées se répartissent uniformément entre les 3 périodes de la journée. Au seuil de signification de 0,01, la différence entre les effectifs observés et théoriques est attribuable aux fluctuations d'échantillonnage.

6. a) H_0 : L'échantillon est représentatif de la population des 18 à 24 ans pour le revenu gagné.

H_1 : L'échantillon n'est pas représentatif de la population des 18 à 24 ans pour le revenu gagné.

- b) 0 \$: 31,2
1 \$ à 24 999 \$: 617,6
25 000 \$ à 49 999 \$: 128,8
50 000 \$ et plus : 22,4

7. H_0 : L'échantillon est représentatif de la population universitaire pour le cycle d'études.

H_1 : L'échantillon n'est pas représentatif de la population universitaire pour le cycle d'études.

$O : 942$	198	60
$T : 922,8$	216	$61,2$

Règle de décision

Rejeter H_0 si $\chi^2 > 5,99$.

Décision et conclusion

Comme $\chi^2 = 1,92 < 5,99$, on ne rejette pas H_0 .

L'échantillon est vraisemblablement représentatif de la population universitaire pour le cycle d'études.

1

2

3

4

5

6

7

8

8. H_0 : La distribution du nombre de filles dans une famille de 4 enfants suit la loi binomiale $B(4; 0,5)$.
 H_1 : La distribution du nombre de filles dans une famille de 4 enfants ne suit pas la loi binomiale $B(4; 0,5)$.
 T : 62,5 250 375 250 62,5

Règle de décisionRejeter H_0 si $\chi^2 > 13,3$.Décision et conclusionComme $\chi^2 = 10,3 < 13,3$, on ne rejette pas H_0 .

Les données échantillonnelles ne permettent pas de rejeter l'affirmation voulant que la distribution du nombre de filles dans les familles de 4 enfants suive la loi binomiale $B(4; 0,5)$.

9. H_0 : La distribution du nombre de défauts par tee-shirt suit la loi de Poisson $Po(0,5)$.
 H_1 : La distribution du nombre de défauts par tee-shirt ne suit pas la loi de Poisson $Po(0,5)$.

T : 545,9 273 68,2 11,3 1,6

Note: La catégorie «4 et plus» a un effectif théorique inférieur à 5 : il faut regrouper les deux dernières catégories avant de calculer le khi-deux. La dernière catégorie devient ainsi «3 et plus», et $dl = 3$.

Règle de décisionRejeter H_0 si $\chi^2 > 7,81$.Décision et conclusionComme $\chi^2 = 5,39 < 7,81$, on ne rejette pas H_0 .

L'hypothèse voulant que la distribution du nombre de défauts par tee-shirt suive la loi de Poisson $Po(0,5)$ est plausible.

10. H_0 : L'échantillon est représentatif des Montréalais de 70 ans et plus pour la variable «âge».
 H_1 : L'échantillon n'est pas représentatif des Montréalais de 70 ans et plus pour la variable «âge».

Âge (en ans)	De 70 à 74	De 75 à 79	De 80 à 84	85 et plus
O	343	352	232	273
% T	30,5 %	26,9 %	22,0 %	20,6 %
T	366,0	322,8	264,0	247,2

Règle de décisionRejeter H_0 si $\chi^2 > 7,81$.Décision et conclusionComme $\chi^2 = 10,7 > 7,81$, on rejette H_0 .

L'échantillon n'est pas représentatif des Montréalais de 70 ans et plus pour la variable «âge».

11. H_0 : La distribution de la quantité annuelle de papier utilisé par l'ensemble des employés suit une loi normale.
 H_1 : La distribution de la quantité annuelle de papier utilisé par l'ensemble des employés ne suit pas une loi normale.

$40 \leq X < 50 : T = 51,5$

$50 \leq X < 60 : T = 59,9$

$X \geq 70 : T = 9,7$

Note: Il faut regrouper les deux 1^{res} classes, car dans la 1^{re} classe, on a $T < 5$.

$dl = 2$

Règle de décisionRejeter H_0 si $\chi^2 > 5,99$.Décision et conclusionComme $\chi^2 = 0,74 < 5,99$, on ne rejette pas H_0 .

La distribution de la quantité annuelle de papier utilisé par l'ensemble des employés suit vraisemblablement une loi normale.

12. H_0 : La distribution du diamètre des tiges de la production suit une loi normale.

H_1 : La distribution du diamètre des tiges de la production ne suit pas une loi normale.

$1,12 \leq X < 1,15 : T = 20,8$

$1,18 \leq X < 1,21 : T = 46,9$

$X \geq 1,27 : T = 16,2$

$dl = 4$

Règle de décisionRejeter H_0 si $\chi^2 > 9,49$.Décision et conclusionComme $\chi^2 = 6,47 < 9,49$, on ne rejette pas H_0 .

La distribution du diamètre des tiges de la production suit vraisemblablement une loi normale.

Exercice de compréhension 6.3

- a) $247 \div 800 = 30,9\%$

- b) 1. H_0 : Chez les lecteurs de livres, il n'y a pas de lien entre le sexe et le nombre de livres lus en un an.

H_1 : Chez les lecteurs de livres, il y a un lien entre le sexe et le nombre de livres lus en un an.

2.

O T	Nombre de livres lus				Total
	De 1 à 4	De 5 à 9	De 10 à 19	20 et plus	
Sexe					
Femmes	143 164,4	105 104,3	134 125,0	150 138,3	532
Hommes	104 82,8	52 52,5	54 63,0	58 69,7	268
Total	247	157	188	208	800
(% T)	(30,9 %)	(19,6 %)	(23,5 %)	(26,0 %)	

Oui, car $T \geq 5$.

3. $\chi^2 = 13,1$

4. Rejeter H_0 si $\chi^2 > 11,3$.

5. Comme $\chi^2 = 13,1 > 11,3$, on rejette H_0 .

Chez les lecteurs de livres, il existe vraisemblablement un lien entre le sexe et le nombre de livres lus en un an.

Nature de la dépendance

50 % ; 53 % ; 42 %.

Exercices 6.2

1. a) 38,9 % ($14 \div 36$)
 - b) i) 6,2 parties, soit 38,9 % des 16 parties jouées à domicile.
 - ii) 7,8 parties, soit 38,9 % des 20 parties jouées à l'extérieur.
 - c) i) Il est impossible de répondre à cette question puisqu'on ne connaît pas la nature du lien entre les deux variables. On sait que le nombre de parties gagnées à domicile n'est pas égal à 38,9 % des parties jouées à domicile, mais de combien est-il? de 45 %? de 34 %?
 - ii) Il est impossible de trouver les effectifs théoriques, car on ne peut pas déterminer le pourcentage théorique sous cette hypothèse.
 2. a) H_0 : Les variables «âge» et «lecture de quotidiens sur Internet» sont indépendantes.
 H_1 : Les variables «âge» et «lecture de quotidiens sur Internet» sont dépendantes.
- | <i>O T</i> | Lecture de quotidiens sur Internet | | |
|---------------------|------------------------------------|--------------|-------|
| | Oui | Non | Total |
| Âge (en ans) | | | |
| Moins de 34 | 27 20,9 | 16 22,1 | 43 |
| De 35 à 54 | 42 36,4 | 33 38,6 | 75 |
| 55 et plus | 28 39,8 | 54 42,2 | 82 |
| Total (% T) | 97 (48,5 %) | 103 (51,5 %) | 200 |
- $\chi^2 = 11,9$
- Règle de décision
 Rejeter H_0 si $\chi^2 > 9,21$.
- Décision et conclusion
 Comme $11,9 > 9,21$, on rejette H_0 .
 Les variables «âge» et «lecture de quotidiens sur Internet» sont dépendantes.
- Nature de la dépendance*
 Globalement, si l'on ne tient pas compte de l'âge, près de 49 % des lecteurs de quotidiens préfèrent la version électronique du journal à la version papier. Or, si l'on tient compte de l'âge, ce pourcentage augmente à 63 % chez les moins de 34 ans et à 56 % chez les 35 à 54 ans. À l'inverse, il diminue à 34 % chez les 55 ans et plus.
- Une tendance se dégage : plus les lecteurs sont jeunes, plus la proportion de ceux qui préfèrent utiliser Internet pour lire des quotidiens est élevée.
- b) 48,5 %; non, elle serait aussi de 48,5 %.
3. a) Non, le fait qu'un des effectifs observés soit inférieur à 5 ne remet pas la validité du test en cause. Ce sont les effectifs théoriques qui doivent être supérieurs ou égaux à 5. Pour cette catégorie, l'effectif théorique est 5,3.
 - b) H_0 : Il n'y a pas de lien entre le niveau de scolarité et la lecture de quotidiens sur Internet.
 H_1 : Il y a un lien entre le niveau de scolarité et la lecture de quotidiens sur Internet.

<i>O T</i>	Lecture de quotidiens sur Internet		
	Oui	Non	Total
Niveau de scolarité			
Primaire	3 5,3	8 5,7	11
Secondaire	23 32,0	43 34,0	66
Collégial	29 27,6	28 29,4	57
Universitaire	42 32,0	24 34,0	66
Total (% T)	97 (48,5 %)	103 (51,5 %)	200

$$\chi^2 = 13,0$$

Règle de décision

Rejeter H_0 si $\chi^2 > 11,3$.

Décision et conclusion

Comme $13,0 > 11,3$, on rejette H_0 .

Il y a un lien entre le niveau de scolarité et la lecture de quotidiens sur Internet.

Nature de la dépendance

Globalement, si l'on ne tient pas compte de la scolarité, près de 49 % des lecteurs de quotidiens préfèrent la version électronique du journal à la version papier. Or, si l'on tient compte de la scolarité, ce pourcentage est plus bas chez ceux qui ont une scolarité de niveau primaire ou secondaire, soit 27 % et 35 % respectivement, et inversement, il est plus élevé chez ceux qui ont une scolarité de niveau collégial ou universitaire, soit 51 % et 64 % respectivement.

Une tendance se dégage : plus le niveau de scolarité est élevé, plus le pourcentage de lecteurs qui utilisent Internet pour lire des quotidiens est élevé.

4. a)

b)

5. a) i) Comme $T=5,6 \geq 5$, la condition d'application du test du khi-deux est respectée. (Le fait que l'effectif observé soit inférieur à 5 n'a pas d'importance.)
 ii) Comme $T=3,3 < 5$, la condition d'application du test du khi-deux n'est pas respectée.
 Note : Si l'on fait quand même le test, tout en sachant que la condition d'application du test du khi-deux n'est pas respectée, la conclusion ne sera pas valide statistiquement.
- b) H_0 : Le temps de course ne dépend pas de l'âge du coureur.
 H_1 : Le temps de course dépend de l'âge du coureur.

1

2

3

4

5

6

7

8

<i>O T</i>	Temps de course (en min)				
	Âge (en ans)	Moins de 75	75; 90	90 et plus	Total
Moins de 30	7 6,0	12 9,7	6 9,4	25	
De 30 à 39	8 5,2	7 8,5	7 8,3	22	
40 et plus	4 7,9	12 12,8	17 12,4	33	
Total (% T)	19 (23,8 %)	31 (38,8 %)	30 (37,5 %)	80	

$$\chi^2 = 7,6$$

Règle de décision

Rejeter H_0 si $\chi^2 > 9,49$.

Décision et conclusion

Comme $7,6 < 9,49$, on ne rejette pas H_0 .

Les données échantillonnelles ne permettent pas de conclure qu'il existe un lien entre l'âge et le temps de course au seuil de signification de 5 %. Les écarts entre les effectifs théoriques et observés ne sont pas statistiquement significatifs; ils sont vraisemblablement attribuables au hasard de l'échantillonnage.

6. a) H_0 : L'attitude face à l'avortement ne dépend pas de la scolarité.

H_1 : L'attitude face à l'avortement dépend de la scolarité.

$$\chi^2 = 17,7$$

Règle de décision

Rejeter H_0 si $\chi^2 > 9,49$.

Décision et conclusion

Puisque $17,7 > 9,49$, on rejette H_0 .

L'attitude face à l'avortement dépend de la scolarité.

Nature de la dépendance

Globalement, si l'on ne tient pas compte de la scolarité, 41 % des répondants sont favorables à l'avortement. Or, lorsque l'on tient compte de la scolarité, ce pourcentage diminue à 28 % chez les gens qui ont moins de 9 années de scolarité et à 39 % chez ceux qui ont de 9 à 12 années de scolarité. À l'inverse, il augmente à 51 % chez les gens qui ont plus de 12 années de scolarité.

Une tendance se dégage : l'appui à l'avortement augmente avec le nombre d'années de scolarité.

- b) Pour les catholiques : on ne rejette pas H_0 .

Pour les protestants : on rejette H_0 .

L'attitude face à l'avortement ne dépend de la scolarité que chez les protestants.

Exercices récapitulatifs

1. a) H_0 : L'échantillon est représentatif de la population pour la nature des dommages.

H_1 : L'échantillon n'est pas représentatif de la population pour la nature des dommages.

Règle de décision

Rejeter H_0 si $\chi^2 > 3,84$.

Décision et conclusion

Comme $\chi^2 = 1,28 < 3,84$, on ne rejette pas H_0 .

Au seuil de signification de 0,05, il n'y a aucune évidence statistique permettant de remettre en cause la représentativité de l'échantillon pour la nature des dommages lors de l'accident.

- b) i) H_0 : Les accidents se répartissent proportionnellement entre la semaine et la fin de semaine.

H_1 : Les accidents ne se répartissent pas proportionnellement entre la semaine et la fin de semaine.

Semaine: $O: 599 \quad T: 571,4$

Fin de semaine: $O: 201 \quad T: 228,6$

Règle de décision

Rejeter H_0 si $\chi^2 > 3,84$.

Décision et conclusion

Comme $\chi^2 = 4,67 > 3,84$, on rejette H_0 .

Les accidents ne se répartissent pas proportionnellement entre la semaine et la fin de semaine. L'analyse des écarts montre que, toutes proportions gardées, il y a vraisemblablement moins d'accidents la fin de semaine.

- ii) H_0 : Les accidents se répartissent uniformément du lundi au vendredi.

H_1 : Les accidents ne se répartissent pas uniformément du lundi au vendredi.

Effectifs théoriques: 119,8

Règle de décision

Rejeter H_0 si $\chi^2 > 9,49$.

Décision et conclusion

Comme $\chi^2 = 6,32 < 9,49$, on ne rejette pas H_0 .

Au seuil de signification de 5 %, il n'y a aucune évidence statistique qui permet de rejeter l'hypothèse voulant que les accidents se répartissent uniformément du lundi au vendredi.

2. H_0 : Le taux de branchement à Internet très haute vitesse dépend de la taille de l'entreprise.

H_1 : Le taux de branchement à Internet très haute vitesse ne dépend pas de la taille de l'entreprise.

<i>O T</i>	Connexion très haute vitesse			
	Nombre d'employés	Oui	Non	Total
De 1 à 9	39 54,3	311 295,8		350
De 10 à 49	40 49,6	280 270,4		320
De 50 à 249	39 32,6	171 177,5		210
250 et plus	35 17,1	75 93,0		110
Total (% T)	153 (15,5 %)	837 (84,5 %)	990	

Règle de décision

Rejeter H_0 si $\chi^2 > 11,3$.

Décision et conclusion

Comme $\chi^2 = 31 > 11,3$, on rejette H_0 .

Le taux de branchement à Internet très haute vitesse dépend de la taille de l'entreprise.

Nature de la dépendance

Globalement, 15,5 % des entreprises branchées à Internet ont une connexion très haute vitesse. Or, lorsque l'on considère la taille de l'entreprise, ce taux est plus bas pour les entreprises de 1 à 9 employés et de 10 à 49 employés, soit 11 % et 13 % respectivement. Inversement, il est plus élevé pour les entreprises de 50 à 249 employés et de 250 employés et plus, soit 19 % et 32 % respectivement.

Une tendance se dégage : plus le nombre d'employés d'une entreprise est élevé, plus le taux de branchement à Internet très haute vitesse est élevé.

- H_0 : La distribution de l'âge dans la population des entrepreneurs suit une loi normale.
- i) $285,0 [(0,420 \cdot 7 - 0,135 \cdot 7) \times 1\ 000]$
ii) $337,3 [(1 - 0,242 \cdot 0 - 0,420 \cdot 7) \times 1\ 000]$
- Rejeter H_0 si $\chi^2 > 11,3$ ($dl = 3$).
- Comme $10,4 < 11,3$, on ne rejette pas H_0 .

La distribution de l'âge des entrepreneurs suit vraisemblablement une loi normale.

Chapitre 7

Exercices de compréhension 7.1

- Y : Consommation de mazout (en litres);
 X : Température extérieure moyenne (en Celsius); Négative.
b) Forte ; plus ; moins.
c) 0,90 ; 90 ; 10.
- a) $\sum xy = 201\ 135$; $\bar{x} = 3,5$; $s_x = 1,87$; $\bar{y} = 9\ 405,33$; $n = 6$.
 $b = \frac{201\ 135 - 6 \times 3,5 \times 9\ 405,33}{5 \times 1,87^2} = 207,22$
($b = 207,03$ avec le mode statistique de la calculatrice.)
 $a = 9\ 405,33 - (207,22 \times 3,5) = 8\ 680,06$
($a = 8\ 680,73$ avec le mode statistique de la calculatrice.)
• Droite de tendance : $y = 8\ 680,1 + 207,2x$
b) Pour $x = 10$, $y = 10\ 752 \$$.

Exercices 7.1

- Positive. b) Nulle.
c) Positive. Généralement, les couples sont formés de personnes dans les mêmes groupes d'âge. La corrélation mesure la tendance «générale» des variables à varier dans le même sens ou dans le sens contraire.
- a) $r = 1$ ou $r = -1$: les deux points sont nécessairement situés sur une même droite.
b) Y : Épaisseur de la glace couvrant un lac
 X : Température extérieure
- a) Avec un coefficient de corrélation linéaire de 0,94, on peut dire qu'il existe un lien très fort entre l'âge auquel un ex-fumeur a cessé de fumer et le taux de mortalité due à un cancer du poumon. Le fait que la corrélation est positive indique la tendance suivante : plus un

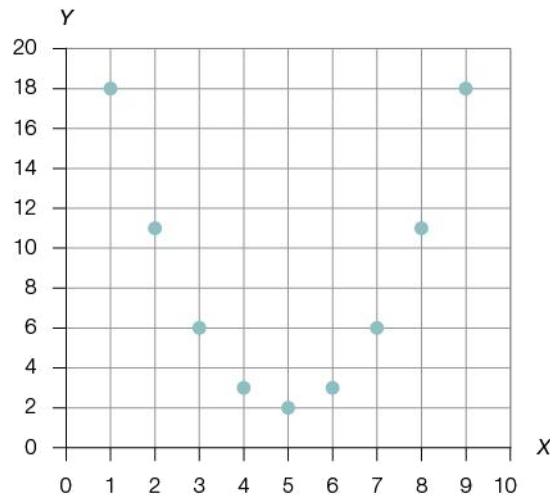
fumeur est âgé au moment où il cesse de fumer, plus le risque de mourir du cancer du poumon est élevé.

b) $r^2 = 0,88$

Interprétation

On peut estimer que l'âge auquel un fumeur cesse de fumer explique 88 % de la variation observée dans le taux de mortalité due à un cancer du poumon. Par conséquent, 12 % de cette variation peut être attribuable à d'autres facteurs.

- $y = 5\ 501,2 + 870,1x$
 $(\sum xy = 81\ 115; \bar{x} = 2; s_x = 1,58; \bar{y} = 7\ 241,4)$
- 11 592 emplois
- 3 761 emplois
- a est le nombre d'emplois estimé pour 2008 ($x = 0$).
b est l'augmentation moyenne du nombre d'emplois par année.
- a) $r = 0$
- Le fait que $r = 0$ nous permet uniquement de déclarer qu'il n'y a pas de dépendance linéaire entre les variables, mais il peut y avoir une dépendance non linéaire.
- Le nuage de points obtenu montre qu'il existe une corrélation non linéaire parfaite entre les deux variables. Celles-ci sont donc dépendantes.



- $r = -0,98$ avec le mode statistique de la calculatrice.

$$(\sum xy = 49\ 070; \bar{x} = 34,6; s_x = 4,77;$$

$$\bar{y} = 296; s_y = 114,15.)$$

Interprétation

La corrélation entre le prix des casquettes et le volume des ventes est très forte. Le fait que la corrélation est négative indique la tendance suivante : plus on augmente le prix des casquettes, plus le volume des ventes diminue.

b) $r^2 = 0,96$

Interprétation

Le prix des casquettes explique 96 % de la variation du volume des ventes. Il y a 4 % de la variation du volume des ventes qui est attribuable à d'autres facteurs.

c) $y = 1\ 107,1 - 23,4x$

d) Environ 124 casquettes.

e) 34 \$

7. a) $r = 1$ avec le mode statistique de la calculatrice.

$$\left(\sum xy = 631\ 480; \bar{x} = 1258; s_x = 394,74; \right.$$

$$\left. \bar{y} = 94,8; s_y = 22,29. \right)$$

Interprétation

La corrélation entre le nombre de kilowattheures consommés et le coût est parfaite. Le fait que la corrélation est positive indique la tendance suivante : plus le nombre de kilowattheures consommés est élevé, plus le coût est élevé.

b) Les points du diagramme de dispersion seraient parfaitement alignés.

c) $r^2 = 1$

Interprétation

Le nombre de kilowattheures consommés explique 100 % de la variation du montant de la facture d'électricité.

d) $y = 23,78 + 0,056x$

e) 102,18 \$

f) 1 147 kWh

g) Cette équation donne la structure de tarification d'Hydro-Québec pour la consommation domestique.

- $a = 23,78$ \$ est le montant fixe à payer pour le service de base avant toute consommation pour 60 jours, soit 0,396 \$ par jour.
- $b = 0,056$ \$ est le coût de l'électricité par kilowattheure. À titre d'information, au-delà des 30 premiers kilowattheures consommés par jour, le coût passe à 0,083 \$/kWh.

8. a) $y = 4,7 + 5,55x$

b) 71,3 %

c) Non. L'année 2018 se situe à 8 ans de la dernière mesure prise, ce qui est beaucoup trop éloigné des années utilisées pour construire la droite de régression : il serait surprenant que la tendance observée de 2002 à 2010 se maintienne jusqu'en 2018. Si l'on remplace x par 18 dans l'équation de la droite de régression, on obtient un taux de récupération de 104,6 % : ce résultat surprenant confirme que le modèle mathématique construit n'est plus valable en 2018.

Exercice récapitulatif

a) X : Distance

Y : Coût

b) $r = 0,99$ avec le mode statistique de la calculatrice.

$$\left(\sum xy = 1851,14; \bar{x} = 11,2; s_x = 2,67; \bar{y} = 22,69; \right.$$

$$\left. s_y = 4,58. \right)$$

Interprétation

La corrélation entre le coût de la course et la distance parcourue est presque parfaite.

Le fait qu'elle est positive indique la tendance suivante : plus la distance à parcourir est grande, plus le coût de la course est élevé.

c) $r^2 = 0,98$

Interprétation

La distance parcourue explique 98 % de la variation du coût d'une course en taxi. Il y a 2 % de la variation du coût qui est attribuable à d'autres facteurs. En fait, le coût dépend aussi du temps d'immobilisation du taxi pendant la course (feux rouges, arrêts obligatoires, etc.), qui est tarifié à 0,63 \$ par minute d'attente.

d) $y = 3,64 + 1,70x$

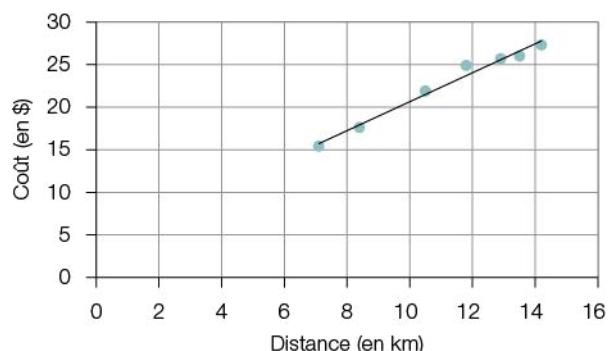
- a correspond au montant du coût de la course qui est indépendant de la distance parcourue : il comprend le tarif de départ, fixé à 3,45 \$, auquel s'ajoute 0,63 \$ par minute d'immobilisation du taxi durant le parcours.

- b est le coût par kilomètre parcouru, fixé à 1,70 \$.

f) 30,84 \$

g) 15,5 km

Coût d'une course en taxi en fonction de la distance parcourue



Chapitre 8

Exercices 8.1

1. a) Le 1^{er} graphique.

b) i) Vrai

ii) Faux. Les femmes syndiquées gagnent plus que les hommes non syndiqués.

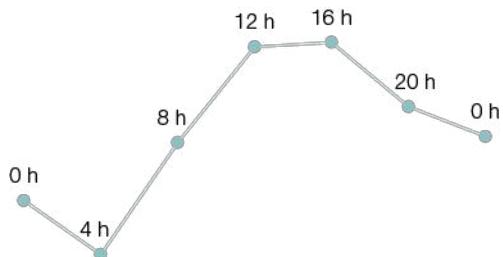
iii) Vrai

c) i) Vrai

ii) Vrai

iii) Faux. Pour les personnes sans études secondaires et pour les diplômés universitaires, c'est l'inverse.

2. a) Forme graphique de la saisonnalité



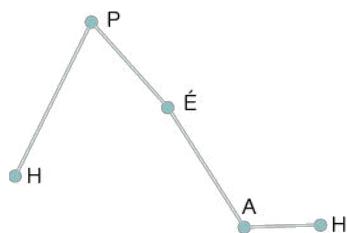
Cycle de la saisonnalité et nombre de périodes

La saisonnalité est quotidienne. Comme on note la température six fois par jour, le cycle saisonnier est de six périodes.

Variation saisonnière de la variable

Généralement, en début de journée, la température est à la baisse jusqu'aux environs de 4 h, soit au moment où elle atteint son minimum. Ensuite, la température augmente rapidement jusqu'à 12 h, et plus doucement jusqu'à 16 h. Par la suite, elle chute d'heure en heure.

b) Forme graphique de la saisonnalité



Cycle de la saisonnalité et nombre de périodes

La saisonnalité est annuelle. Le cycle saisonnier est de quatre périodes.

Variation saisonnière de la variable

Généralement, c'est au trimestre du printemps que le volume des ventes de voitures neuves atteint un maximum. Par la suite, le volume des ventes diminue pour atteindre son plus bas niveau aux trimestres d'automne et d'hiver.

3. a) La série présente une tendance curviligne croissante.

b) On observe trois cycles:

- 1^{er} cycle : de 1976 à 1983 ;
- 2^e cycle : de 1983 à 1994 ;
- 3^e cycle : de 1994 à 2010.

4. a) Hiver 2010 : série 3 : 8,5.

Printemps 2011 : série 2 : 10,8 ; série 3 : 10,7.

Hiver 2012 : série 2 : 11,7 ; série 3 : 11,0.

Été 2013 : série 2 : 10,8 ; série 3 : 10,2.

b) Courbe noire : série 1.

Courbe bleue : série 2.

Courbe orange : série 3.

c) Le lissage par moyenne mobile pondérée (série 2).

5. a) Indice de saisonnalité : automne : 1,174 ; hiver : 0,904.

Données désaisonnalisées :

4. 10,1	8. 11,6	12. 11,2	16. 11,0
5. 10,5	9. 11,5	13. 10,5	17. 12,1

- b) De 2010 à 2014, le volume des ventes du trimestre d'été est approximativement égal à 79,6 % de la moyenne trimestrielle annuelle des ventes, soit 20,4 % de moins que celle-ci.

- c) Les données désaisonnalisées indiquent une tendance à la hausse de la moyenne trimestrielle annuelle des ventes : elle est passée de 10 800 \$ à 11 100 \$.

- d) On observe une première baisse des ventes en 2012, au trimestre d'automne : les données désaisonnalisées indiquent que la moyenne trimestrielle des ventes a été approximativement de 112 000 \$ alors qu'elle était de 122 000 \$ au trimestre précédent. Cette baisse s'est poursuivie jusqu'au trimestre d'été de 2013, où l'on a observé pour la première fois une augmentation de la moyenne trimestrielle des ventes.

Exercices récapitulatifs

1. a) Moyenne mobile pondérée, trimestre d'été 2011 : 13,638.

$$\frac{\text{Données brutes}}{\text{Données lissées}} = 0,917$$

Indice de saisonnalité 2008-2013, trimestre d'été : 0,937.

Taux désaisonnalisé, trimestre d'été :

2008 : 11,4 2010 : 13,9 2012 : 13,6
2009 : 16,5 2011 : 13,3

- b) Le plus élevé : 15,9 % pour les trimestres d'hiver 2009, 2010 et 2012.

Le plus faible : 10,7 % pour le trimestre d'été 2008.

- c) 4

2. a) Non. La moyenne trimestrielle augmente : le taux désaisonnalisé passe de 12,8 % à 13,0 %.

- b) Non. La moyenne trimestrielle diminue : le taux désaisonnalisé passe de 15,3 % à 14,6 %.

- c) L'indice de saisonnalité du trimestre d'hiver est 1,090 : généralement, chez les Québécois de 15 à 24 ans, le taux de chômage du trimestre d'hiver correspond à environ 109 % de la moyenne trimestrielle annuelle du taux de chômage, soit 9 % de plus que celle-ci.

- d) 13,9

- e) Hiver : 12,6 Été : 12,169
Printemps : 12,996 Automne : 11,928

3. a) On observe une tendance curviligne légèrement décroissante.

b) Forme graphique de la saisonnalité



Cycle de la saisonnalité et nombre de périodes

On observe une saisonnalité annuelle sur un cycle de quatre périodes.

Variation saisonnière de la variable

Généralement, c'est au trimestre d'hiver que le taux de chômage est le plus élevé et au trimestre d'été ou d'automne qu'il est le plus bas.

4. 1^{er} cycle : de 1979 à 1990 ;

- 2^e cycle : de 1990 à 2000 ;

- 3^e cycle : de 2000 à 2007.

BIBLIOGRAPHIE

- ALALOUF, S., D. LABELLE et J. MÉNARD.
Introduction à la statistique appliquée, 2^e éd.,
Montréal, Addison-Wesley, 1990.
- ALLARD, J. *Concepts fondamentaux de la statistique*,
Montréal, Addison-Wesley, 1992.
- BAILLARGEON, G. *Méthodes statistiques avec
applications dans différents secteurs de l'entreprise*,
Trois-Rivières, Éditions SMG, 1984.
- BAILLARGEON, G., et L. MARTIN. *Statistique
appliquée à la psychologie*, 2^e éd., Trois-Rivières,
Éditions SMG, 1989.
- BÉLISLE, J.-P., et J. DESROSIERS. *Introduction à la
statistique*, Boucherville, Gaëtan Morin éditeur,
1983.
- D'ASTOUS, A. *Le projet de recherche en marketing*,
Montréal, Chenelière-McGraw-Hill, 1995.
- LAPIN, L. *Statistique de gestion*, Laval, Études vivantes,
1987.
- MARTINEAU, G. *Statistique non paramétrique appliquée
aux sciences humaines*, Montréal, Éditions Sciences
et culture, 1990.
- SATIN, A., et W. SHAstry. *L'échantillonnage : un guide
non mathématique*, 2^e éd., n° 12-602-XPF au
catalogue, Statistique Canada.

Adresses électroniques de données statistiques

- Association canadienne du logiciel de divertissement:
theesa.ca
- Association des centres jeunesse du Québec:
www.acjq.qc.ca
- Bureau d'assurance du Canada:
www.ibc.ca
- CEFARIO:
www.cefrio.qc.ca
- Centre d'études sur les médias:
www.cem.ulaval.ca
- Collège des médecins du Québec:
www.cmq.org
- Commission des transports du Québec:
www.ctq.gouv.qc.ca
- Conseil québécois du théâtre:
www.cqt.ca
- Éduc'alcool:
educalcool.qc.ca
- Fondation de l'entrepreneurship:
www.entrepreneurship.qc.ca

- HabiloMédias:
habilomedias.ca
- Hydro-Québec:
www.hydroquebec.com
- Industrie Canada:
www.ic.gc.ca
- Infopresse:
www.infopresse.com
- Institut de la statistique du Québec:
www.stat.gouv.qc.ca
- Le site officiel des Canadiens de Montréal:
canadiens.com
- Loto-Québec:
lotoquebec.com
- Ministère de l'Économie, de l'Innovation et des Exportations:
www.mdeie.gouv.qc.ca
- Ministère de l'Éducation, du Loisir et du Sport:
www.mels.gouv.qc.ca
- Ministère de l'Enseignement supérieur:
www.mesrst.gouv.qc.ca
- Ministère de la Culture et des Communications:
www.mcc.gouv.qc.ca
- Ministère de la Famille:
www.mfa.gouv.qc.ca
- Ministère des Ressources naturelles et de la Faune
du Québec:
www.mern.gouv.qc.ca
- Recyc-Québec:
www.recyc-quebec.gouv.qc.ca
- Revenu Québec:
www.revenuquebec.ca
- Secrétariat à l'adoption internationale du Québec:
adoption.gouv.qc.ca
- Service de protection contre l'incendie de la
Ville de Québec (SPCIQ):
www.ville.quebec.qc.ca/incendie
- Société de l'assurance automobile du Québec:
www.saaq.gouv.qc.ca
- Statistique Canada:
www.statcan.gc.ca
- TECHNOCompétences:
www.technocompetences.qc.ca
- Tourisme Québec:
www.tourisme.gouv.qc.ca
- Vélo Québec:
www.velo.qc.ca

INDEX

A

Addition, règle d', 95, 126
Aire sous la courbe normale, 167
Amplitude d'une classe, 22
Analyse de régression, 305
Approximation de la loi binomiale
 par une loi de Poisson, 160
 par une loi normale, 175
Arrangement, 120, 127

B

Bimodalité, 48
Binomiale (v. *Loi binomiale*)
Bornes de l'intervalle, 205

C

Calcul
 de la moyenne
 avec les données brutes, 39
 avec les effectifs, 40
 avec les pourcentages, 41
 du khi-deux, 272, 289, 296
Calculatrice
 méthodes pour générer des nombres aléatoires, 187
 utilisation des touches de mémoire pour calculer le khi-deux, 277
 utilisation du mode statistique de la, 70
 (deux variables), 309

Caractéristiques des lois du khi-deux, 274

Catégories, 3, 4

Causalité, 290, 304

Centiles, 57, 82

Chronogramme, 316

Classe(s), 21

 amplitude d'une, 22
 démarche pour construire des, 22, 24
 données groupées en, 53, 82
 données non groupées en, 50, 51, 82
 limite inférieure d'une, 22
 limite supérieure d'une, 22
 modale, 47
 notation des, 22
 ouverte, 28

Coefficient

 de détermination, 307, 313
 de Pearson, 302
 de variation, 68, 82

Combinaison, 121, 127

Composantes d'une série chronologique, 318

Condition d'application d'un test du khi-deux, 272, 289

Construction d'un test

 d'ajustement du khi-deux, 271, 296
 d'hypothèse
 sur une moyenne, 246, 264
 sur un pourcentage, 252, 264
 d'indépendance du khi-deux, 288, 296

Contexte d'une expérience aléatoire

 binomiale, 145, 146
 de Poisson, 157

Correction de continuité, 176

Corrélation, 301

 linéaire, 300, 301, 313
 coefficients de, 302, 313
 non linéaire, 301

Cote z , 74, 75, 82

Courbe
 de fréquences cumulées, 32, 34
 de Gauss, 163
 normale, 163
 aire sous la, 167
 équation de la, 164

Cycles, 319

D

Déciles, 57, 82

Degrés de liberté, 211, 273, 282, 289, 296

Démarche

 de résolution de problèmes
 de probabilité, 96, 126
 de variables aléatoires, 181
 pour construire
 des classes, 22, 24
 un test d'ajustement du khi-deux, 271, 296
 un test d'hypothèse
 sur une moyenne, 246, 264
 sur un pourcentage, 252, 264
 un test d'indépendance du khi-deux, 288, 296

Dépendance, 290, 292, 304, 313

 entre deux variables quantitatives, 301

Détermination, coefficient de, 307, 313

Diagramme

 à rectangles, 14, 15, 34
 à rectangles horizontaux, 15, 34
 à rectangles verticaux, 14, 15, 34
 circulaire, 15, 16, 34
 de dispersion, 300
 en arbre, 110, 115
 en bâtons, 19, 34
 linéaire, 16, 17, 34

Dispersion relative, 68

Distribution

 bimodale, 48
 conditionnelle, tableau de, 106
 d'échantillonnage
 de la moyenne, 191, 235
 de \hat{P} , 220, 252
 d'un pourcentage, 219, 220
 de probabilité, 134, 136, 181
 tableau de, 13, 18, 21, 81
 règles de présentation d'un, 14
 uniforme, 271

Données

 groupées en classes, 53, 82
 homogénéité des, 69
 non groupées en classes, 50, 51, 82

Droite

 de régression, 305
 de tendance, 308

E

Écart

 absolu moyen, 63
 type, 62, 63, 64, 82
 corrigé, 65
 de la loi de Poisson, 157
 de la population lorsque celui-ci est inconnu, estimation de l', 209
 d'une loi binomiale, 151, 153
 d'une variable aléatoire, 140, 141

- Échantillon(s), 2
 appariés, 258
 dépendants, 258
 représentativité d'un, 279, 296
 statistiques d'un, 191
 taille
 choix de la, 213, 228
 de grande, 203
 de petite, 211
 minimale d'un, 214, 229
 sur la marge d'erreur, effet de la variation de la, 214
- Échantillonnage
 accidentel, 189
 à l'aveuglette, 189
 aléatoire simple, 187
 de \hat{P} , distribution d', 220, 252, 260
 de volontaires, 190
 d'une moyenne, distribution d', 191, 235
 d'un pourcentage, distribution d', 219, 220
 méthodes d'
 non probabiliste, 189
 probabiliste, 187
 panel Web (ou Internet), 190
 par grappes, 189
 par quotas, 190
 sondage en ligne, 190
 stratifié, 189
 systématique, 188
- Échelle(s)
 de mesure, 6, 8
 de rapport, 7
 d'intervalle, 7
 nominale, 6
 ordinaire, 6
- Effectif(s), 14
 observés, 272
 théoriques, 271, 272, 288, 296
- Effet de la variation
 de la taille de l'échantillon sur la marge d'erreur, 214
 du niveau de confiance sur la marge d'erreur, 208
- Éléments équiprobables, 90
- Équation
 de la courbe normale, 164
 de la droite de régression, 305
- Équiprobabilité (*v. Éléments équiprobables*)
- Espace échantillonnal, 89
- Espérance
 de la loi de Poisson, 157, 181
 de la loi normale, 181
 d'une binomiale, 151, 181
 d'une variable aléatoire, 140, 141
 et jeux de hasard, 142
- Estimation
 de l'écart type de la population lorsque celui-ci est inconnu, 209
 d'une moyenne par intervalle de confiance, 203, 205, 235, 247
 d'un pourcentage par intervalle de confiance, 224, 236
 ponctuelle de la moyenne, 209, 216, 236
- Étendue, 61
 d'une série statistique, 23
- Événement(s), 89
 antécédent de x , 136
 certain, 92
 contraire, 92
 dépendants, 103
 impossible, 92
 incompatibilité de deux, 103
 incompatibles, 93, 103
 indépendance de deux, 103
 indépendants, 103, 126
 intersection de deux, 92
 probabilité d'un, 89, 90, 126
 union de deux, 93
- Expérience
 aléatoire, 89
 binomiale, 146
 de Poisson, 157
- F**
- Facteur de correction, 195, 196, 222
 Factorielle, simplification d'une, 119
 Fenêtre, 323, 324
- Fonction de probabilité, 135
 de la loi de Poisson, 157
 d'une loi binomiale, 147
 propriétés d'une, 137
- Fréquence(s)
 absolue, 14
 rectification de, 30
 relative, 14
- G**
- Gauss, courbe de, 163
- H**
- Hasard, loi du, 191, 273
 Histogramme, 25, 27, 34
 à classes
 égales, 25
 inégales, 29
- Homogénéité des données, 69
- Hypothèse
 alternative, 244
 nulle, 243
 test d', 242
- I**
- Incompatibilité de deux événements, 103
 Indécis, répartition des, 230
 Indépendance de deux événements, 103
 Indice de saisonnalité, 327
 Inférence statistique, 88
 Intersection de deux événements, 92
 Intervalle de confiance, 205
 estimation
 d'une moyenne par, 203, 205, 235, 247
 d'un pourcentage par, 224, 236
- K**
- Khi-deux
 calcul du, 272
 condition d'application d'un test du, 272, 289
 critique, 274
 loi(s) du, 273
 caractéristiques des, 274
 table de la, 349
 test d'ajustement du, 270
 construction d'un, 271, 296
 règle de décision d'un, 274
- test d'indépendance du, 286
 construction d'un, 288, 296
- valeur du, 273, 289
- L**
- Limite
 inférieure d'une classe, 22
 supérieure d'une classe, 22
- Lissage
 d'une série chronologique, 321, 333
 par moyenne mobile, 323, 333
 exponentiel, 325, 333

Loi(s)
 binomiale, 145, 181
 approximation de la
 par une loi de Poisson, 160
 par une loi normale, 175
 fonction de probabilité d'une, 147
 table de la, 337
 de Poisson, 157, 181
 approximation d'une loi binomiale par une, 160
 fonction de probabilité de la, 157
 table de la, 157, 342
 de probabilité pour une moyenne d'échantillon, 192
 de Student, 211
 table de la, 211, 348
 du hasard, 191, 273
 du khi-deux, 273
 caractéristiques des, 274
 table de la, 349
 normale, 162, 181
 approximation d'une loi binomiale par une, 175
 centrée réduite, 167, 168
 table de la, 167, 347

M
 Majorité, 48
 Marge d'erreur, 205
 effet de la variation
 de la taille de l'échantillon sur la, 214
 du niveau de confiance sur la, 208
 Médiane, 50, 82
 de données
 groupées en classe, 53, 82
 non groupées en classe, 50, 51, 82
 Mesures
 de dispersion, 61, 82
 de position, 56, 74, 82
 de tendance centrale, 39, 55, 81
 Méthode(s)
 d'échantillonnage, 187, 189
 des moindres carrés, 305
 du rapport à la moyenne mobile pondérée, 327
 non probabilistes, 189
 pour générer des nombres aléatoires
 au moyen de la calculatrice, 188
 au moyen du logiciel Excel, 187
 probabilistes, 187
 Modalités, 3
 Mode, 47, 55, 81
 statistique d'une calculatrice, 70, 309
 Moindres carrés, méthode des, 305
 Moyenne(s), 39, 55, 81
 avec les données brutes, 39
 avec les effectifs, 40
 avec les pourcentages, 41
 d'échantillon, loi de probabilité pour une, 192
 distribution d'échantillonnage d'une, 191, 235
 estimation d'une, par intervalle de confiance, 203, 205, 235, 247
 estimation ponctuelle de la, 209, 216, 236
 mobile, 322, 323
 lissage d'une série chronologique par une, 323, 333
 pondérée, 324
 non pondérée, 322, 333
 pondérée, 45, 81, 333
 représentation graphique de la, 42
 test d'hypothèse sur une, 242, 264
 démarche pour construire un, 246, 264
 règle de décision pour un, 245
 test sur l'égalité de deux, 256
 Multiplication
 principe de, 118, 127
 règle de, 108

N
 Niveau de confiance, 205
 sur la marge d'erreur, effet de la variation du, 208
 Notation
 des classes, 22
 factorielle, 119
 sigma, 40
 Nuage de points, 300

O
 Ogive, 32, 34

P
 Panel Web (ou Internet), 190
 Paramètres d'une population, 191
 Pas de sondage, 188
 Pearson, coefficient de, 302
 Permutation, 119, 127
 Pivot, 42
 Pluralité, 48
 Point critique, 245
 Poisson
 expérience aléatoire de, 157
 loi de, 157, 181
 approximation de la loi binomiale par une, 160
 fonction de probabilité de la, 157
 table de la, 157, 342
 Polygone de fréquences, 27, 34, 75
 Population, 2
 Pourcentage(s)
 distribution d'échantillonnage d'un, 219, 220
 estimation d'un, par intervalle de confiance, 224, 236
 rectifié, 31
 test d'hypothèse sur un, 251, 264
 démarche pour construire un, 252, 264
 test sur l'égalité de deux, 260
 théorème central limite pour un, 221
 Principe de multiplication, 118, 127
 Probabilité(s), 88, 89
 classique 90, 126
 conditionnelle, 100, 101, 126
 distribution de, 134, 136, 181
 d'une variable aléatoire continue, 166, 181
 d'un événement, 89, 90, 126
 empirique, 90, 91, 126
 fonction de, 135
 de la loi de Poisson, 157
 d'une loi binomiale, 147
 propriétés d'une, 137
 fréquentiste, 90, 91
 loi de, pour une moyenne d'échantillon, 192
 propriétés des, 94
 résolution de problème(s) de, 96, 126
 Propriétés
 des probabilités, 94
 d'une fonction de probabilité, 137

Q
 Quantiles, 57, 82
 Quartiles, 57, 82
 Quintiles, 57, 82

R
 Rapport à la moyenne mobile pondérée, méthode du, 327
 Recensement, 2
 Rectification de fréquences, 30
 Région, 157
 Règle(s)
 d'addition, 95, 126

- de décision
 d'un test d'ajustement du khi-deux, 274
 d'un test d'hypothèse sur une moyenne, 245
 d'un test d'indépendance, 290
 de multiplication, 108
 de présentation d'un tableau de distribution, 14
- Régression
 analyse de, 305
 droite de, 305
 linéaire, 305, 313
- Répartition des indécis, 230
- Représentation graphique
 de la moyenne, 42
 de l'intervalle de confiance, 226
 d'une série chronologique, 316
 d'une variable qualitative, 14
 d'une variable quantitative continue, 25
 d'une variable quantitative discrète, 19
- Représentativité d'un échantillon, 279, 296
- Résolution de problèmes de probabilité, 96, 126
- Risque d'erreur, 205
- S**
- Saisonnalité, 320
 indice de, 327
- Série
 chronologique, 316, 333
 composantes d'une, 318
 cyclique, 319
 lissage d'une, 321, 333
 lissage d'une, par moyenne mobile, 323, 333
 lissage exponentiel d'une, 325, 333
 désaisonnalisée, 327, 333
 étendue de la, 23
 statistique, 12
- Seuil de signification, 275
 d'un test, 244
- Sigma, notation, 40
- Simplification d'une factorielle, 119
- Sondage, 2
 en ligne, 190
- Statistiques d'un échantillon, 191
- Student, loi de, 211
 table de la, 211, 348
- Sturges, table de, 23
- T**
- Table
 binomiale, utilisation de la, 153
 de la loi binomiale, 337
 de la loi de Poisson, 157, 342
 de la loi de Student, 211, 348
 de la loi du khi-deux, 349
 de la loi normale centrée réduite, 167, 347
 de Sturges, 23
- Tableau
 à double entrée, 286
 de contingence, 286
 de distribution, 13, 18, 21, 81
 conditionnelle, 106
 règles de présentation d'un, 14
- Taille
 de l'échantillon, 203, 211, 213, 228
 minimale de l'échantillon, 214, 229
- Tendance
 à long terme, 318
 centrale, mesures de, 39, 55, 81
 croissante, 318
 curviligne, 318
 décroissante, 318
 de forme linéaire, 318
 de forme non linéaire, 318
- stable, 318
- Test
 bilatéral, 244
 d'ajustement du khi-deux, 270
 construction d'un, 271, 296
 règle de décision, 274
 d'hypothèse, 242
 démarche pour construire un, 246, 264
 seuil de signification du, 244
 sur une moyenne, 242, 264
 démarche pour construire un, 246, 264
 règle de décision pour un, 245
 sur un pourcentage, 251, 264
 démarche pour construire un, 252, 264
 d'indépendance du khi-deux, 286, 288
 construction d'un, 288, 296
 règle de décision pour un, 290
 du khi-deux, condition d'application d'un, 272, 289
 hypothèses du, 243
 non paramétrique, 270
 paramétrique, 241
 sur l'égalité de deux moyennes, 256
 sur l'égalité de deux pourcentages, 260
 unilatéral à droite, 244
 unilatéral à gauche, 244
- Théorème central limite, 195, 196
 pour un pourcentage, 221
- Tirage sans remise, 147
- U**
- Union de deux événements, 93
- Unité statistique, 2
- V**
- Valeur(s)
 désaisonnalisée, 327
 du khi-deux, 273, 289
 possibles pour une cote z , 75
- Variable(s), 3
 aléatoire, 133
 continue, 134, 181
 probabilité d'une, 166, 181
 discrète, 134, 181
 écart type d'une, 140, 141
 espérance d'une, 140, 141
 dépendance entre deux, 301
 dépendantes, 288
 indépendantes, 288
 qualitative, 3, 4, 6, 81
 nominale, 4, 6, 81
 ordinale, 4, 6, 81
 représentations graphiques d'une, 14
 quantitative, 3, 5, 6
 continue, 5, 6, 81
 représentation graphique d'une, 25
 discrète, 5, 6, 81
 représentation graphique d'une, 19
- Variance, 62, 63
 d'une binomiale, 151
- Variation(s)
 aléatoires, 320
 coefficients de, 68, 82
 cycliques, 319, 333
 de la taille de l'échantillon sur la marge d'erreur, effet de la, 214
 du niveau de confiance sur la marge d'erreur, effet de la, 208
 saisonnières, 320, 333
 uniques, 320, 333
- Z**
- Zone de rejet, 245

Notions de statistique présente les diverses méthodes utilisées pour analyser des données, calculer les risques liés à des situations où le hasard intervient, effectuer une étude par sondage et tester une hypothèse.

À l'instar des éditions précédentes, la troisième édition maintient l'approche pédagogique originale qui a fait son succès :

- un texte clair et concis à la portée des étudiants ;
- une présentation visuelle des concepts ;
- un souci constant de mettre en évidence le sens et la cohérence des notions ;
- une approche intuitive des notions à l'aide de mises en situation ;
- une pédagogie participative avec des exemples à compléter ;
- une démarche d'évaluation continue caractérisée par des exercices de compréhension ;
- une consolidation des apprentissages par des exercices de fin de section et de fin de chapitre contenant un grand nombre de problèmes basés sur des données réelles.

Un cahier de laboratoires Excel conçu pour accompagner le manuel est aussi offert chez l'éditeur.

QUOI DE NEUF DANS CETTE TROISIÈME ÉDITION ?

- Une mise en page plus dynamique, tout en couleurs.
- Des données réactualisées et de nouveaux sujets d'étude plus près des intérêts des étudiants.
- Un contenu enrichi par de nouveaux éléments :
 - au chapitre 1 : une présentation du diagramme linéaire et un résumé des divers types de graphiques ;
 - au chapitre 2 : une section consacrée à l'analyse combinatoire ;
 - au chapitre 5 : le test sur l'égalité de deux moyennes avec des échantillons dépendants.

Christiane Simard a enseigné les mathématiques au collégial pendant 35 ans. Au cours de sa carrière, elle a développé une nouvelle approche pédagogique visant à mettre en évidence le sens et la cohérence des notions.

Madame Simard est également l'auteure des ouvrages suivants : *Notions de statistique. Laboratoires Excel, 4^e édition*; *Méthodes quantitatives, 5^e édition*; *Méthodes quantitatives. Laboratoires Excel, 4^e édition* et *Méthodes quantitatives avancées, 2^e édition*.

ISBN 978-2-89710-813-7



9 782897 108137