

Qu'est ce que la statistique?

La statistique est la science dont le but **est de donner un sens aux données**. L'étude statistique comporte généralement 4 étapes :

- (1) la collecte des données ;
- (2) le traitement des données collectées, aussi appelé la statistique descriptive.
- (3) l'analyse et l'interprétation des données aussi appelée l'inférence statistique, qui s'appuie sur la statistique mathématique.
- (4) la diffusion des résultats d'analyse.

Dans ce cours, l'emphase est mise sur l'étape 2 (chapitre 1) mais surtout sur l'étape 3 (chapitres 2-10).

Applications de la statistique : pharmacologie, psychologie, médecine, environnement, cour de justice, sondages, physique, chimie, sciences sociales, marketing, finance, économétrie, etc.

Chapitre 1

Statistiques descriptives

1.1 Introduction: variables et distributions

Unité statistique (ou unité) : objet pour lequel nous sommes intéressés à recueillir de l'information. Peut être un individu, une compagnie, etc.

Population : ensemble d'unités que l'on cherche à étudier (la population des personnes atteintes du VIH au Canada, la population des électeurs au Québec, la population de voitures fabriquées dans une chaîne de production donnée, la population des accidents observés à une certaine intersection dans la ville de Montréal, etc.)

Échantillon : n'importe quel sous ensemble de la population

Variable : caractéristique d'une unité qui peut prendre différentes valeurs (modalités) pour différentes unités.

Variable qualitative : classe les individus dans un groupe ou une catégorie. Par exemple, le sexe d'un individu (homme, femme) ou le niveau de scolarité d'un individu (primaire, secondaire, universitaire)

Variable quantitative : variable dont les valeurs sont numériques. Les valeurs prises par une telle variable peuvent être continues (température, poids d'un individu) ou discrète (années de scolarité, nombre d'enfants dans un ménage).

Distribution d'une variable : La distribution d'une variable est une correspondance entre les valeurs de la variable et leurs **fréquences** ou leurs **fréquences relatives**. La fréquence représente le nombre d'observations appartenant à une catégorie et la fréquence relative est la fréquence divisée par le nombre total d'observations. La fréquence relative est parfois multipliée par 100, de façon à représenter un pourcentage.

Exemple 1.1 Le tableau 1.1 présente le salaire annuel (en millions de dollars américains) des 40 CEO (Chief Executive Officers) les mieux payés en 2006 ainsi que leur âge et leur plus haut diplôme obtenu. Ces données ont été publiées dans le magazine Forbes, édition du 8 mai 2006.

Ce jeu de données comprend 3 variables qualitatives (Nom, Compagnie et Diplôme) ainsi qu'une variable quantitative discrète (Rang) et 2 variables quantitatives continues (Salaire et Age).

Tableau 1.1 Salaires des CEO (Forbes, 2006)

<i>Rang</i>	<i>Nom</i>	<i>Compagnie</i>	<i>Salaire</i>	<i>Age</i>	<i>Diplôme</i>
1	Fairbank	CapitalOne	249.42	55	MBA
2	Semel	Yahoo	230.55	63	MBA
3	Silverman	Cendant	139.96	65	Droit
4	Karatz	KBHome	135.53	60	Droit
5	Fuld	LehmanBros	122.67	60	MBA
6	Irani	OccidentalPetro	80.73	71	PhD
7	Ellison	Oracle	75.33	61	Aucun
8	Thompson	Symantec	71.84	57	Maitrise
9	Crawford	CaremarkRx	69.66	57	Bacc.
10	Mozilo	Countrywide	68.96	67	Bacc.
11	Chambers	CiscoSystems	62.99	56	MBA
12	Dreier	RylandGroup	56.47	58	Bacc.
13	Frankfort	Coach	55.99	60	MBA
14	Hovnanian	HovnanianEnt	47.83	48	MBA
15	Drosdick	Sunoco	46.19	62	Maitrise
16	Toll	TollBrothers	41.31	65	Droit
17	Ulrich	Target	39.64	63	Bacc.
18	Rollins	Dell	39.32	53	MBA
19	Cazalot	MarathonOil	37.48	55	Bacc.
20	Novak	YumBrands	37.42	53	Bacc.
21	Papa	EOGResources	36.54	59	MBA
22	Termeer	Genzyme	36.38	60	MBA
23	Adkerson	FreeportCopper	35.41	59	MBA
24	Sharer	Amgen	34.49	58	Maitrise
25	Sugarman	IStar	32.94	43	MBA
26	David	UnitedTech	32.73	64	MBA
27	Simpson	XTOEnergy	32.19	57	MBA
28	Lanni	MGMMirage	31.54	63	MBA
29	Jacobs	Qualcomm	31.44	64	PhD
30	Bollenbach	HiltonHotels	31.43	63	MBA
31	Mulva	ConocoPhillips	31.34	59	MBA
32	Mack	MorganStanley	31.23	61	Bacc.
33	Williams	Aetna	30.87	57	Maitrise
34	Lesar	Halliburton	29.36	53	MBA
35	Hanway	Cigna	28.82	54	MBA
36	Cayne	BearStearns	28.4	72	Aucun
37	Amos	Aflac	27.97	54	Bacc.
38	Thiry	DaVita	27.89	50	MBA
39	Rowe	Exelon	26.9	60	Droit
40	Cornelius	Guidant	25.18	62	MBA

Le tableau 1.2 représente la distribution de la variable «*Diplôme* » dans l'exemple 1.1

Tableau 1.2 *Distribution de la variable «Diplôme » dans l'exemple 1.1*

<i>Valeurs</i>	<i>Fréquence</i>	<i>Fréquence Relative</i>
Aucun	2	0,05
Bacc.	8	0,2
Droit	4	0,1
Maitrise	4	0,1
MBA	20	0,5
PhD	2	0,05
	40	1

Il existe de nombreuses méthodes graphiques permettant d'illustrer la distribution d'une variable.

1.2 Quelques méthodes graphiques

Ici, nous mentionnons quelques méthodes graphiques :

Pour variable qualitative :

- (i) Diagramme circulaire (Pie chart, en anglais)
- (ii) Diagramme à bâtons (Bar chart, en anglais)

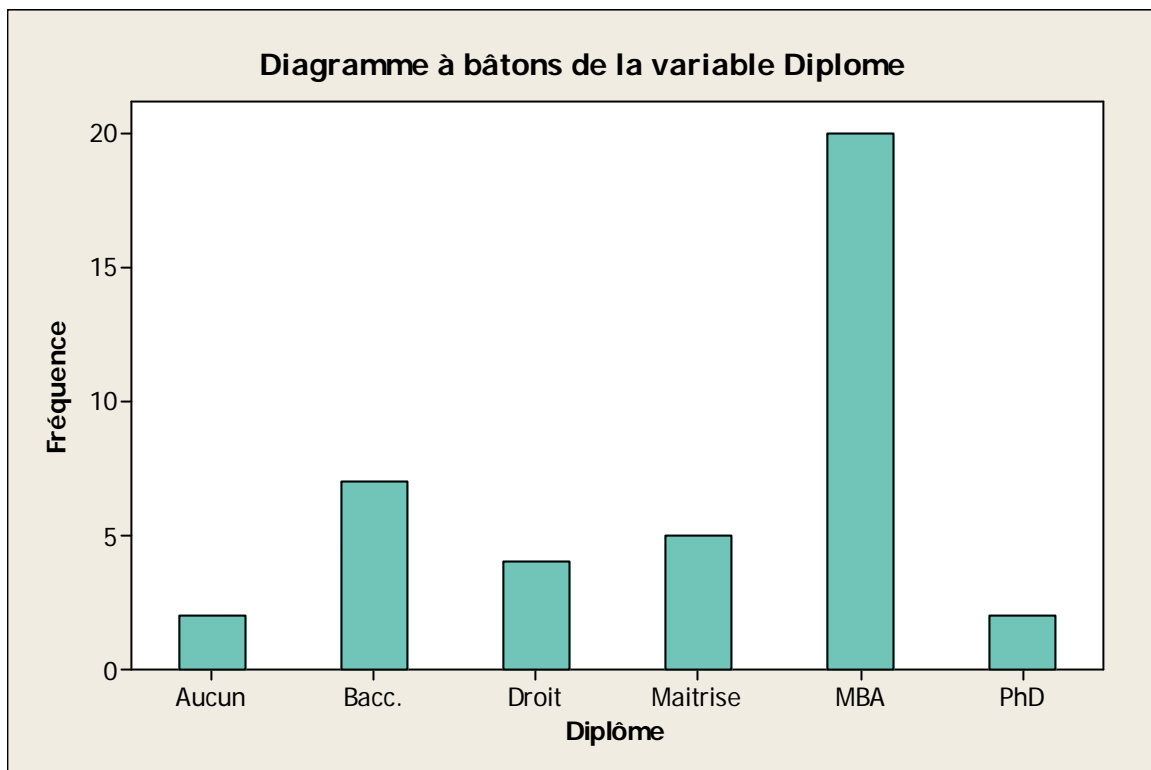
Pour variable quantitative :

- (i) l'histogramme (Histogram, en anglais)
- (ii) le graphique « tiges et feuilles » (Stem-and-leaf, en anglais)
- (iii) le diagramme en boîte (Box plot, en anglais)
(voir section 1.5)

Le diagramme à bâtons est une représentation courante de la distribution d'une variable qualitative.

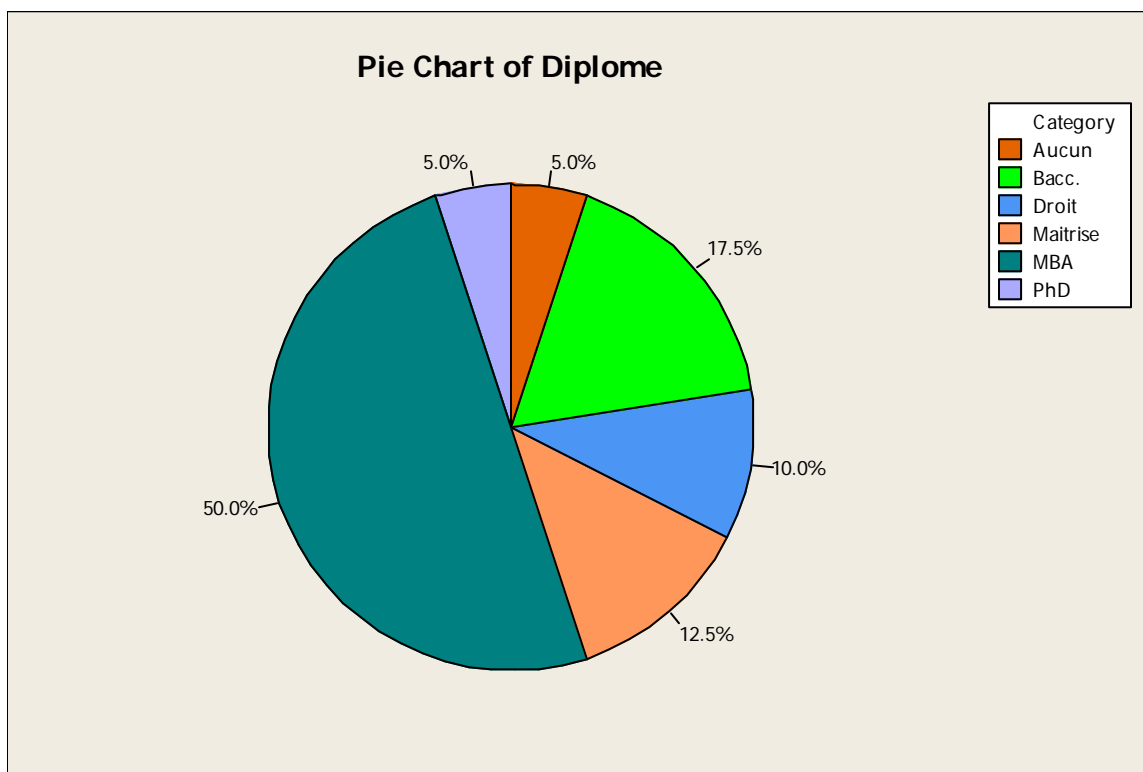
Dans l'exemple 1.1, on avait exhibé la distribution de la variable « Diplôme ». Ci-dessous, on a un diagramme à bâtons.

Figure 1.1 Diagramme à bâtons pour la variable « Diplôme » dans l'exemple 1.1



Le **diagramme circulaire** est une autre représentation courante de la distribution d'une variable qualitative.

Figure 1.2 Diagramme circulaire pour la variable « Diplôme » dans l'exemple 1.1



L'histogramme est la représentation la plus courante de la distribution d'une variable quantitative. Comment construire un histogramme?

1. Choisir un nombre de classe, habituellement entre 5 et 10.

2. Déterminer l'étendue (étendue = plus grande valeur – plus petite valeur). On obtient la largeur de la classe en divisant l'étendue par le nombre de classes choisi à l'étape 1.

3. Arrondir vers le haut la largeur de la classe obtenu à l'étape 2 à une valeur appropriée (si nécessaire).

4. Déterminer les bornes des classes. La plus petite classe doit inclure la plus petite donnée. Si une observation est sur les bornes, compte-t-elle à droite ou à gauche? Le livre les classe à gauche mais il est possible de les classer à droite.

5. Construire un tableau comprenant les classes, leur fréquence respective ainsi que leur fréquence relative respective.

6. Construire l'histogramme en mettant les intervalles sur l'axe horizontal et où les fréquences (ou les fréquences relatives) représente la hauteur des bâtons.

Exemple 1.2 Les données suivantes représentent la moyenne académique de 30 étudiants au département de mathématiques et de statistique

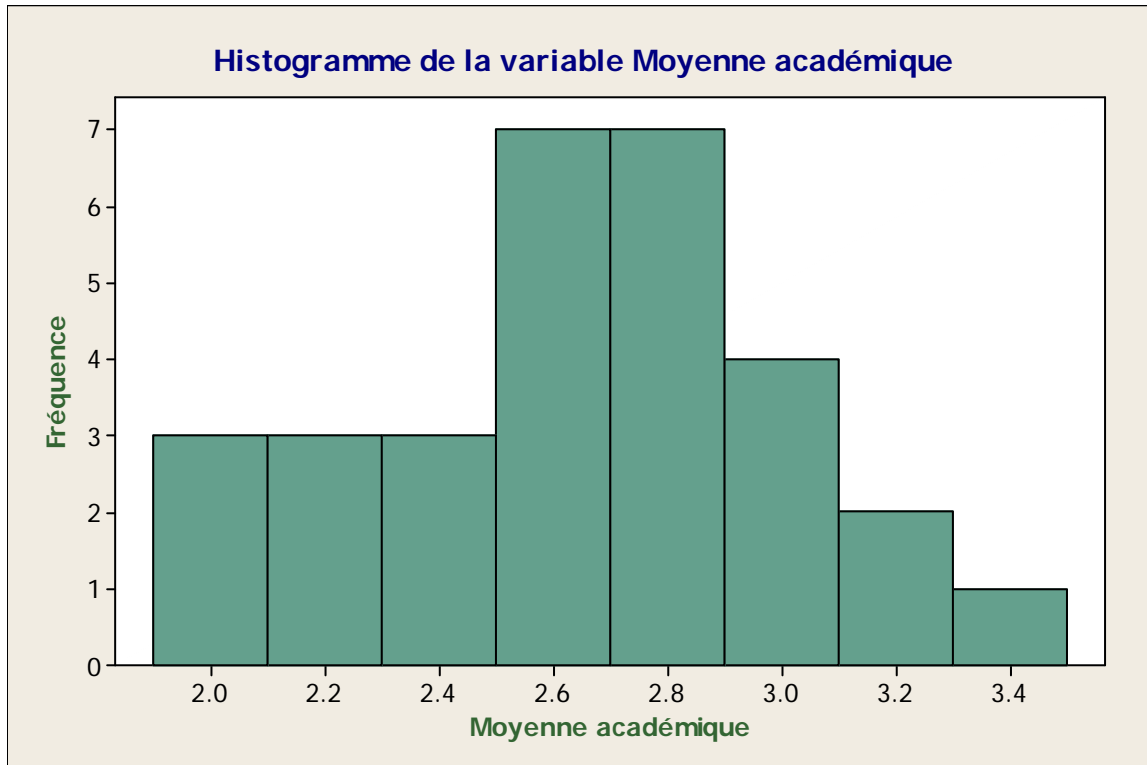
2,0 3,1 1,9 2,5 1,9 2,3 2,6 3,1 2,5 2,1
 2,9 3,0 2,7 2,5 2,4 2,7 2,5 2,4 3,0 3,4
 2,6 2,8 2,5 2,7 2,9 2,7 2,8 2,2 2,7 2,1

Solution :

1. On choisit 8 classes.
2. Étendue = $3,4 - 1,9 = 1,5$. La largeur approximative des classes est donc : $1,5 / 8 = 0,1875$.
3. Arrondir 0.1875 à 0.2. Donc, la largeur de la classe est égale à 0,2.
4. La première classe doit contenir la plus petite valeur, 1,9. Donc la première classe contiendra les données qui tombent dans l'intervalle $[1,9; 2,1)$, etc.
5. On construit le tableau suivant :

Classe	Intervalle	Fréquence	Fréquence Relative
1	1,9 to < 2,1	3	3/30
2	2,1 to < 2,3	3	3/30
3	2,3 to < 2,5	3	3/30
4	2,5 to < 2,7	7	7/30
5	2,7 to < 2,9	7	7/30
6	2,9 to < 3,1	4	4/30
7	3,1 to < 3,3	2	2/30
8	3,3 to < 3,5	1	1/30

Figure 1.3 Histogramme de la variable Moyenne académique dans l'exemple 1.2



Remarques :

(1) Si on avait un très grand nombre d'observations et que l'on utilisait un grand nombre de classes, chacune d'une largeur très étroite, alors l'aspect général de l'histogramme serait une courbe assez lisse.

(2) Le choix du nombre de classes a souvent un impact important sur le « look » de l'histogramme.

(3) Distribution symétrique

Distribution asymétrique à droite

Distribution asymétrique à gauche

Le graphique « tiges et feuilles » est une autre représentation courante de la distribution d'une variable quantitative.

Comment construire une graphique « tiges et feuilles »?

1. Séparer chaque nombre en une tige qui contient tous les chiffres sauf le dernier et une feuille, soit le dernier chiffre. Les tiges ont autant de chiffres que nécessaire, alors que la feuille n'a qu'un seul chiffre.
2. On place les tiges sur une colonne verticale avec la plus petite tige en haut et on fait une ligne à droite de cette colonne.
3. On écrit chaque feuille à droite de sa tige en ordre croissant.

Exemple 1.3

Voici les notes de l'examen final pour le cours STT1700. On cherche à construire un graphique « tiges et feuilles ».

53	98	54	45
55	85	70	57
48	50	84	49
69	64	73	91
50	100	82	58
83	84	96	52
16	78	70	37
68	83	44	81
61	49	59	72

Solution : Les notes varient de 16 à 100. Les tiges seront donc données par : 1, 2, 3, ..., 10.

Figure 1.4 Graphique « Tiges et feuilles »

1	6
2	
3	7
4	4 5 8 9 9
5	0 0 2 3 4 5 7 8 9
6	1 4 8 9
7	0 0 2 3 8
8	1 2 3 3 4 4 5
9	1 6 8
10	0

Avantages : Le graphique « tiges et feuilles » exhibe non seulement les fréquences et la forme de la distribution (comme un histogramme) mais également les données elles mêmes.

Remarque : Les méthodes graphiques sont intéressantes et utiles afin de décrire des données. Cependant, les graphiques ne donnent pas d'information suffisamment précise pour mener une inférence (par exemple, intervalle de confiance, test d'hypothèse, etc.) Nous avons donc besoin de mesures numériques. Ceci fera l'objet des sections 1.3 et 1.4.

1.3 Mesures de tendance centrale

Une mesure de tendance centrale est un indice de la position du centre d'une série de données ou d'une distribution. Elle donne une idée de l'ordre de grandeur des données.

Nous considérons 3 mesures de tendance centrale.

Le mode

Le mode d'une série de données est défini comme la donnée qui a la plus grande fréquence (i.e., la donnée qui apparaît le plus de fois).

Exemple 1.6 : Voici le poids en kg de 15 vaches
425 489 505 398 478 489 500 401 490 399 415 504 433
351 451

Ici, le mode est 489 qui apparaît 2 fois, contrairement à toutes les autres données qui n'apparaissent qu'une seule fois.

Remarque : Le mode n'est pas unique. Dans l'exemple 1.6, si on ajoute la donnée 401, alors il y aura 2 modes : 489 et 401.

La moyenne arithmétique

Soit x_1, x_2, \dots, x_n une série de n données. Leur moyenne arithmétique est définie par

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Remarque : Soit une série de n données, y_1, \dots, y_n et supposons que les y_i s'écrivent comme $y_i = a + bx_i$. Si l'on connaît la moyenne arithmétique des x_i , alors il est facile d'obtenir la moyenne arithmétique des y_i :

$$\bar{y} = a + b\bar{x}$$

Exemple 1.5 Supposons que l'on dispose des températures pour 10 journées estivales en degrés Celsius :

22, 28, 18, 29, 34, 20, 25, 37, 16, 24.

La moyenne des 10 observations est égale à 25.3. Maintenant, une personne vous demande de lui donner la moyenne des 10 jours en degrés Fahrenheit car elle ne comprend pas bien l'échelle des Celsius. La règle suivante est bien connue :

$$1 \text{ F} = (9/5) \times 1 \text{ C} + 32$$

Afin de calculer la moyenne des 10 jours en Fahrenheit, on pourrait transformer en Fahrenheit chacune des données et calculer la moyenne des données transformées. On peut obtenir la moyenne en degrés Fahrenheit plus rapidement en calculant :

$$(9/5) \times 25.3 + 32 = 77.54$$

La médiane

La médiane est la donnée centrale d'une série, lorsque les données sont rangées en ordre croissant ou décroissant.

Lorsque les données sont en nombre impair, la médiane est la donnée centrale. Par exemple, la médiane des données

1, 3, 5, 7, 9, 10, 13

est 7.

Lorsque les données sont en nombre pair, la médiane est la moyenne des deux données centrales. Par exemple, la médiane des données

1, 3, 5, 7, 9, 10, 13, 15

est $(7 + 9)/2 = 8$.

Remarques :

(1) Que le nombre de données soit pair ou impair, le nombre d'observations en dessous de la médiane est égal au nombre d'observations au dessus de la médiane.

(2) Contrairement à la moyenne arithmétique, la médiane n'est pas affectée par la présence de **valeurs atypiques** (ou valeurs aberrantes). Par exemple, dans l'exemple précédent, la moyenne arithmétique est égale à 7,875. Si on remplace la dernière valeur, 15, dans l'exemple précédent par 315, la médiane reste inchangée alors que la moyenne arithmétique est maintenant égale à 45,375. On dira alors que la médiane est une mesure de tendance centrale **robuste à la présence de valeurs aberrantes**.

1.4 Mesures de dispersion

Habituellement, il n'est pas suffisant de rapporter une mesure de tendance centrale car cette dernière ne peut, à elle seule, donner une idée complète d'une série de données ou d'une distribution. On rapportera également des mesures de dispersion.

Il existe plusieurs mesures de dispersion :

L'étendue

L'étendue est définie comme la différence entre la plus grande donnée et la plus petite donnée.

Exemple 1.6 On a la série de donnée suivante :

18 13 11 22 1 4 6 17 8

L'étendue est égale à $22-1=21$.

Écart-type et variance

Comme mesure de la dispersion d'une série de données, on utilise l'écart-type; une quantité associée à l'écart-type est la variance.

Soit x_1, x_2, \dots, x_n une série de n données. Leur **variance** est définie par

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

L'**écart-type** est la racine carrée de la variance, soit

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Formules de calcul:

On démontre facilement que

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}$$

Remarques :

(1) $\sum_{i=1}^n (x_i - \bar{x}) = 0.$

(2) Soit $y_i = a + bx_i$ où a et b sont des constantes. Soit s_x^2 la variance des x_i et s_y^2 la variance des y_i . On démontre aisément les relations suivantes:

$s_y^2 = b^2 s_x^2, \text{ et donc } s_y = b s_x.$

Revenons à l'exemple 1.5. L'écart-type des températures en degrés Celsius est égal à $s_x = 6.78$. Si l'on veut l'écart-type des données en degrés Fahrenheit, il suffit d'appliquer la règle précédente pour obtenir : $s_y = (9 / 5) \times 6.78 = 12.204$.

(3) La variance est fonction de la moyenne des observations. Comme la moyenne, **la variance n'est pas robuste à la présence de valeurs aberrantes.**

(4) Il est parfois utile de connaître la position relative d'une observation. Par exemple, dans le contexte d'un examen, vous voudriez peut être savoir comment votre note se compare à celle des autres étudiants dans la classe. Un exemple d'une mesure de position relative est la **cote z** :

$$z = \frac{x - \bar{x}}{s_x}$$

Notons que la cote z est une transformation linéaire des x_i du type $a + bx_i$, où $a = -\bar{x}/s_x$ et $b = 1/s_x$. Il s'ensuit que la moyenne et l'écart-type de z sont $\bar{z} = 0$ et $s_z = 1$.

La cote z est une distance standardisée entre une donnée et la moyenne des données. **Notons que la cote z n'a pas d'unité.**

Exemple 1.7 Considérons les données de l'exemple 1.5. Après calculs, on obtient $s_x = 6.78$. On a le tableau suivant :

x_i	z_i
22	-0.48650
28	0.39805
18	-1.07620
29	0.54547
34	1.28259
20	-0.78135
25	-0.04423
37	1.72486
16	-1.37104
24	-0.19165
$\bar{x} = 25.3; s_x = 6.78$	$\bar{z} = 0; s_z = 1$

L'écart interquartile (interquartile range)

Définition : Le $p^{\text{ième}}$ percentile ou quantile d'un jeu de données présenté en ordre croissant est la valeur telle qu'au plus $p\%$ des valeurs sont en dessous d'elle et au plus $(100-p)\%$ sont au dessus.

Les percentiles les plus utilisés sont le $25^{\text{ième}}$, le $50^{\text{ième}}$ et $75^{\text{ième}}$ appelés respectivement le premier quantile (Q_1), la médiane (Q_2) et le troisième quantile Q_3 .

L'écart interquartile est défini selon

$$IQR = Q_3 - Q_1$$

Comment obtenir Q_1 et Q_3 ?

- Q_1 est la donnée en position $\frac{(n+1)}{4}$ et Q_3 est la donnée en position $\frac{3(n+1)}{4}$.
- Rien ne garantit que les valeurs $\frac{(n+1)}{4}$ et $\frac{3(n+1)}{4}$ seront entières. Dans ce cas, les positions des quartiles sont déterminées par interpolation.

Exemple 1.8 On a 26 données

1 1 2 3 3 3 4 4 5 5 5 5 5 6 6 6 7 7 7 8 9 9 9 9 9 9

La médiane est donnée par $\frac{5+6}{2} = 5,5$

La position de Q_1 est donnée par $27/4 = 6,75$. Donc Q_1 est donnée par la valeur à $3/4$ de la distance entre les valeurs 3 et 4 et on a $Q_1 = 3,75$.

De manière similaire, la position de Q_3 est donnée par $0,75 \times 27 = 20,25$. Donc Q_3 est donnée par la valeur à $1/4$ de la distance entre les valeurs 8 et 9 et on a $Q_3 = 8,25$.

Donc,

$$IQR = Q_3 - Q_1 = 8,25 - 3,75 = 4,5.$$

Remarques:

(1) Contrairement à s^2 , l'IQR est une robuste aux valeurs aberrantes.

(2) Le résumé de 5 chiffres (five-number summary dans le livre) est composé de : min, Q_1 , médiane, Q_3 , max.

1.5 Une autre méthode graphique : Le diagramme en boîte ou boxplot

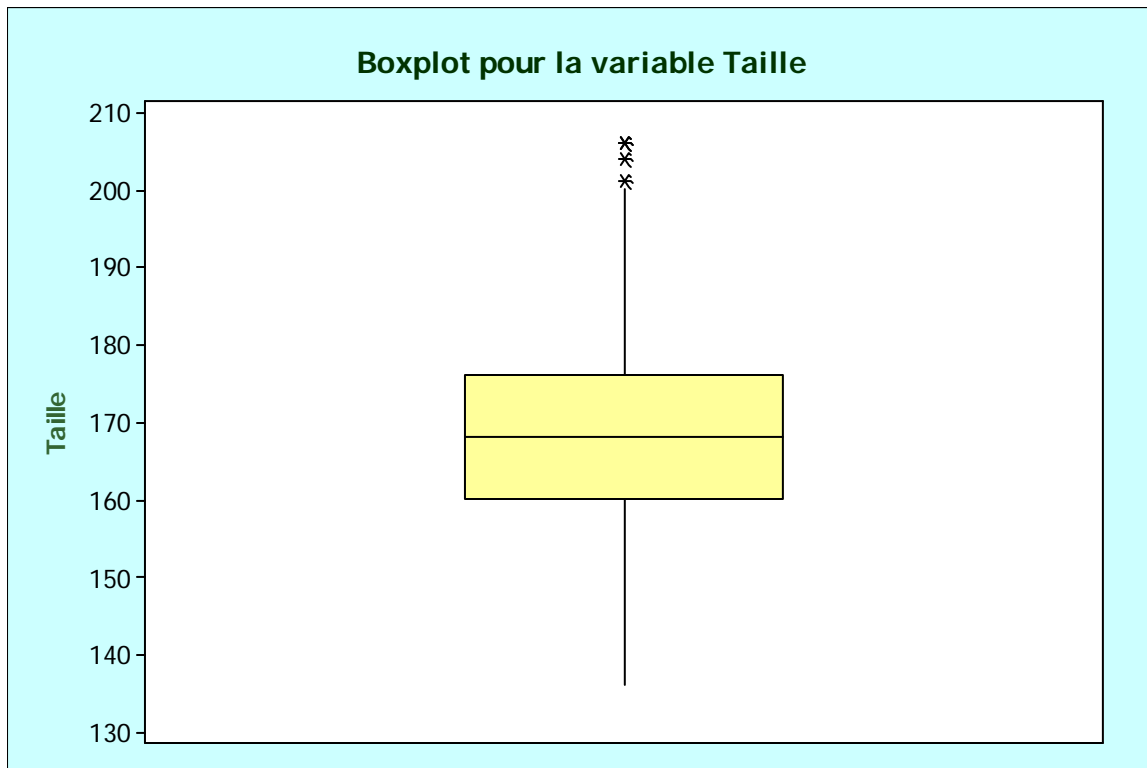
Le boxplot est une autre méthode graphique permettant d'étudier la distribution d'une variable quantitative. Le boxplot incorpore le résumé de 5 chiffres (minimum, Q_1 , médiane, Q_3). La figure 1.7 représente le boxplot décrivant la taille (en cm) de 1000 hommes et 1000 femmes.

- La boîte est délimitée par Q_1 et Q_3 . La longueur de la boîte représente donc l'IQR.
- La ligne dans la boîte représente la médiane. Si la ligne coupe la boîte en 2 rectangles égaux, alors la distribution est symétrique. Sinon elle est asymétrique.
- Les étoiles représentent les données qui sont jugées aberrantes. Une donnée x est jugée aberrante si

$$x < Q_1 - 1.5IQR \text{ ou si } x > Q_3 + 1.5IQR$$

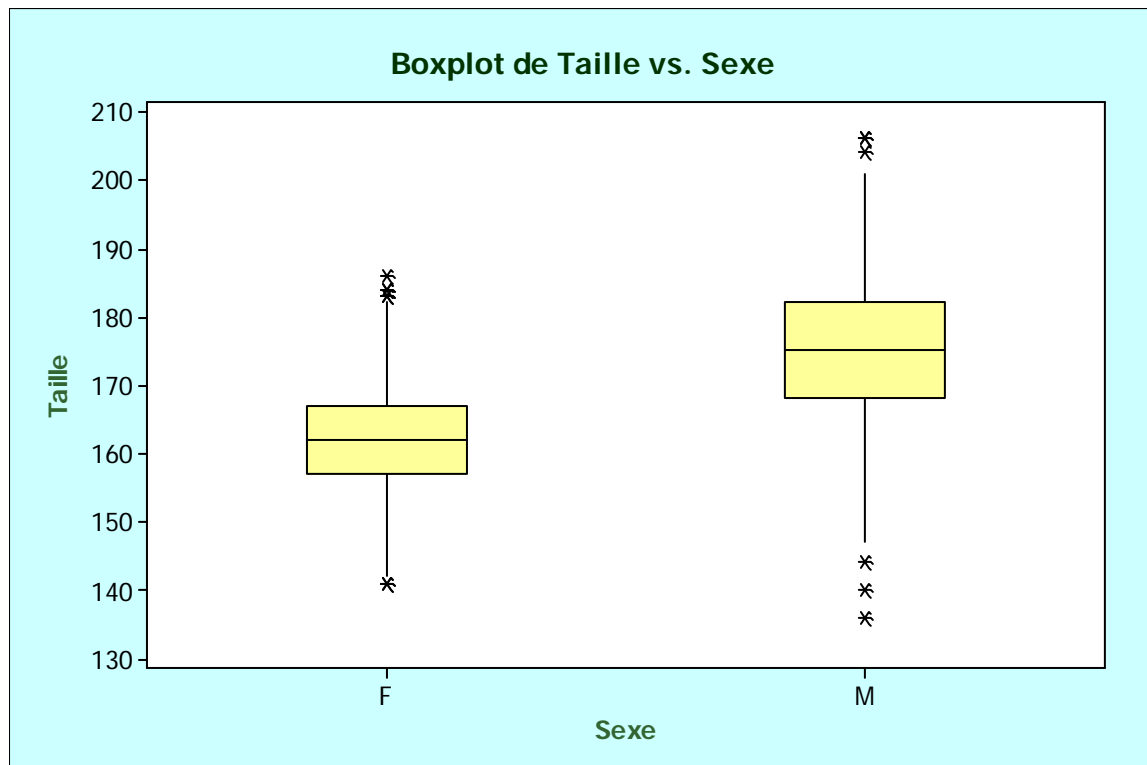
- Les points $Q_1 - 1.5IQR$ et $Q_3 + 1.5IQR$ sont souvent appelés « clôtures » (fences, en anglais)

Figure 1.7 Boxplot pour la variable Taille



On peut également faire des « side-by-side boxplots » qui permettent de comparer la distribution d'une variable quantitative selon les modalités d'une variable qualitative. La Figure 1.8 représente 2 « side by side boxplots » décrivant la taille (en cm) selon le sexe.

Figure 1.8 Boxplot pour la variable Taille selon la variable Sexe



1.6 Une règle empirique : la règle 68-95-99

Un histogramme est **en forme de cloche** s'il possède un seul mode, s'il est symétrique et si ses queues diminuent graduellement.

Si un histogramme est en forme de cloche, alors la règle empirique suivante, appelée **la règle 68-95-99**, s'applique.

Soit une série de n données dont la distribution est en forme de cloche. Alors :

- (i) l'intervalle $(\bar{x} - s, \bar{x} + s)$ contient **approx. 68%** des données.
- (ii) l'intervalle $(\bar{x} - 2s, \bar{x} + 2s)$ contient **approx. 95%** des données.
- (iii) l'intervalle $(\bar{x} - 3s, \bar{x} + 3s)$ contient **approx. 99%** des données.

Remarques :

- (1) Si la distribution n'est pas en forme de cloche, on peut toujours appliquer la règle 68-95-99, mais les résultats risquent d'être fort imprécis.
- (2) Les données comprises dans l'intervalle $(\bar{x} - s, \bar{x} + s)$ sont celles dont la cote z se situe entre -1 et 1. De manière similaire, les données comprises dans l'intervalle $(\bar{x} - 2s, \bar{x} + 2s)$ sont celles dont la cote z se situe entre -2 et 2 et les données comprises dans l'intervalle $(\bar{x} - 3s, \bar{x} + 3s)$ sont celles dont la cote z se situe entre -3 et 3.
- (3) Cette règle sera justifiée plus formellement au Chapitre 4.

1.7 Corrélation et droite des moindres carrés

Jusqu'ici, nous avons mis l'accent sur la description d'une variable (qualitative ou quantitative). En pratique, la relation entre deux variables quantitatives continues est fréquemment étudiée. Cette relation est exposée clairement à l'aide d'un **nuage de points**. Considérons l'exemple suivant :

Exemple 1.9 Le tableau 1.3 présente, pour un ensemble de 41 maisons vendues à Outremont au printemps 1981, les valeurs de deux variables:

x : L'évaluation municipale, en milliers de dollars.

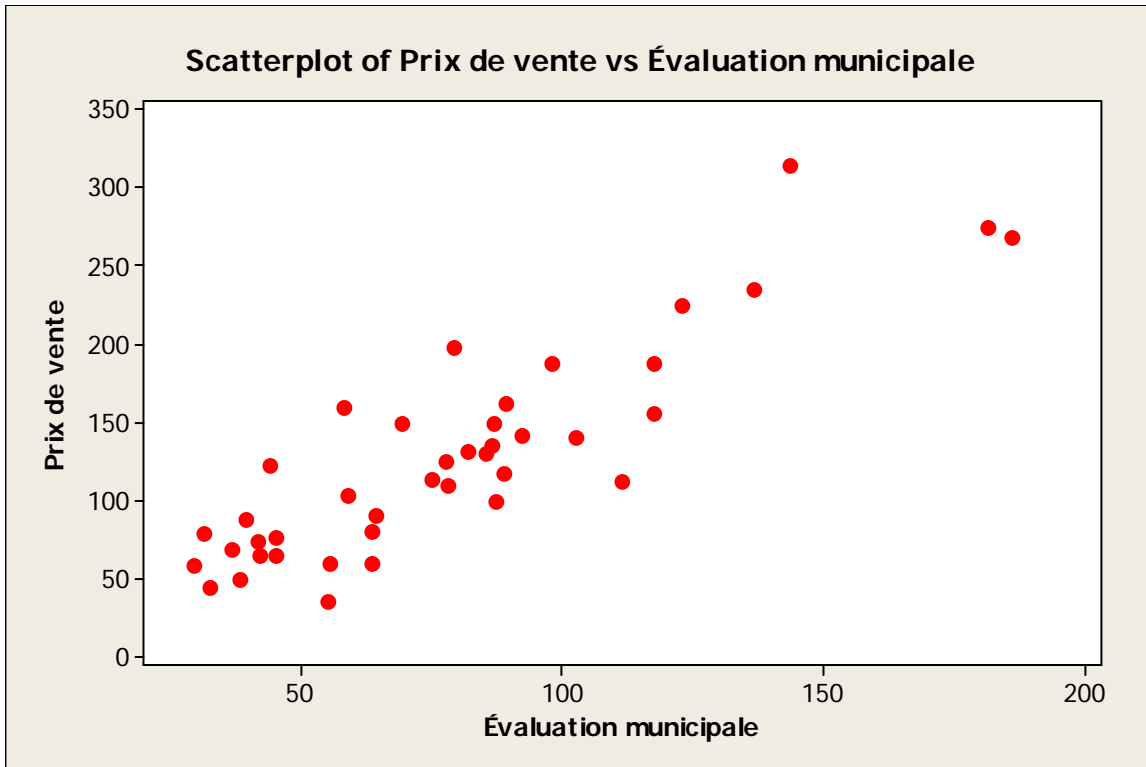
y : Le prix à la vente, en milliers de dollars.

Tableau 1.3 *Prix de vente (y)
et évaluation municipale (x)
de 41 maisons vendues à
Outremont*

x	Y	x	y	x	Y	X	y
45,3	65	136,4	235	88,8	117	29,3	58
55,6	60	77,5	125	58,1	160	82,0	132
102,5	140	111,2	112	98,0	188	79,0	198
38,2	50	186,1	268	39,3	88	64,4	90
63,6	60	69,3	150	58,8	103	78,0	110
63,6	80	32,5	45	42,2	65	75,1	114
41,6	74	92,3	142	45,2	77	86,8	149
31,3	79	36,6	69	117,4	188	143,5	315
181,4	275	87,4	100	117,3	156	85,4	130
89,3	162	44,0	123	122,9	225	86,6	135
54,9	36						

La Figure 1.9 montre clairement qu'il y a une relation entre les deux variables. De plus, cette relation semble linéaire. Mais certaines relations sont plus fortes que d'autres.

Figure 1.9 Nuage de points
*Relation entre l'évaluation municipale et le prix de vente
à Outremont*



Un indice de la force d'une relation linéaire est le **coefficient de corrélation** r , une mesure définie par

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$
$$= \frac{1}{(n-1)} \sum_{i=1}^n \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}$$

Remarques :

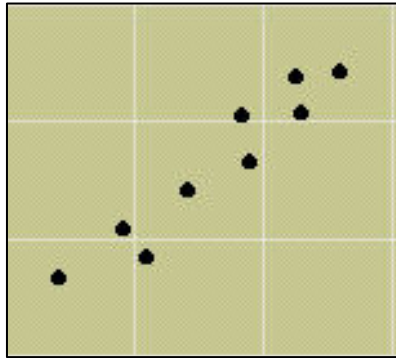
- (1) $-1 \leq r \leq 1$
- (2) $|r| = 1$ si et seulement si il existe des constantes a et b telles que $y_i = a + bx_i$ pour tout i , c'est-à-dire, si et seulement si les points du nuage se situent tous sur une même droite.
- (3) $r > 0$ lorsqu'on a une **association positive** et $r < 0$ lorsqu'elle est négative.

(4) La corrélation est définie comme le produit croisée (divisée par $(n-1)$) des quantités $\left(\frac{x_i - \bar{x}}{s_x}\right)$ et $\left(\frac{y_i - \bar{y}}{s_y}\right)$. La corrélation demeure donc inchangée si on change les unités de mesure de l'une et/ou de l'autre variable.

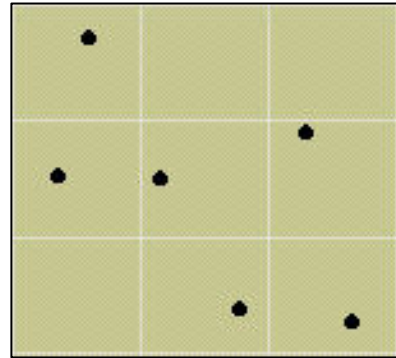
(5) **Attention :** le coefficient de corrélation donne une mesure de la force de la relation entre 2 variables si cette dernière est linéaire. Si la relation n'est pas linéaire, on ne peut utiliser le coefficient de corrélation.

(6) Le coefficient de corrélation n'est pas robuste à la présence d'observations atypique. En effet, la présence d'une seule valeur atypique peut faire varier le coefficient de corrélation de manière drastique.

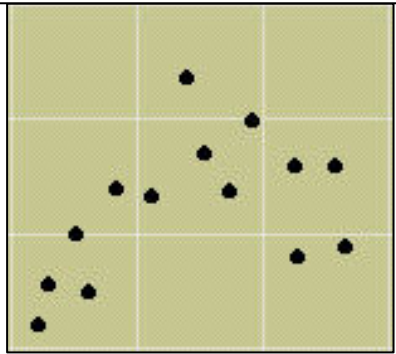
Figure 1.10 Différent types de relations



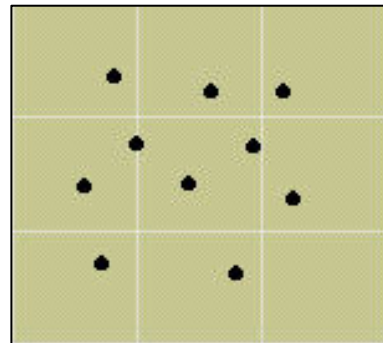
Relation linéaire positive et forte



Relation linéaire négative et faible



Relation non-linéaire



Aucune relation

Formule de calcul

La formule suivante permet de calculer le coefficient de corrélation plus aisément:

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \sqrt{\sum y_i^2 - n\bar{y}^2}}$$

Dans l'exemple 1.7, on calcule $r = 87,6\%$, ce qui indique que la relation est relativement forte.

Droite des moindres carrés

Lorsque le nuage de points montre qu'il existe une relation entre deux variables, et que **cette relation est linéaire**, il est bon de l'exprimer à l'aide de l'une équation d'une droite,

$$y = a + bx.$$

Cette droite doit passer le plus près possible des points du nuage. Pour préciser cette notion, nous devons définir une mesure de la distance entre le nuage et la droite. Celle que nous adoptons est une quantité D définie par:

$$D = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

où $\hat{y}_i = a + bx_i$ est le point sur la droite d'abscisse x_i .

La quantité D est donc la somme des carrés des distances verticales, $e_i = y_i - \hat{y}_i$, entre les points du nuage et la droite.

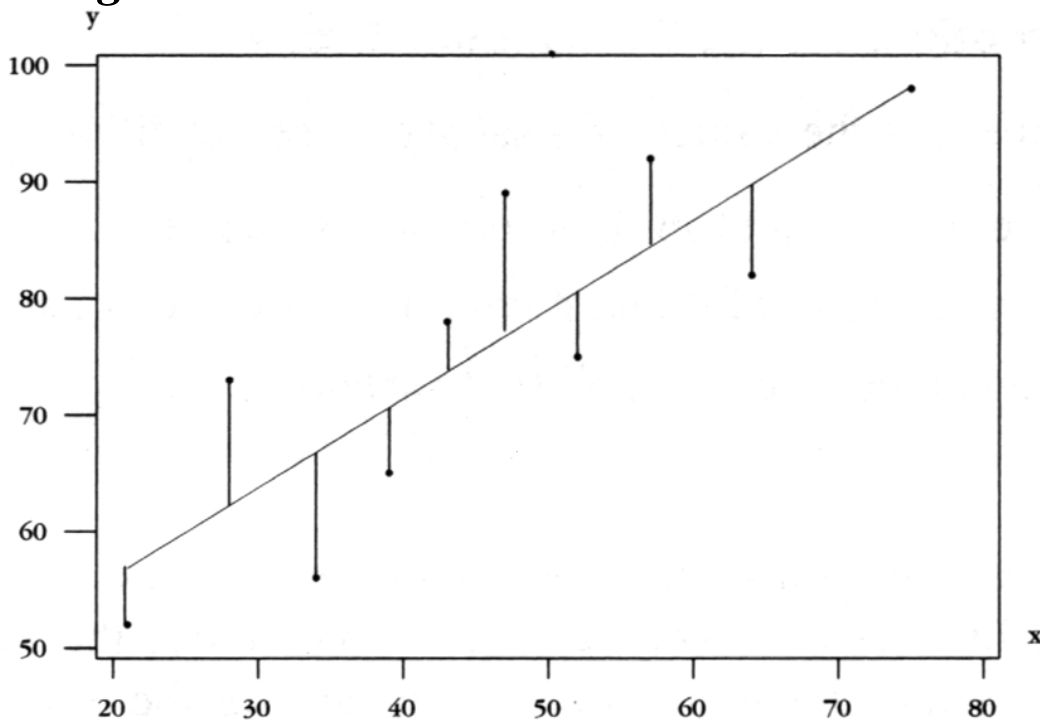
La **droite des moindres carrés** est celle qui minimise D .

Le problème est donc

$$\begin{aligned} \text{minimiser } D &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - a - bx_i)^2 \end{aligned}$$

par rapport à a et b .

Figure 1.11 Minimiser les distances verticales



Les valeurs de a et b qui minimisent D satisfont:

$$\begin{aligned}\frac{\partial D}{\partial a} &= -2 \sum_{i=1}^n (y_i - a - bx_i) = 0, \\ \frac{\partial D}{\partial b} &= -2 \sum_{i=1}^n (y_i - a - bx_i)x_i = 0.\end{aligned}$$

La première équation donne

$$\bar{y} = a + b\bar{x} \Rightarrow a = \bar{y} - b\bar{x}.$$

En substituant cette expression à a dans la deuxième équation, nous obtenons

$$\begin{aligned}\sum_{i=1}^n (y_i - a - bx_i)x_i &= \sum_{i=1}^n (y_i - \bar{y} + b\bar{x} - bx_i)x_i \\ &= \sum_{i=1}^n [y_i - \bar{y} - b(x_i - \bar{x})]x_i \\ &= \sum_{i=1}^n (y_i - \bar{y})x_i - b \sum_{i=1}^n (x_i - \bar{x})x_i\end{aligned}$$

ce qui entraîne

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})x_i}{\sum_{i=1}^n (x_i - \bar{x})x_i}.$$

Il existe une autre expression pour le numérateur. Notez que

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n x_i (y_i - \bar{y}) - \sum_{i=1}^n \bar{x} (y_i - \bar{y}) \\ &= \sum_{i=1}^n x_i (y_i - \bar{y}) - \bar{x} \sum_{i=1}^n (y_i - \bar{y}) \\ &= \sum_{i=1}^n x_i (y_i - \bar{y}),\end{aligned}$$

utilisant le fait que $\sum_{i=1}^n (y_i - \bar{y}) = 0$. On montre de la même façon que

$$\sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Nous obtenons alors

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Les coefficients de la droite des moindres carrés sont donc:

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \text{ et } a = \bar{y} - b\bar{x}.$$

Voici une formule de calcul de b :

$$b = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}.$$

Remarques :

(1) La droite des moindres carrés passe par le point (\bar{x}, \bar{y}) .

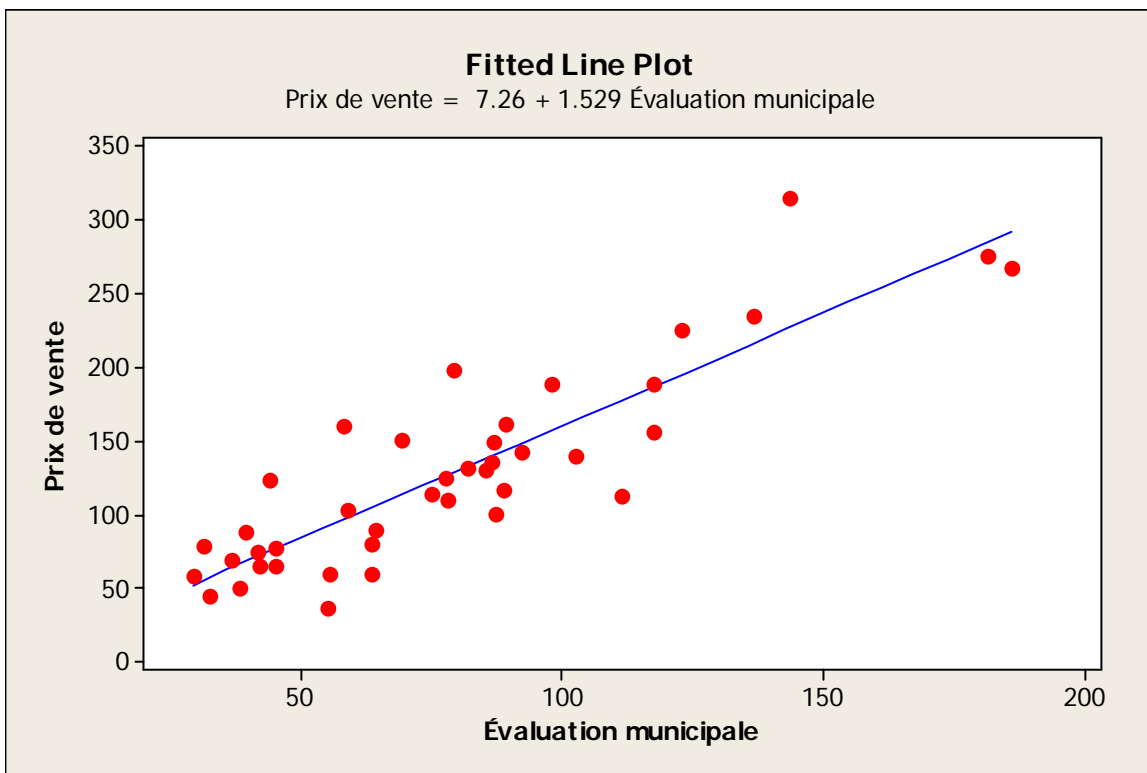
(2) En comparant les expressions de b et de r , on constate que r et b sont de même signe et que $r = 0 \Leftrightarrow b = 0$. Nous avons la relation suivante:

$$r = b \frac{s_x}{s_y}, \text{ et donc } b = r \frac{s_y}{s_x}$$

Donc $r > 0$ si et seulement si la droite des moindres carrés est de pente positive, et $r = 0$ si et seulement si la droite des moindres carrés est horizontale.

(3) Le coefficient a est appelée **ordonnée à l'origine** et représente la valeur sur la droite lorsque $x = 0$. Le coefficient b représente **la pente** de la droite des moindres carrés. Donc, lorsque l'on augmente x d'une unité, la variable y croît de b unités si $b > 0$ et décroît de b unités si $b < 0$.

Figure 1.12 Droite des moindres carrés dans l'exemple 1.7



Dans l'exemple 1.7, on a $a = 7,26$ et $b = 1,529$. La droite de régression est donc donnée par

$$y = 7,26 + 1,529 x$$

Par exemple, pour une maison dont l'évaluation municipale est de 180 000\$, on prédit que le prix de vente sera égal à 282 480\$ puisque

$$y = 7,26 + 1,529 \times 180 = 282,48.$$

Chapitre 2

Probabilités

2.1 - Définitions, axiomes et propriétés

Avant de donner des définitions formelles, essayons de comprendre la signification du mot **probabilité**.

Considérons les expériences suivantes :

- 1) On lance une pièce de monnaie. Intuitivement, on sait que la probabilité d'obtenir FACE est $\frac{1}{2}$.

Quelle est l'interprétation de $P(\text{FACE}) = \frac{1}{2}$?

- 2) Le lecteur de météo au télé-journal nous dit que demain, il y a 35% de chance de précipitations; c'est-à-dire, la probabilité qu'il pleuve demain est 35%. Comment interpréter cette affirmation?

- 3) On lance un dé équilibré. On sait que la probabilité d'obtenir 2 est $\frac{1}{6}$. Quelle est l'interprétation de $P(\text{obtenir } 2) = \frac{1}{6}$?

Définitions

Expérience aléatoire Une expérience aléatoire est une expérience dont on ne peut pas prédire les résultats avec certitude.

Espace échantillon L'ensemble des résultats possibles d'une expérience aléatoire est appelé espace échantillon. Il sera dénoté par Ω .

Événement Un événement A est un sous-ensemble de l'espace échantillon Ω .

Événement élémentaire Un événement élémentaire est un événement qui ne peut être décomposé.

Remarque : Les définitions d'événement et d'événement élémentaire implique qu'un événement est une union d'événements élémentaires.

Exemple 2.1

On tire une personne aléatoirement dans la population et on recueille son groupe sanguin. On a $\Omega = \{A, B, AB, O\}$. Les événements $E_1 = \{A\}$, $E_2 = \{B\}$, $E_3 = \{AB\}$ et $E_4 = \{O\}$ sont les événements élémentaires.

Exemple 2.2

1)Expérience: Lancer deux pièces de monnaie. $\Omega = \{PP, PF, FP, FF\}$.

Les quatre événements élémentaires sont : $E_1 = \{PP\}$, $E_2 = \{PF\}$, $E_3 = \{FP\}$ et $E_4 = \{FF\}$.

Événement	Signification courante
$\{PP,PF\}$	Le premier lancer donne une pile
$\{PP,PF,FP\}$	Obtenir au moins une pile
$\{PF,FP\}$	Obtenir exactement une pile

2) Expérience: Lancer un dé. $\Omega = \{1,2,3,4,5,6\}$

Événement	Signification courante
$\{1,2,3\}$	Le résultat est inférieur ou égal à 3
$\{2,4,6\}$	Le résultat est pair
$\{1,5\}$	Avoir "1" ou "5"

Combinaisons d'événements

<i>Opération</i>	<i>Sens concret</i>
<i>Réunion</i>	$A \cup B$ équivaut à l'énoncé « A ou B s'est produit ».
<i>Intersection</i>	$A \cap B$ équivaut à l'énoncé « A et B se sont produits ».
<i>Complémentation</i>	A^c équivaut à l'énoncé « A ne s'est pas produit ».
<i>Différence</i>	$A - B$ (ou $A \setminus B$) équivaut à l'énoncé « A s'est produit mais pas B ». Notez que $A \setminus B = A \cap B^c$.

Lois de Morgan:

$$(A \cap B)^c = A^c \cup B^c$$

et

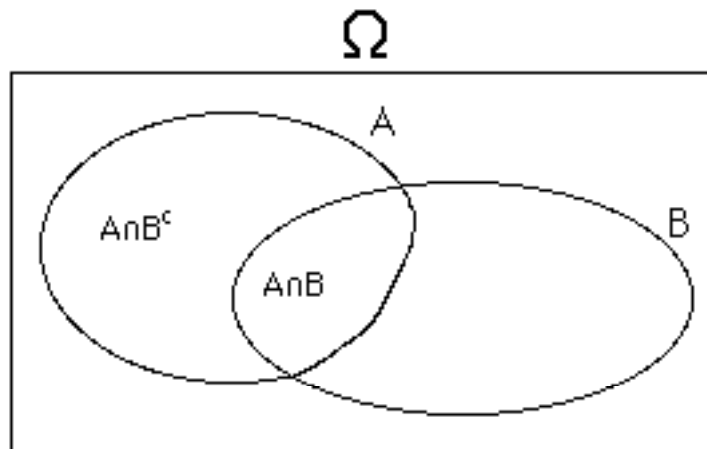
$$(A \cup B)^c = A^c \cap B^c.$$

Décomposition utile:

$$A = (A \cap B) \cup (A \cap B^c),$$

où $(A \cap B)$ et $(A \cap B^c)$ sont disjoints, c'est-à-dire,

$$(A \cap B) \cap (A \cap B^c) = \emptyset.$$



Événement impossible et événement certain.

L'ensemble vide \emptyset et l'ensemble Ω sont respectivement, l'événement impossible et l'événement certain.

Événements disjoints ou incompatibles

Deux événements A et B sont dits **disjoints** ou **mutuellement exclusifs**, ou **incompatibles**, s'ils ne peuvent pas se produire en même temps.

Formellement,

A et B sont dits incompatibles si $A \cap B = \emptyset$

Exemple 2.3 On tire au hasard une personne au hasard dans une certaine population.

Considérons les événements suivants:

A: La personne choisie a les yeux bleus

B: La personne choisie a les cheveux
blonds

C: La personne choisie a les yeux verts

<i>Événement</i>	<i>Sens concret</i>
$A \cup C$	
$A \cap B$	
A^c	
$A - B$	
$(A \cap B)^c$	
$A^c \cup B^c$	

Axiomes

Définition : Une fonction P qui fait correspondre à chaque événement $A \subset \Omega$ un nombre réel $P(A)$ est appelée une **probabilité** si elle satisfait les axiomes suivants:

A1 $P(A) \geq 0$ pour tout événement A

A2 Si A_1, A_2, \dots, A_k sont des événements disjoints deux à deux, alors

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k)$$

A3 $P(\Omega) = 1$

Propriétés

1 $P(A^c) = 1 - P(A)$

2 $P(\emptyset) = 0$: la probabilité de l'événement impossible est 0.

$$3 \quad P(A) = P(A \cap B) + P(A \cap B^c)$$

$$4 \quad \text{Si } B \subset A, \text{ alors } P(B) \leq P(A)$$

$$5 \quad P(A) \leq 1 \text{ pour tout \u00e9v\u00e9nement } A \subset \Omega$$

$$6 \quad P(A - B) = P(A) - P(A \cap B)$$

$$7 \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Exemple 2.4 Soit A et B deux événements tels que

$$P(A) = 0,3, P(B) = 0,5 \text{ et } P(A \cup B) = 0,7.$$

Déterminer : a) $P(A \cap B)$; b) $P(A^c \cap B^c)$;

c) $P(B \cap A^c)$.

Solution :

Exemple 2.5 Un étudiant prend un cours de biologie et un cours de statistique. La probabilité qu'il réussisse le cours de biologie est 0,5 alors que la probabilité qu'il réussisse celui de statistique est 0,7. La probabilité qu'il réussisse les deux cours est 0,3.

Déterminer :

- a) La probabilité qu'il réussisse au moins un cours.
- b) Il coule les deux cours.
- c) Il coule le cours de statistique mais réussit le cours de biologie.

Solution :

2.2 Attribution des probabilités

A chaque résultat $\omega \in \Omega$ on fait correspondre une probabilité $p(\omega)$, un nombre qui satisfait

$$0 \leq p(\omega) \leq 1$$

pour tout $\omega \in \Omega$. De plus, la somme des probabilités de tous les éléments de Ω est égale à 1

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

Remarque Comment attribue-t-on ces probabilités? Il y a deux façons de le faire, l'une a priori, l'autre empirique.

Méthode a priori Les probabilités « a priori » sont des probabilités qui semblent intuitivement « raisonnables », généralement à cause de certains aspects physiques de l'expérience. Cette intuition conduit normalement à un modèle particulièrement simple, appelé modèle **d'équiprobabilité**, dans lequel on attribue à chaque résultat la même probabilité. Il s'agit là d'une supposition — a priori raisonnable — au sujet de la nature du phénomène observé. Par exemple, lorsqu'on lance un sou, il est naturel de supposer, à moins d'évidence contraire, que la probabilité d'avoir face est égale à la probabilité d'avoir pile. Cette probabilité vaut alors

1/2. En général, si Ω contient n résultats équiprobables, alors chacun a probabilité $1/n$.

Méthode empirique L'hypothèse d'équiprobabilité est une hypothèse scientifique qu'on doit tôt ou tard confronter à l'expérience; et elle peut être remise en question à la suite de certains faits d'observation empirique. Il est naturel, par exemple, de supposer équiprobables les résultats « garçon » et « fille » lorsqu'on observe une naissance; pourtant, les nombreuses données sur les naissances montrent que cette hypothèse n'est pas strictement vraie. On constate en effet que la proportion de garçons à la naissance est plutôt de 51 % et non de 50 %. Dans la plupart des applications l'intuition ne fournit même pas une première approximation. Seules les données d'enquêtes peuvent fournir l'information. Quelle est, par exemple, la probabilité qu'une personne tirée dans la population canadienne ait 70 ans ou plus? On n'en aurait aucune idée sans les données du recensement qui révèlent que le pourcentage de Canadiens de 70 ans et plus est de 7,6 %. La probabilité voulue est donc de 0,076. De même, les données actuarielles nous permettent d'estimer que la probabilité qu'une femme de 18 ans vive jusqu'à l'âge de 65 ans et au-delà est de 0,87.

Probabilité d'un événement

Ayant défini la probabilité d'un résultat, nous pouvons définir la probabilité d'un événement: la probabilité d'un événement A est la somme des probabilités des résultats contenus dans A :

$$P(A) = \sum_{\omega \in A} P(\omega)$$

Sous l'hypothèse d'équiprobabilité, cette règle prend une forme particulièrement simple:

$$P(A) = \frac{\text{Card}(A)}{\text{Card}(\Omega)}$$

où $\text{Card}(\cdot)$ désigne la cardinalité d'un ensemble, le nombre d'éléments qu'il contient.

2.3 Probabilités conditionnelles

Afin de saisir la notion de probabilité conditionnelle, considérons les exemples suivants :

- 1) On tire une personne au hasard au Canada. On sait que P (la personne tirée soit francophone) ≈ 0.2 . Si après avoir tiré cette personne, on sait qu'elle vient du Québec, alors P (la personne tirée soit francophone) ≈ 0.8 .
- 2) On tire une carte dans un jeu de 52 cartes. On sait P (la carte tirée soit un Roi) $= \frac{4}{52} = \frac{1}{13}$. Si vous avez vu que la carte tirée est une figure noire, alors P (la carte tirée soit un Roi) $= \frac{2}{6} = \frac{1}{3}$.

Remarques Dans les exemples précédents, soit A : la personne est francophone et B : la personne tirée vient du Québec. Alors, $P(A|B) \approx 0.8$.

Soit A : la carte tirée est un Roi et B : la carte tirée est une figure noire. Alors, $P(A|B) = 1/3$.

Définition La **probabilité conditionnelle** d'un événement B étant donné un événement A , dénotée par $P(B|A)$, est définie par

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

à condition que $P(A) \neq 0$.

Remarque La probabilité conditionnelle $P(B|A)$ n'est pas définie si $P(A) = 0$, puisque la définition ci-dessus entraînerait alors une division par 0. Cette restriction est conforme au sens intuitif de probabilité conditionnelle: on ne saurait imposer comme condition qu'un événement impossible se soit réalisé.

La notion de probabilité conditionnelle permet de donner une formule générale de la probabilité de l'intersection de deux événements A et B , formule qui découle immédiatement de la définition formelle de probabilité conditionnelle:

$$P(A \cap B) = P(A)P(B | A)$$

Puisqu'on peut échanger A et B dans cette formule, on a aussi:

$$P(A \cap B) = P(B)P(A | B)$$

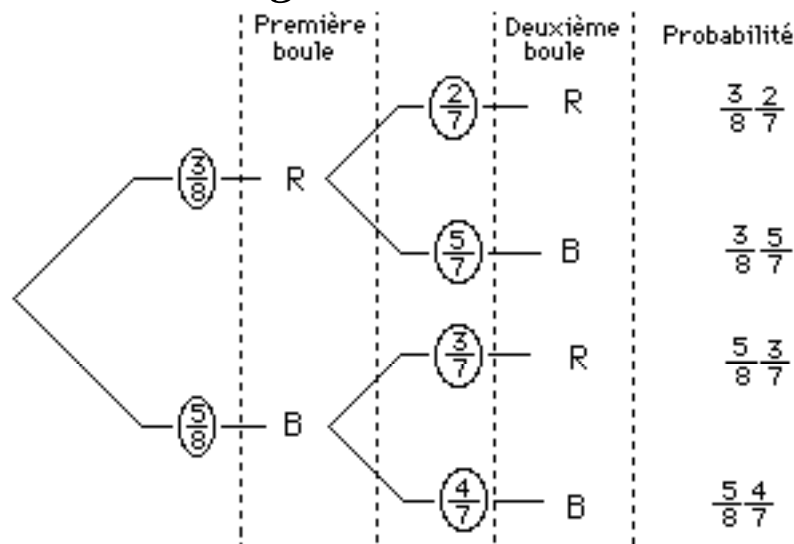
Exemple 2.6

Une personne est choisie au hasard dans une colonie de vacances. Soit Ω l'espace échantillon, $\Omega = \{\text{l'ensemble de tous les vacanciers}\}$. Soit A l'événement " la personne choisie s'est inscrite au tennis", B l'événement " la personne choisie s'est inscrite au golf". On peut identifier les probabilités $P(A)$ et $P(B)$ aux pourcentages de ceux qui jouent au tennis et au golf, respectivement; de même, la probabilité $P(A \cap B)$ est le pourcentage de personnes qui jouent au tennis et au golf. La probabilité conditionnelle de B étant donné A , $P(B|A)$, qui est égale à $P(A \cap B)/P(A)$, représente le pourcentage de ceux qui jouent au golf parmi ceux qui jouent au tennis.

Exemple 2.7 Tirages successifs sans remise

D'une urne qui contient 3 boules rouges et 5 blanches, on tire sans remise deux boules, l'une après l'autre. Soit $A =$ "la première boule est rouge" et $B =$ "la deuxième boule est blanche". Alors $P(A) = 3/8$. Quant à $P(B)$, on ne l'obtient pas immédiatement. Ce qu'on peut obtenir immédiatement, c'est $P(B|A)$, la probabilité que B se produise, sachant que A s'est produit. Si A s'est produit, la première boule tirée est rouge, il ne reste plus que 7 boules dans l'urne, dont deux sont rouges et 5 blanches; donc $P(B|A) = 5/7$ (voir la figure 2.3.1). De même, $P(B|A^C) = 4/7$.

Figure 2.3.1 Tirage de deux boules sans remise



Exemple 2.8 Afin de déterminer les intentions de vote de la population, 100 personnes ont été interviewées et on leur a demandé pour quel parti politique A, B, C, elles allaient voter. Les données sont exhibées dans le tableau ci-dessous.

Parti Sexe	A	B	C
Hommes	13	21	19
Femmes	20	8	19

Si on tire une personne au hasard dans ce groupe, déterminer les probabilités suivantes :

- a) La personne tirée vote pour A.
- b) La personne tirée vote pour A si on sait que c'est une femme.
- c) La personne tirée vote pour B ou C si on sait que c'est un homme
- d) La personne tirée est une femme si on sait qu'elle vote pour C.

Solution : Soient

A : la personne tirée vote pour A.

B : la personne tirée vote pour B.

C : la personne tirée vote pour C.

H : la personne tirée est un homme.

F : la personne tirée est une femme.

2.4 Indépendance d'événements

Deux événements A et B sont indépendants si la probabilité conditionnelle de B étant donné A est égale à la probabilité inconditionnelle de B , c'est-à-dire, si

$$P(B|A) = P(B).$$

Puisque $P(A \cap B) = P(A)P(B|A)$, cette condition est équivalente à

$$P(A \cap B) = P(A) P(B).$$

Et c'est plutôt cette égalité qui servira de définition formelle:

Définition: Deux événements A et B sont dits **indépendants** si et seulement si

$$P(A \cap B) = P(A) P(B)$$

Cette définition est équivalente à chacune des deux suivantes:

$$P(A|B) = P(A) \quad , \quad P(B|A) = P(B),$$

à condition que les probabilités conditionnelles impliquées soient définies, c'est-à-dire, que $P(B) \neq 0$ pour la première, et que $P(A) \neq 0$ pour la deuxième.

Remarques

(1) Ne pas confondre les notions d'événements incompatibles et événements indépendants!

(2) Si A et B sont indépendants, alors

- (i) A et B^C sont indépendants
- (ii) A^C et B sont indépendants
- (iii) A^C et B^C sont indépendants

Exemple 2.9 *Épreuves indépendantes*

On lance un dé deux fois. Quelle est la probabilité d'avoir un nombre inférieur à 3 suivi d'un nombre supérieur à 5?

Solution : Soit $A =$ « le premier lancer donne un nombre inférieur à 3 » et $B =$ « le deuxième lancer donne un nombre supérieur à 5 ». On cherche donc $P(A \cap B)$. Les événements A et B sont indépendants car ils correspondent à deux épreuves qui n'exercent pas d'influence l'une sur l'autre. Donc $P(A \cap B) = P(A)P(B)$, et puisque $P(A) = 1/3$, $P(B) = 1/6$, $P(A \cap B) = (1/3)(1/6) = 1/18$.

Dans plusieurs applications, les épreuves indépendantes prennent la forme de tirages successifs avec remise.

Exemple 2.10 *Indépendance et dépendance: tirages avec et sans remise*

D'une population qui contient 100 pièces fabriquées dont 12 sont défectueuses on tire successivement deux pièces. Soit $A =$ "la première pièce est défectueuse" et $B =$ "la deuxième pièce n'est pas défectueuse". Déterminer $P(A \cap B)$ en supposant que les tirages se font a) avec remise, b) sans remise.

Solution : Il est évident que $P(A) = 0,12$ et $P(B) = 0,88$.

a) Les événements A et B sont indépendants puisque les tirages se font avec remise et donc $P(A \cap B) = P(A)P(B) = 0,12 \times 0,88 = 0,1056$;

b) $P(A \cap B) = P(A)P(B|A) = (12/100)(88/99) = 0,1067$.

Exemple 2.11 On tire au hasard une personne d'une certaine population. Considérons les événements suivants:

A: La personne choisie a les yeux bleus

B: La personne choisie a les yeux bruns

C: La personne choisie a les cheveux blonds

D: La personne choisie est en faveur de la peine capitale pour tout meurtre

E: La personne choisie est en faveur de la peine capitale pour le meurtre d'un policier

Dire si les propositions suivantes sont vraies ou fausses. Discutez.

a) *A* et *B* sont indépendants

b) *A* et *B* sont incompatibles

c) $P(A \cup B) = 0$

d) $E \subset D$

e) $P(A|C) > P(A)$

f) *A* et *D* sont indépendants

g) *B* et *D* sont incompatibles

h) $P(D) > P(E)$

i) $P(D|E) = P(D)/P(E)$

j) $P(E|D) = 1$

k) $P(D \cup E) = P(E)$.

Généralisation de la notion d'indépendance à plusieurs événements

La notion d'indépendance se généralise à plusieurs événements A_1, \dots, A_n . Il faudrait, entre autres, que

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2) \dots P(A_n)$$

mais cela ne suffit pas.

On dit des événements A_1, A_2, \dots, A_n qu'ils sont **mutuellement indépendants** si pour $k = 2, \dots, n$, on a

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_k})$$

pour tout choix i_1, \dots, i_k de k entiers parmi les entiers $1, 2, 3, \dots, n$.

Exemple 2.12 *Indépendance de 3 événements*

Les événements A , B , et C sont mutuellement indépendants si et seulement si les conditions suivantes sont vérifiées

$$P(A \cap B) = P(A)P(B), P(A \cap C) = P(A)P(C),$$

$$P(B \cap C) = P(B)P(C),$$

ainsi que

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

Chapitre 3

Variables aléatoires discrètes

3.1 Variables aléatoires

Une variable aléatoire, généralement dénotée par une lettre majuscule comme X , Y , Z , est une caractéristique numérique des résultats d'une expérience. C'est une fonction qui fait correspondre un nombre à chaque élément de l'espace échantillon.

Définition

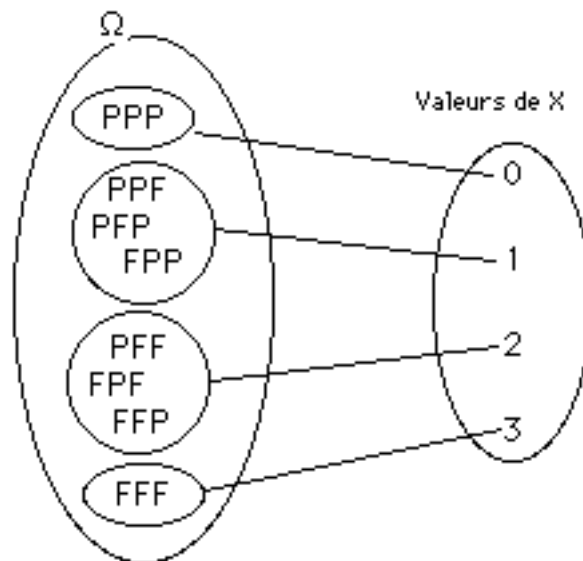
Une variable aléatoire est une fonction X qui fait correspondre à chaque élément $\omega \in \Omega$ un nombre $X(\omega)$.

Exemple 3.1 Le nombre X de faces lorsqu'on lance une pièce de monnaie trois fois, est une variable aléatoire: c'est une correspondance entre les éléments de l'espace échantillon et les valeurs de la variable aléatoire. La correspondance pour X est la suivante:

Ici, on a

$$\Omega = \{PPP, PPF, PFP, FPP, PFF, FFP, FPF, FFF\}$$

Correspondance entre les éléments de Ω et les valeurs de X



Si $\omega = PPP$, alors $X(\omega) = 0$ alors que si $\omega = FPP$, on a $X(\omega) = 1$.

Si on a déjà établi une probabilité sur Ω , il est aisé de la "transporter" à l'ensemble des valeurs de X . Supposons, par exemple, qu'on ait admis *l'équiprobabilité* des 8 résultats dans Ω . Alors,

$$P[X = 0] = P[\{(PPP)\}] = 1/8$$

$$P[X = 1] = P[\{(PPF),(PFP),(FPP)\}] = 3/8$$

$$P[X = 2] = P[\{(FFP), (FPF), (PFF)\}] = 3/8$$

$$P[X = 3] = P[\{(FFF)\}] = 1/8$$

Nous distinguerons deux types de variables aléatoires: les variables aléatoires **discrètes** et les variables aléatoires **continues**.

Dans ce chapitre, on considère le cas des variables aléatoires discrètes. Le cas des variables aléatoires continues sera traité au chapitre 4.

Variables aléatoires discrètes : ce sont celles dont les valeurs forment un ensemble fini, ou **infini dénombrable** (comme l'ensemble des entiers, par exemple).

Définition Le **support discret** d'une variable aléatoire X est l'ensemble \mathcal{D} des valeurs x dont la probabilité est non nulle:

$$\mathcal{D} = \{x \mid P(X = x) > 0\}$$

Parfois le nombre de valeurs est un nombre infini (dans la plupart des modèles qui traitent des files d'attente, par exemple, le nombre d'arrivées à un comptoir de service pendant un certain intervalle de temps est une variable qui prend les valeurs 0,1,2, ..., sans fin.)

Fonction de masse

Si x désigne une valeur quelconque d'une variable aléatoire X , il est théoriquement possible de calculer la probabilité que X prenne la valeur x , dénotée par $P[X = x]$ ou $p(x)$.

Définition La fonction de masse $p(x)$ d'une variable aléatoire discrète X est une fonction qui fait correspondre à chaque valeur x de X la probabilité que X prenne la valeur x :

$$p(x) = P[X = x]$$

Exemple 3.2 Nombre de FACE en trois lancers

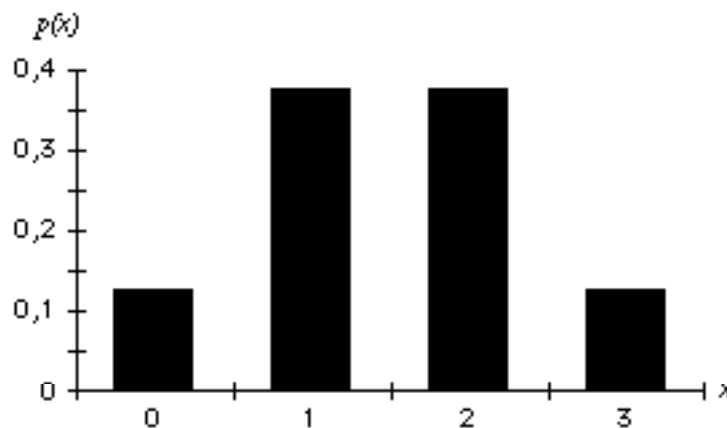
On lance trois pièces de monnaie. Soit X le nombre de FACE obtenus. Déterminer la fonction de masse p de X .

Solution : X est précisément la variable décrite au début de cette section et nous avons déjà déterminé sa fonction de masse, puisque nous avons calculé $P[X = x]$ pour $x = 0, 1, 2, 3$. Les valeurs x et les probabilités correspondantes $p(x)$ constituent la fonction de masse qui peut être présentée sous forme de tableau:

x	0	1	2	3	
$p(x)$	1/8	3/8	3/8	1/8	1

Figure 3.1

Distribution de X : nombre de FACE en trois lancers



Remarque Une fonction de masse n'est pas nécessairement symétrique comme celle de la figure ci-dessus.

Exemple 3.3 Deux dés: somme des résultats

On lance deux dés. Soit X la somme des deux résultats obtenus. Déterminer la fonction de masse p de X

Solution : Le support de X est l'ensemble $\{2, 3, \dots, 12\}$. On admet l'équiprobabilité des 36 résultats dans Ω .

x	2	3	4	5	6	7	8	9	10	11	12	
$p(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

Propriétés d'une fonction de masse

Une fonction de masse $p(x)$ doit satisfaire les conditions suivantes:

$$p(x) \geq 0 \text{ et } \sum_x p(x) = 1$$

Fonction de répartition

La fonction de répartition est une autre fonction associée à une variable aléatoire. En voici une définition:

Définition Soit X une variable aléatoire discrète. La *fonction de répartition* de X est une fonction F définie par

$$F(x) = P(X \leq x) \text{ pour tout } x \in \mathfrak{R}$$

La fonction de répartition donne pour chaque point x , la probabilité accumulée jusqu'à ce point:

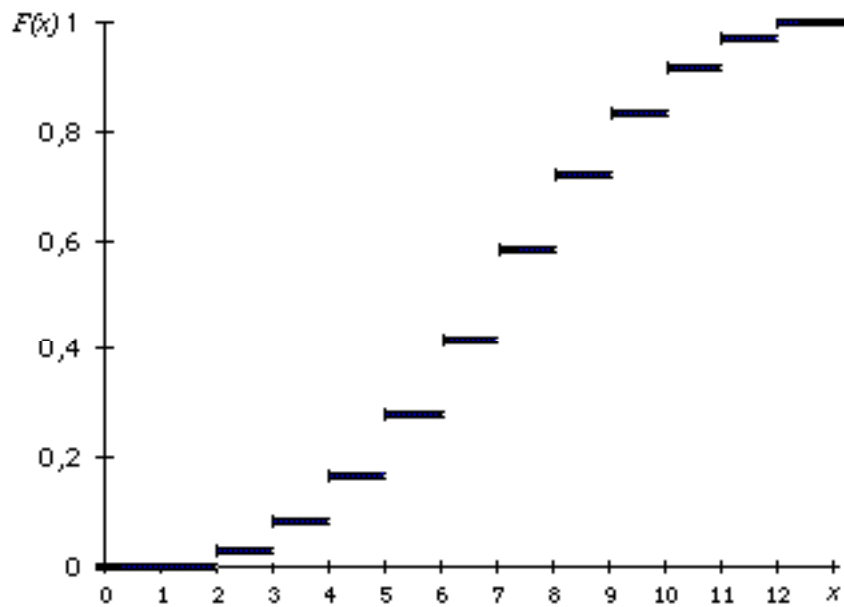
$$F(x) = \sum_{k \leq x} p(k)$$

Voici la fonction de répartition de la variable définie à l'exemple 3.3.

$$F(x) = \begin{cases} 0 & \text{si } x < 2 \\ 1/36 & \text{si } 2 \leq x < 3 \\ 3/36 & \text{si } 3 \leq x < 4 \\ 6/36 & \text{si } 4 \leq x < 5 \\ 10/36 & \text{si } 5 \leq x < 6 \\ 15/36 & \text{si } 6 \leq x < 7 \\ 21/36 & \text{si } 7 \leq x < 8 \\ 26/36 & \text{si } 8 \leq x < 9 \\ 30/36 & \text{si } 9 \leq x < 10 \\ 33/36 & \text{si } 10 \leq x < 11 \\ 35/36 & \text{si } 11 \leq x < 12 \\ 1 & \text{si } 12 \leq x \end{cases}$$

La figure 3.2 présente la fonction de masse de la variable définie à l'exemple 3.3.

Figure 3.2
*Fonction de répartition de
X: la somme des résultats obtenus en deux lancers d'un
dé*



Propriétés d'une fonction de répartition

Une fonction de répartition jouit des propriétés suivantes.

1. $0 \leq F(x) \leq 1$;
2. F est croissante;
3. $\lim_{x \rightarrow -\infty} F(x) = 0$ et $\lim_{x \rightarrow \infty} F(x) = 1$;
4. $F(x)$ est continue à droite.

3.2 - Espérance et variance d'une variable aléatoire

Définition Soit X une variable de fonction de masse p . Alors **l'espérance mathématique de X** , dénotée par $E(X)$, est définie par

$$E(X) = \sum_x xp(x)$$

L'espérance de X est aussi appelée **moyenne** de X . Habituellement, l'espérance mathématique d'une variable aléatoire X est dénotée par $E(X) \equiv \mu_X$ ou plus simplement par $E(X) \equiv \mu$.

Exemple 3.4 Une compagnie d'assurance offre une police d'annulation de voyage (seule cause d'annulation acceptée: maladie). La prime est de 72 \$ par personne; le coût pour la compagnie d'assurance est de 800 \$ lorsqu'il y a annulation. Supposons que, d'après les statistiques, la probabilité qu'un client tombe malade (et donc annule son voyage) est de 0,02. Déterminer l'espérance du gain de la compagnie lorsqu'elle assure une personne (négligez tous frais, sauf le versement de 800 \$ s'il y a lieu).

Solution :

Exemple 3.5 Le jeu de la roulette utilise un plateau comprenant 38 cases dont 36 sont numérotées 1, 2, ..., 36, et les deux dernières sont numérotées 0 et 00. Le croupier lance une boule dans le sens inverse de la rotation de la roulette qui va s'arrêter sur un numéro (le numéro gagnant!) de la roulette. On considère deux versions du jeu (en pratique, il existe plusieurs autres versions).

a) La version « **Straight Up** » : Vous misez 1 dollar sur un des numéros. Si vous gagnez, vous remportez 35 fois votre mise. Soit X votre gain. Déterminez $E(X)$.

b) La version « **Split** » : Vous misez 5 dollars sur deux numéros adjacents. Si vous gagnez, vous remportez 17 fois votre mise. Soit X votre gain. Déterminez $E(X)$.

Espérance mathématique d'une fonction de X

Soit Y une variable aléatoire définie par $Y = \varphi(X)$, où X est une variable discrète. Alors l'espérance de Y peut être calculée par la formule

$$E[\varphi(X)] = \sum_x \varphi(x) p(x)$$

Par exemple, si $\varphi(X) = X^2$, on a

$$E[X^2] = \sum_x x^2 p(x).$$

Ou encore, si $\varphi(X) = e^X$, on a

$$E[e^X] = \sum_x e^x p(x).$$

Interprétation de l'espérance

On lance un dé. Soit X le résultat obtenu. Il n'est pas difficile de montrer que $E(X) = 3.5$.

Quelle est l'interprétation de $E(X) = 3.5$?

Variance d'une variable aléatoire

L'espérance d'une variable aléatoire est un indice de la *position* de sa distribution. Nous définissons maintenant un indice de la *dispersion* d'une variable aléatoire.

Définition Soit X une variable de fonction de masse p et de moyenne μ .

Alors **la variance** de X , dénotée par $Var(X)$, est définie par

$$Var(X) = E[(X - \mu)^2] = \sum_x (x - \mu)^2 p(x)$$

Habituellement, la variance d'une variable aléatoire X est dénotée par σ^2 ou σ_x^2 .

Théorème 3.1 Soit X une variable aléatoire de moyenne μ . Alors

$$Var(X) = E(X^2) - \mu^2$$

Démonstration :

Définition L'écart-type d'une variable aléatoire X , dénotée par σ ou σ_X , est la racine carrée de sa variance:

$$\sigma_X = \text{Écart-type de } X = \sqrt{\text{Var}(X)}$$

Exemple 3.6 Dans l'exemple 3.5, on s'intéresse à déterminer la variance du gain pour les deux versions du jeu de roulette.

Solution :

Fonction affine d'une variable aléatoire

Si X est une variable aléatoire et a et b sont des constantes, alors $Y = a + bX$ est aussi une variable aléatoire.

Théorème 3.2 Soit X est une variable aléatoire et a et b des constantes. Alors

$$\mu_Y = E(a + bX) = a + bE(X) = a + b\mu_X$$

$$\sigma_Y^2 = \text{Var}(a + bX) = b^2 \text{Var}(X) = b^2 \sigma_X^2$$

$$\sigma_Y = |b| \sigma_X$$

Démonstration :

Exemple 3.7 *Espérance, variance et écart-type d'une fonction affine*

Vous donnez ordre à votre courtier de vous acheter 12 actions de la compagnie ABC au prix du marché X . Supposons que $\mu_X = 27$, $\sigma_X = 3$. Vous recevrez une facture dont le montant Y est la valeur de vos actions, plus une commission forfaitaire de 50 \$. Déterminer l'espérance et l'écart-type de Y .

Solution :

$Y = 50 + 12X$. Alors $\mu_Y = 50 + 12\mu_X = 50 + 12(27) = 374$ \$. $\sigma_Y = |12|\sigma_X = 12(3) = 36$ \$.

3.3 - Plusieurs variables

Un même contexte expérimental peut donner lieu à plusieurs variables aléatoires. Si, par exemple, on tire au hasard un ménage dans un quartier, on peut observer X , le revenu du ménage; ou Y , le nombre d'enfants; ou encore Z , le nombre de chambres à coucher. Il arrive également qu'on définisse de nouvelles variables aléatoires comme fonctions de variables observées. Le plus souvent, ce sont des sommes qu'on calculera, ou des fonctions linéaires. Par exemple, si X est le revenu du père de famille, Y celui de la mère, alors dans les enquêtes sociales ou économiques on s'intéressera particulièrement au revenu du couple, $Z = X + Y$. Que peut-on dire de l'espérance ou de la variance d'une somme de variables aléatoires? Il y a un théorème qui montre comment calculer l'espérance d'une fonction linéaire de variables aléatoires:

Théorème 3.3 Soient X_1, \dots, X_n n variables aléatoires de moyennes μ_1, \dots, μ_n . Soient a_1, \dots, a_n n constantes. Alors

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i) = \sum_{i=1}^n a_i \mu_i$$

Corollaires

- 1 $E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mu_i$: l'espérance d'une somme de variables aléatoires est égale à la somme des espérances.

- 2 Supposons que $\mu_1 = \dots = \mu_n = \mu$. Alors
 - a) $E\left(\sum_{i=1}^n X_i\right) = n\mu$.
 - b) Si $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ alors $E(\bar{X}) = \mu$.

- 3 Soit X et Y deux variables aléatoires, et a et b deux constantes. Alors

$$E(aX+bY) = aE(X) + bE(Y),$$

$$E(X+Y) = E(X) + E(Y)$$

$$E(X-Y) = E(X) - E(Y)$$

Il existe des résultats semblables pour la variance d'une fonction linéaire de variables aléatoires, mais nous les énoncerons ici dans un cas particulier, le cas où les variables sont **indépendantes**.

Théorème 3.4 Soit X_1, \dots, X_n n variables aléatoires **indépendantes** de moyennes μ_1, \dots, μ_n et de variances $\sigma_1^2, \dots, \sigma_n^2$. Soit a_1, \dots, a_n n constantes. Alors

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) = \sum_{i=1}^n a_i^2 \sigma_i^2$$

Corollaires

1. $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \sigma_i^2$: la variance d'une somme de variables aléatoires **indépendantes** est égale à la somme des variances.

2. Supposons que $\sigma_1^2 = \dots = \sigma_n^2 = \sigma^2$. Alors

a) $\text{Var}\left(\sum_{i=1}^n X_i\right) = n\sigma^2$.

b) Si $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, alors $\text{Var}(\bar{X}) = \sigma^2/n$.

3. Soit X et Y deux variables aléatoires **indépendantes**, et a et b deux constantes. Alors

a) $\text{Var}(aX+bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$

b) $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$

c) $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y)$

Exemple 3.8 On suppose que le poids (en kg) des adultes se distribue avec une moyenne de 64 et un écart-type de 12. Soit X le poids total de 14 personnes qui s'entassent dans un ascenseur. Calculez l'espérance mathématique et la variance de X .

Solution :

Exemple 3.9 Pour chacune des paires de variables aléatoires X et Y , dites si d'après vous $\sigma_X > \sigma_Y$ ou si $\sigma_Y > \sigma_X$:

a) X : La valeur d'une action que vous venez d'acheter, dans une semaine;

Y : La valeur d'une action que vous venez d'acheter, dans un an.

b) X : La température le 1^e janvier prochain à Montréal;

Y : La température le 1^e janvier prochain à Nairobi.

c) X : Le poids d'une personne choisie au hasard dans une école de garçons;

Y : Le poids d'une personne choisie au hasard dans une école mixte.

d) X : Le temps que vous mettez à vous rendre à l'université à pied;

Y : Le temps que vous mettez à vous rendre à l'université en métro.

e) X : Le temps d'attente dans une file où il n'y a qu'une personne devant vous;

Y : Le temps d'attente dans une file où il y a 2 personnes devant vous.

f) X : La proportion d'objets défectueux dans un échantillon de 10 objets tirés d'une certaine population;

Y : La proportion d'objets défectueux dans un échantillon de 100 objets tirés d'une certaine population.

g) X : Le nombre d'objets défectueux dans un échantillon de 10 objets tirés *sans* remise d'une certaine population;

Y : Le nombre d'objets défectueux dans un échantillon de 10 objets tirés *avec* remise d'une certaine population;

h) X : Le revenu moyen de 10 familles choisies au hasard dans une population;

Y : Le revenu moyen de 100 familles choisies au hasard dans une population.

Théorème 3.5 Soit X_1, \dots, X_n n variables aléatoires indépendantes de moyennes μ_1, \dots, μ_n Alors

$$E(X_1 \times X_2 \times \dots \times X_n) = E(X_1) \times E(X_2) \times \dots \times E(X_n)$$

3.4 – Une loi discrète importante : la loi binomiale

On constate dans les applications classiques de la statistique que des expériences assez diverses peuvent avoir un noyau théorique commun — un ensemble de caractéristiques essentielles qui rend possible des **groupements de variables de même loi**, c'est-à-dire des variables dont la fonction de masse est mathématiquement de même forme. Dans ce chapitre, nous étudions deux lois importantes : **la loi binomiale et la loi multinomiale** (voir section 3.5).

La loi binomiale s'applique à toute expérience qui satisfait les conditions suivantes:

(i) Elle est composée d'une suite de n épreuves indépendantes.

(ii) Chaque épreuve peut donner lieu à deux résultats, « succès » et « échec ». Ces épreuves sont souvent appelées « épreuves de Bernoulli ».

(iii) La probabilité p de succès à chaque épreuve reste fixe.

Si X est le nombre de succès obtenus au cours d'une telle expérience, alors X est de loi binomiale de paramètres n et p .

On écrit $X \sim \mathfrak{B}(n ; p)$ pour signifier: « X suit une loi binomiale de paramètres n et p ».

Exemples 3.10

1. On tire un échantillon de 15 pièces dans un lot de pièces fabriquées. X est le nombre de pièces défectueuses.
2. Dans un sondage d'opinion on interroge 500 personnes choisies au hasard dans une population. X est le nombre de ceux qui répondent « oui » à la question « Êtes-vous en faveur d'un enseignement religieux dans les écoles? ».
3. On observe 25 naissances dans un hôpital. X est le nombre de garçons parmi les nouveau-nés.
4. On teste une nouvelle pilule auprès de 15 personnes souffrant de migraines. X est le nombre de sujets qui ont trouvé la pilule efficace.

Si $X \sim \mathfrak{B}(n ; p)$ alors la fonction de masse de X est donnée par

$$p(x) = P\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x},$$
$$x = 0, 1, \dots, n.$$

où

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Théorème 3.6 Si $X \sim \mathfrak{B}(n ; p)$, alors

$$E[X] = np \text{ et } \text{Var}[X] = np(1-p) = npq.$$

Démonstration :

Exemple 3.11

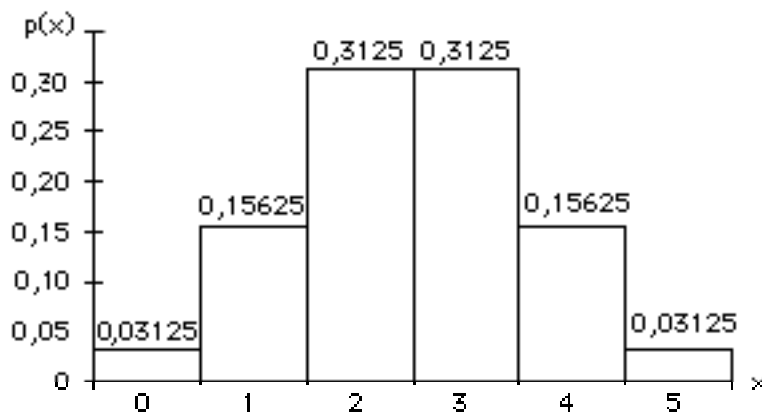
On choisit au hasard une famille, parmi les familles ayant 5 enfants. Quelle est la probabilité qu'il y ait moins de deux filles dans cette famille?

Solution : En supposant l'indépendance entre les naissances (relativement au sexe de l'enfant), le nombre de filles dans une famille de cinq enfants est une $\mathfrak{B}(5 ; 0,5)$. Par conséquent, la probabilité cherchée est:

$$P(X \leq 1) = P\{X = 0\} + P\{X = 1\} = \binom{5}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 + \binom{5}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^4 = 6/32.$$

Figure 3.2

Fonction de masse d'une variable $X \sim \mathfrak{B}(5 ; 0,5)$



Exemple 3.12 Détermination de n

Les 25 employés d'un certain bureau organisent une loterie. Ils sont numérotés de 1 à 25 et chaque semaine un numéro est choisi au hasard parmi les nombres de 1 à 25. L'employé qui porte ce numéro gagne un prix de 25 \$. Jean se demande combien de semaines la loterie doit durer afin qu'il ait au moins 70% des chances de gagner le prix durant cette période.

Solution : Soit n le nombre de semaines où les employés feront cette loterie. Durant cette période, Jean peut gagner 0, 1, ..., ou n fois. Évaluons la probabilité que Jean ne gagne aucune fois durant cette période. Par la formule vue précédemment, on a

$$P(X=0) = \binom{n}{0} \left(\frac{1}{25}\right)^0 \left(\frac{24}{25}\right)^n = \left(\frac{24}{25}\right)^n < 0,30$$

On veut donc que $n > \frac{\ln(0,30)}{\ln(24/25)} = 29,4933$. La loterie devra donc durer au moins trente semaines pour que Jean ait au moins 70% des chances de gagner une fois.

Exemple 3.13 : Dans un village où ont été entreposés des déchets chimiques, on constate que 8 personnes ont été atteintes d'une certaine sorte de cancer dans une période de 5 ans. Étant donné que la population du village n'est que de 8 000, ce nombre semble excessif. Une commission chargée de déterminer si les déchets chimiques ont contribué à hausser le taux prélevé des données sur les populations de plusieurs villages de taille et situation comparables. La commission découvre que durant la même période, il y a eu 588 cas dans un bassin de population de 2 350 000 habitants. Considérer ce taux comme un taux normal (et connu sans erreur) pour calculer la probabilité d'avoir 8 cas ou plus dans une population de 8 000. Expliquer ce que ce calcul peut contribuer à la question posée par la commission.

Solution :

Exemple 3.14 : Depuis 1988, le gouvernement américain a exigé la peine de mort dans 36 cas impliquant des trafiquants de drogues. Des 36 trafiquants, 4 seulement étaient de race blanche. On sait par ailleurs que 75% des accusés (de trafic de drogues) sont blancs. Donner un argument aussi complet que possible pour démontrer qu'il y a une discrimination systémique en faveur des blancs. *Les données proviennent des promoteurs d'un projet de loi appelé The Racial Justice Act qui a été à l'étude au Congrès américain. Le but de ce projet était de permettre à une personne de race noire condamné à mort de présenter pour sa défense une argumentation semblable à celle que vous donnerez. Le projet n'a pas été adopté.*

Solution :

3.5 – Une généralisation la loi binomiale : la loi multinomiale

L'expérience de Bernoulli est une expérience où les seuls résultats possibles sont $\{Succès; \acute{E}chec\}$. Une loi binomiale est constituée d'expériences de Bernoulli. On peut facilement s'imaginer des contextes où les résultats sont plus nombreux.

Par exemple, dans une enquête telle que celle sur la promotion de la santé faite par Statistique Canada, on demande aux gens d'indiquer leur activité principale au cours des douze derniers mois. Les réponses possibles sont: $\{Travailleur; \acute{A} \text{ la recherche d'un emploi; \acute{E}tudiant; Retraité; Ménagère; Autre\}$. Il y a donc 6 résultats possibles. On peut dénoter par p_1 la probabilité qu'un individu réponde qu'il est «travailleur», par p_2 la probabilité qu'il soit à la recherche d'un emploi, ..., et par p_6 la probabilité qu'il ait une autre activité.

Notons qu'un individu donné ne peut choisir qu'une seule occupation. Ainsi, à un individu i donné, on peut associer un vecteur \mathbf{X}_i de 6 composantes, où apparaît un 1 dans la position 1 et des 0 ailleurs s'il est travailleur, un 1 en position 2 et des 0 ailleurs s'il est à la recherche d'un emploi, etc. Il s'agit d'une simple extension de l'expérience de Bernoulli. Si nous interrogeons n personnes afin de connaître leur occupation, on pourra déterminer combien parmi elles travaillent, combien étudient, etc. Nous pouvons présenter cette information sous la forme d'un vecteur: $(x_1; x_2; \dots; x_k)$. Cette notation signifie que parmi les n individus interrogés, x_1 se déclarent travailleurs, x_2 se considèrent étudiants, etc.

Une expérience aléatoire est une **expérience de Bernoulli généralisée** si, et seulement si, elle conduit à k résultats possibles. Les k résultats possibles sont, par convention, $\{R_1; R_2; \dots; R_k\}$ et nous dénotons par $p_k = P\{R_k\}$. On dénote le résultat d'une expérience de Bernoulli généralisée par un vecteur ayant des zéros partout sauf en position i , où nous plaçons un 1 signifiant que R_i s'est réalisé.

La fonction de masse de loi multinomiale fait intervenir le coefficient multinomial défini de la façon suivante:

$$\binom{n}{x_1, \dots, x_k} = \frac{n!}{x_1! \dots x_k!}$$

Notons que $\binom{n}{n-x} = \frac{n!}{x!(n-x)!} = \binom{n}{x}$

de sorte que le coefficient binomial n'est autre qu'un coefficient multinomial avec $k = 2$.

Définition Une variable **multinomiale** est une variable aléatoire (vectorielle), $\mathbf{X} = (X_1, X_2, \dots, X_k)$, comptant le nombre d'occurrences de chacune des k catégories de résultat dans une suite de n expériences de Bernoulli généralisées indépendantes et où la probabilité d'appartenir à la i^e catégorie est p_i . Chacune des composantes prend sa valeur dans l'ensemble $\{0; 1; \dots; n\}$. Notons que $x_1 + x_2 + \dots + x_k = n$. De plus, on a

$$P(\mathbf{X} = (x_1, x_2, \dots, x_k)) = \binom{n}{x_1 \ x_2 \ \dots \ x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

Nous dénoterons ce fait par

$$\mathbf{X} = (X_1, X_2, \dots, X_k) \sim MN(n; p_1; p_2; \dots; p_k).$$

Exemple 3.15 Dans une certaine province trois partis sont en lice. Si dans la population 50% des gens favorisent le parti A, 40% le parti B, et 10% le parti C, calculons la probabilité que sur 6 personnes choisies au hasard dans cette province, 3 favorisent A, 1 favorise B et 2 favorisent C.

Solution : Ici, $\mathbf{X} \sim MN(6; 0,5; 0,4; 0,1)$.

Par conséquent,

$$\begin{aligned} P[\mathbf{X} = (3, 1, 2)] &= \binom{n}{3 \ 1 \ 2} (0,5)^3 (0,4)^1 (0,1)^2 \\ &= \frac{6!}{3!1!2!} \times 0,125 \times 0,4 \times 0,01 = 0,03 \end{aligned}$$

Remarque : Lorsque $\mathbf{X} = (X_1, X_2, \dots, X_k) \sim MN(n; p_1; p_2; \dots; p_k)$, chacune des composantes X_j du vecteur \mathbf{X} suit une loi binomiale de paramètres n et p_j . Autrement dit, on a $X_j \sim B(n, p_j)$, $j = 1, \dots, k$. Il s'ensuit que $E(X_j) = np_j$.

Chapitre 4

Variables aléatoires continues

4.1 Fonctions de répartition et de densité

Pour les variables aléatoires discutées au chapitre précédent, on peut identifier une série (finie ou infinie) de valeurs isolées x_1, x_2, \dots auxquelles correspondent, par la fonction de masse, des probabilités $P(X = x_1), P(X = x_2), \dots$. Ce sont des variables **discrètes** ou **discontinues**. Les variables aléatoires **continues**, en revanche, prennent une infinité de valeurs que l'on ne peut dénombrer. Typiquement, l'ensemble des valeurs d'une variable aléatoire continue est un intervalle (borné ou non) de nombres réels. Pour ces variables, il n'existe pas de fonction de masse. Il existe toujours, cependant, une fonction de répartition:

Exemples de variables aléatoires continues :

- (1) Soit X le poids (en kgs) d'un individu tiré au hasard dans une population.
- (2) Soit Y la durée de vie d'une ampoule tirée au hasard dans un lot d'ampoules.

Fonction de répartition

La fonction de répartition $F : \mathbb{R} \rightarrow [0 ; 1]$, est définie, dans le cas continu comme dans le cas discret, par

$$F(x) = P[X \leq x], \quad x \in \mathbb{R}$$

Une fonction de répartition jouit des propriétés suivantes.

1. $0 \leq F(x) \leq 1$;
2. F est non décroissante;
3. $\lim_{x \rightarrow -\infty} F(x) = 0$ et $\lim_{x \rightarrow \infty} F(x) = 1$;
4. $F(x)$ est continue à droite.

Fonction de densité

Une variable aléatoire X est dite *continue* si elle possède une *fonction de densité*, c'est-à-dire, s'il existe une fonction f telle que pour $x \in \mathbb{R}$.

$$F(x) = \int_{-\infty}^x f(t) dt$$

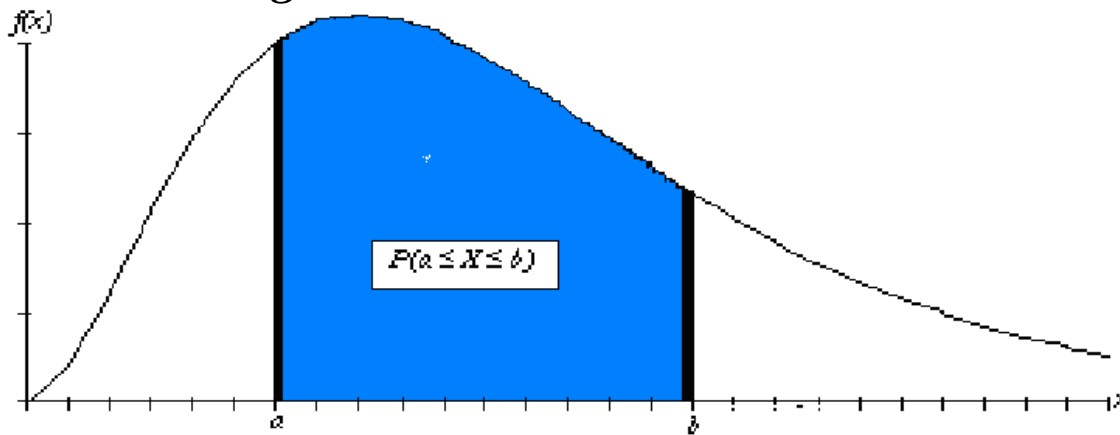
où f satisfait les conditions suivantes:

1. $f(x) \geq 0, \forall x \in \mathbb{R}$;

2. f admet au plus un nombre fini de discontinuités sur chaque intervalle fini de \mathbb{R} ;

3.
$$\int_{-\infty}^{\infty} f(t)dt = 1.$$

Figure 4.1 *Fonction de densité*



Si $a \leq b$, $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Si f est continue, alors f est la dérivée de F : $f(x) = \frac{d}{dx} F(x)$

Remarque Dans un modèle continu, un point unique a toujours probabilité nulle: $P(X = x) = 0$ pour tout x . Par conséquent, toute inégalité stricte peut être remplacée par une égalité non stricte, et inversement, sans que la probabilité change. Ainsi, si X est continue, $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$

Remarque *Relation entre une fonction de masse et une fonction de densité*

Le premier graphique ci-dessous (Figure 4.2) présente la distribution des tailles (en pouces) d'un groupe d'adultes. Les tailles sont arrondies à l'entier le plus proche; il s'agit donc d'une variable discrète qui ne prend que des valeurs entières comme 64, 65, 66, etc.

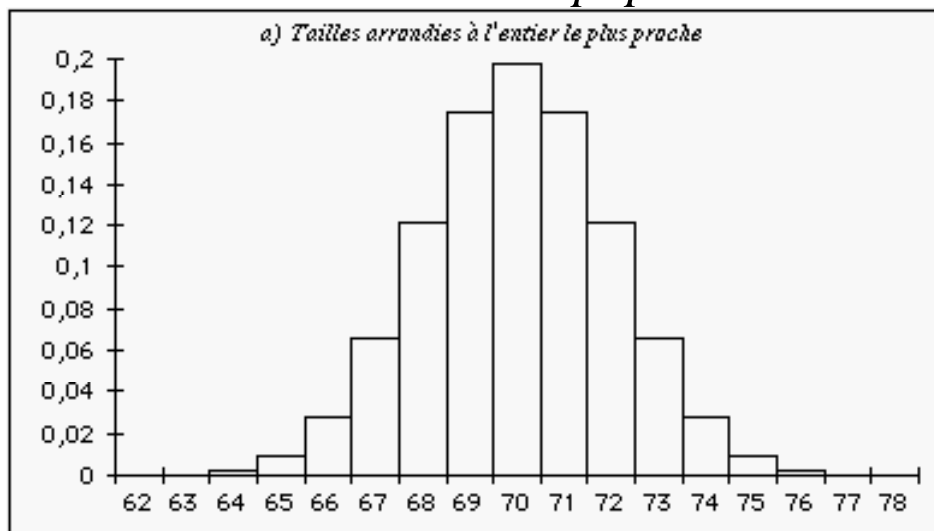
Dans la représentation graphique, les probabilités, indiquées sur l'axe vertical, sont proportionnelles aux hauteurs des rectangles. Mais puisque les bases sont de même largeur, les probabilités sont également proportionnelles aux aires des rectangles. Nous voulons préserver cette propriété des représentations par histogramme.

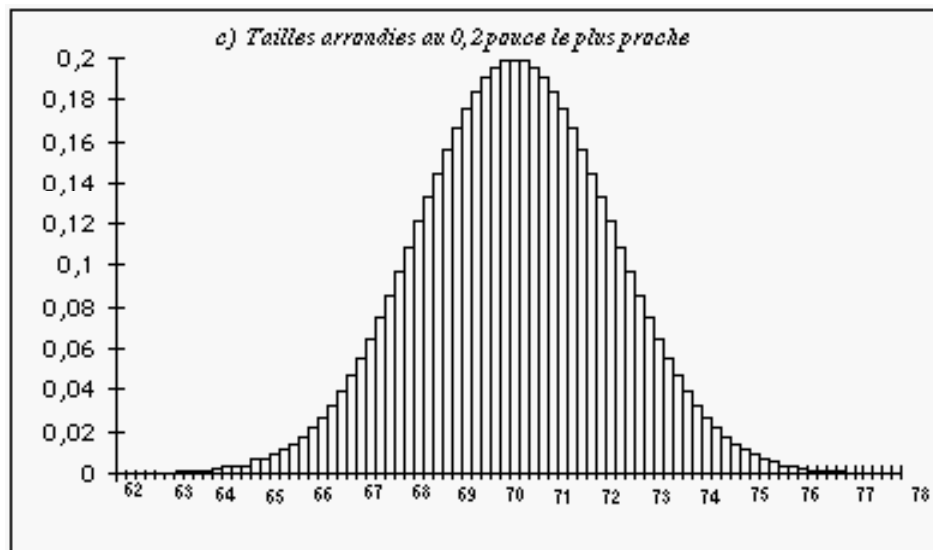
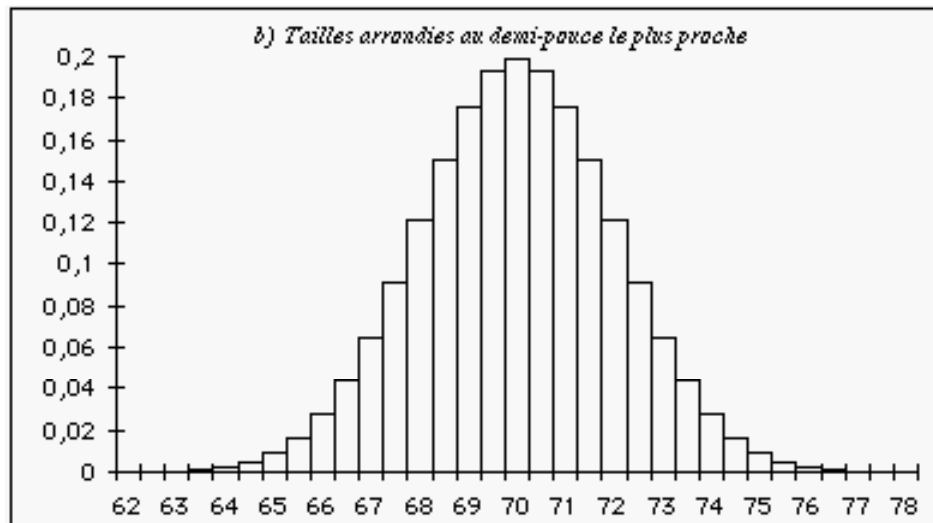
Dans le deuxième graphique, les tailles sont arrondies au demi-pouce le plus proche. Il s'agit donc d'une variable discrète encore, prenant les valeurs 64; 64,5; 65; 65,5, etc. On a gardé la même échelle, mais cette fois-ci, la hauteur d'un rectangle, indiquée sur l'axe vertical, représente la probabilité divisée par la largeur (0,5) de l'intervalle; donc la hauteur représente la densité de probabilité et non la probabilité comme telle. Ici aussi, la probabilité d'une valeur est représentée par l'aire d'un rectangle, c'est-à-dire, par la densité multipliée par la base.

Dans le troisième graphique, le processus se poursuit avec une variable dont l'écart entre les valeurs est de 0,2 pouces. On conçoit bien qu'à la limite l'histogramme converge vers une courbe dans laquelle la probabilité d'un intervalle est donnée par l'aire de la surface sous la courbe au-dessus de l'intervalle.

Figure 4.2 *Une fonction de densité est une limite d'histogrammes*

Distribution des tailles d'une population d'adultes





Espérance et variance d'une variable aléatoire continue

On remplace simplement les sommes (utilisées dans le cas de variable aléatoires discrètes) par des intégrales.

Si X est une variable aléatoire continue, alors

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

et

$$V(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx$$

Les propriétés des espérances et variances (sommations, etc.) restent valides dans le cas continu grâce à la propriété de linéarité des intégrales.

4.3 Loi normale

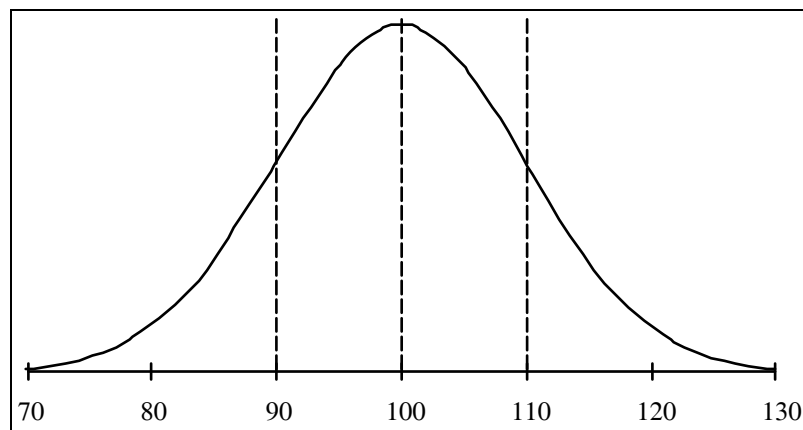
La **loi normale** est l'une des lois les plus importantes de la statistique. Non seulement permet-elle de modéliser des variables courantes dans le quotidien et dans les sciences, telles la taille et parfois le poids d'individus, mais elle joue un rôle capital au niveau de l'inférence statistique. Une définition, en termes de la fonction de densité, suit:

Définition: Une variable aléatoire X est distribuée selon une loi normale de paramètres μ et σ^2 , notée $X \sim \mathfrak{N}(\mu ; \sigma^2)$, si elle a pour densité la fonction

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)(x-\mu)^2/\sigma^2}$$

La figure 4.3 présente le graphique de la densité d'une loi $\mathfrak{N}(100 ; 100)$. Une densité normale est une courbe symétrique par rapport à sa moyenne et présente des points d'inflexion de part et d'autre de celle-ci, à une distance d'un écart-type.

Figure 4.3 Loi normale $\mathfrak{N}(100; 100)$



On peut vérifier que $\int_{-\infty}^{\infty} f(x)dx = 1$

Théorème 4.1 Si $X \sim \mathfrak{N}(\mu ; \sigma^2)$, alors $E(X) = \mu$ et $Var(X) = \sigma^2$.

Théorème 4.2 Toute fonction linéaire d'une variable normale est normale

Si $X \sim \mathfrak{N}(\mu ; \sigma^2)$ et $Y = a + bX$ alors

$$Y \sim \mathfrak{N}(a + b\mu ; b^2 \sigma^2)$$

Corollaire Si $X : \mathfrak{N}(\mu ; \sigma^2)$, alors

$$Z = \frac{X - \mu}{\sigma} \sim \mathfrak{N}(0 ; 1)$$

Comment calculer $P(a < X < b)$ si $X \sim \mathfrak{N}(\mu ; \sigma^2)$?

On peut bien sûr appliquer la définition en obtient

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)(x-\mu)^2/\sigma^2} dx$$

Cette manière de faire, est tout à fait valide, bien que relativement complexe.

Calculs de probabilités

En pratique, on va plutôt transformer X en Z et utiliser des tables fournissant les aires sous une courbe normale particulière, la courbe $\mathfrak{N}(0 ; 1)$.

Exemple 4.1 Calcul des probabilités pour une normale centrée-réduite

Soit $Z \sim \mathcal{N}(0 ; 1)$. Calculer

- a) $P[Z > 1,25]$;
- b) $P[Z \leq -1]$;
- c) $P[1,15 < Z \leq 2,11]$;
- d) $P[0 < Z \leq 1]$;
- e) $P[-2 < Z \leq 1]$,

Solution :

Pour des variables normales de moyenne et variance quelconques, il faut pouvoir «ramener» une normale arbitraire $\mathfrak{N}(\mu ; \sigma^2)$ à une $\mathfrak{N}(0 ; 1)$.

Exemple 4.2 *Calcul des probabilités pour une normale quelconque*

Supposons que les montants correspondant à une population de factures sont de moyenne $\mu = 200$ \$ et d'écart-type $\sigma = 80$ \$. En supposant que les montants des factures sont de loi normale, déterminer la probabilité qu'une facture tirée au hasard corresponde à un montant compris entre 40 \$ et 280 \$.

Solution :

Variables normales indépendantes

Considérons des variables aléatoires indépendantes X_1, \dots, X_n . Le théorème suivant affirme que si elles sont toutes normales, alors toute *combinaison linéaire* — et donc en particulier leur somme — est normale. En voici une formulation précise.

Théorème 4.3 Soit X_1, \dots, X_n n variables aléatoires indépendantes, $X_i \sim \mathfrak{N}(\mu_i; \sigma_i^2)$, $i = 1, \dots, n$, et a_1, \dots, a_n des constantes. Si $X = \sum_{i=1}^n a_i X_i$, alors

$$X \sim \mathfrak{N}\left(\sum_{i=1}^n a_i \mu_i; \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

Cas particuliers importants du Théorème 4.3 :

(1) Les constantes a_i sont toutes égales à 1. Dans ce cas, X n'est que la somme

$$T = \sum_{i=1}^n X_i ;$$

(2) Les constantes a_i sont toutes égales à $1/n$. Dans ce cas, X est la moyenne arithmétique : $\bar{X} = \sum_{i=1}^n X_i / n$.

Nous obtenons le corollaire suivant en substituant 1 et $1/n$ à a_i dans le résultat du théorème 4.3 :

Corollaire Si X_1, \dots, X_n sont indépendantes, chacune de même loi $\mathfrak{N}(\mu ; \sigma^2)$, alors

$$T \sim \mathfrak{N}(n\mu ; n\sigma^2) \text{ et } \bar{X} \sim \mathfrak{N}\left(\mu ; \frac{\sigma^2}{n}\right)$$

Exemple 4.3 On prélève un échantillon de $n = 15$ factures d'une très grande population de moyenne $\mu = 300$ \$ et écart-type $\sigma = 60$ \$. Quelle est la probabilité que la valeur moyenne de l'échantillon se situe à 10 \$ ou moins de la moyenne de la population)? On suppose que la population est normale.

Solution :

4.4 Théorème limite central

Dans la section précédente, nous avons vu que si les variables aléatoires indépendantes X_1, \dots, X_n sont de loi normale, alors leur moyenne \bar{X} est de loi normale. Ce théorème est utile pour traiter des problèmes d'échantillonnage, dans lesquels la moyenne d'un échantillon est utilisée pour estimer la moyenne d'une population. Mais l'hypothèse que chacune des variables X_1, \dots, X_n est elle-même normale signifie que la distribution de la population échantillonnée est normale. Cette hypothèse est plutôt restrictive: **les populations sont rarement normales.**

Il existe, cependant, un théorème fondamental qui a pour remarquable conclusion que la somme et la moyenne de n variables aléatoires indépendantes suivent approximativement une loi normale, **même si la population n'est pas normale**: la seule condition exigée est que n soit assez grand. Il s'agit du **Théorème Limite Central (TLC)**.

Afin de mieux comprendre le TLC, considérons les exemples suivants :

Exemple 4.4 On lance un dé n fois. On obtient donc un jeu de données composé de n observations. Si X désigne le résultat obtenu lors d'un lancer, sa fonction de masse est donnée par :

Lançons le dé $n = 10$ fois. On obtient alors un échantillon avec 10 observations. Répétons cette expérience (lancer un dé 10 fois) 50 fois. On obtient donc 50 échantillons, chacun composé de 10 observations.

Dans chaque échantillon (aléatoire), on calcule la moyenne échantillonnale \bar{X} . Notons que \bar{X} est une variable aléatoire puisque l'échantillon est aléatoire.

Question Puisque \bar{X} est une variable aléatoire, quelle est sa distribution? Cette distribution est appelée **distribution d'échantillonnage**.

Tableau 4.1 50 lancers de 10 dés

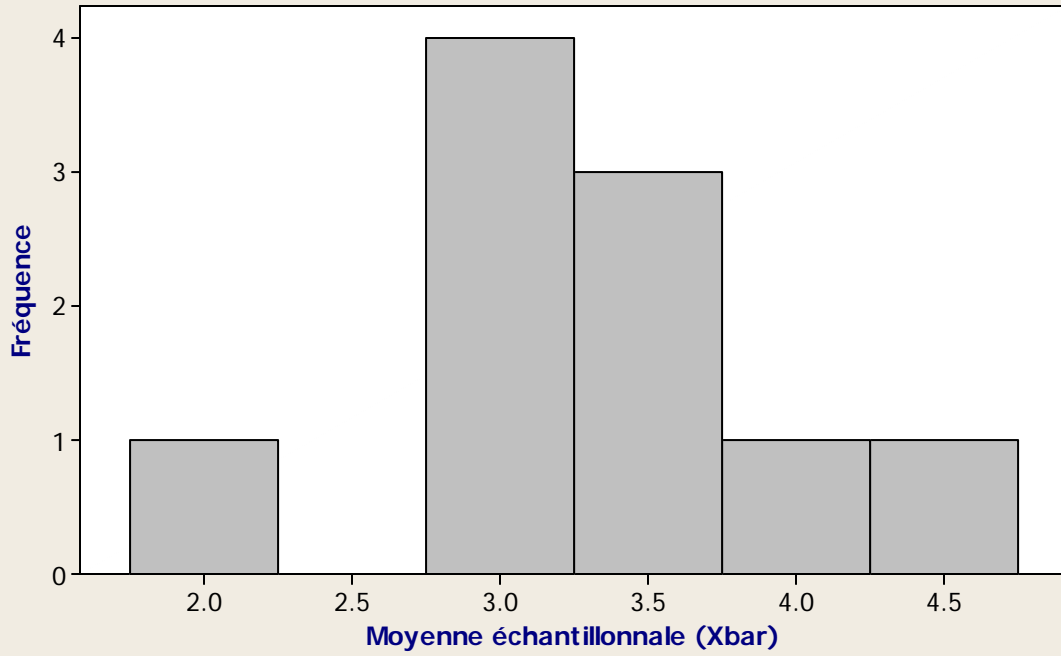
Échantillon no.	Observations	Moyenne échantillonnale
1	$(x_1^{(1)}, x_2^{(1)}, \dots, x_{10}^{(1)})$	\bar{X}_1
2	$(x_1^{(2)}, x_2^{(2)}, \dots, x_{10}^{(2)})$	\bar{X}_2
3	$(x_1^{(3)}, x_2^{(3)}, \dots, x_{10}^{(3)})$	\bar{X}_3
.	.	.
.	.	.
50	$(x_1^{(50)}, x_2^{(50)}, \dots, x_{10}^{(50)})$	\bar{X}_{50}

Les figures suivantes montrent la distribution de \bar{X} pour R échantillons de taille $n=10$ et $n=30$. Ici, $R=10;100;1000$ et 10000 .

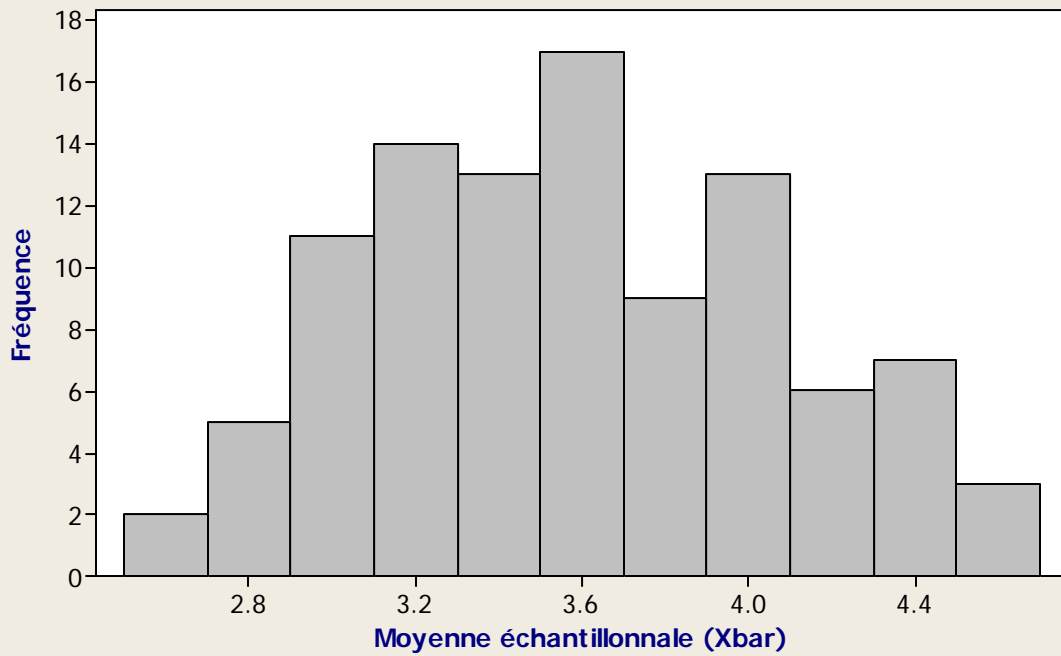
On voit que pour $n=10$, la distribution de \bar{X} tend vers une distribution en forme de cloche centrée en 3,5 à mesure que R augmente. Par conséquent, la distribution de \bar{X} tend vers une loi normale avec $n=10$. Bien sûr, ce résultat reste vrai lorsqu'on augmente la taille de l'échantillon (par exemple, $n=30$ et $n=100$).

Remarque Dans cet exemple, on avait commencé par générer des échantillons à partir d'une distribution uniforme (i.e., les résultats d'un dé). Si on avait commencé avec une autre distribution, est-ce que la distribution de \bar{X} aurait eu le même comportement?

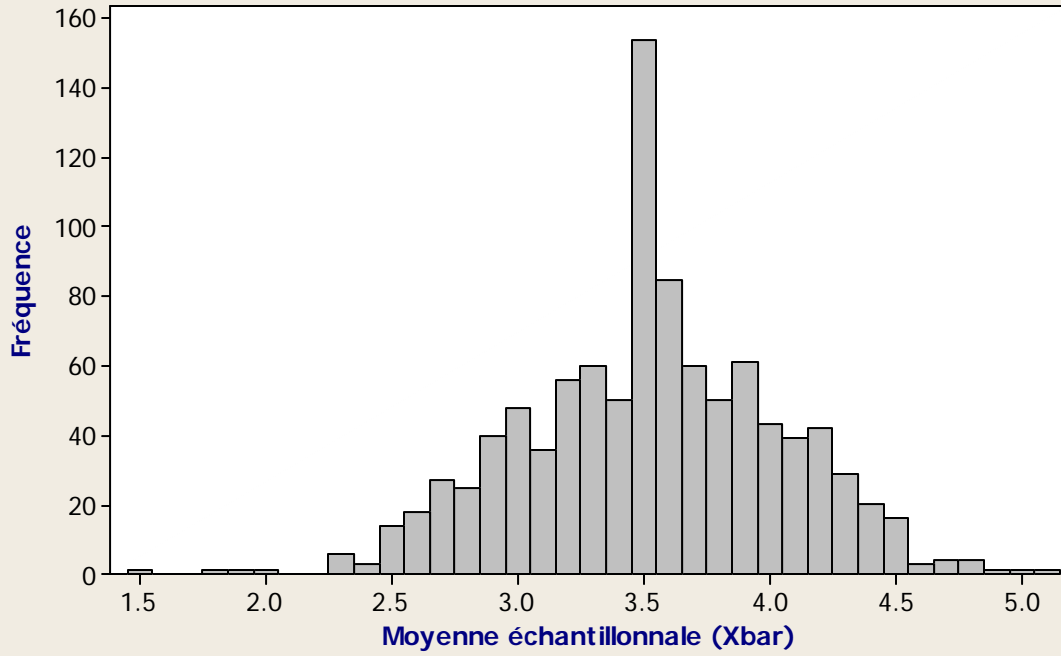
Histogramme pour 10 échantillons avec $n = 10$



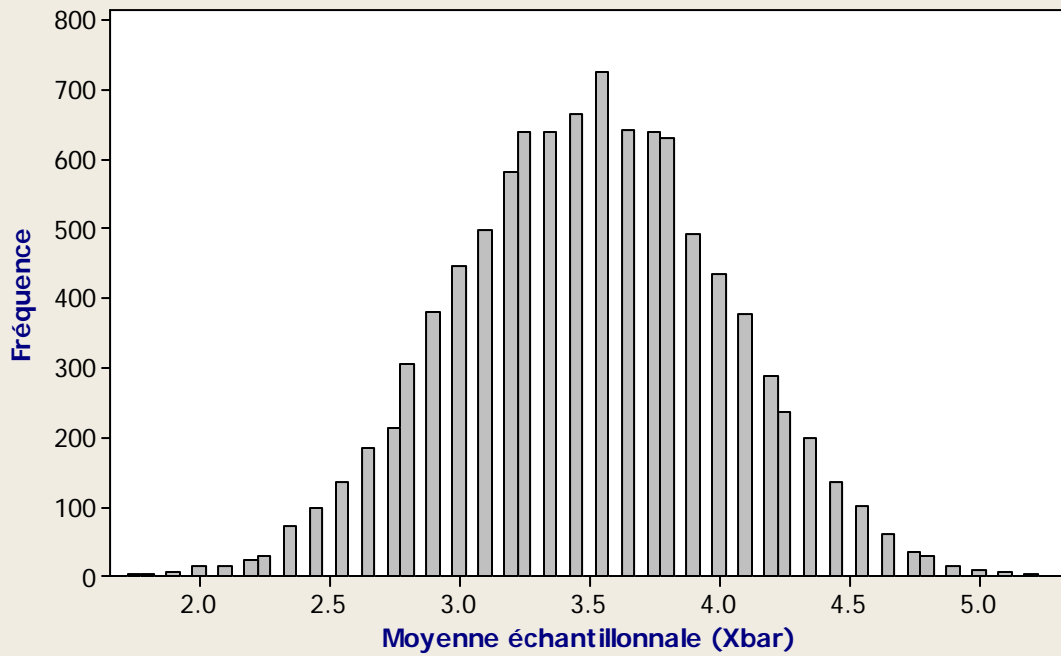
Histogramme pour 100 échantillons avec $n = 10$



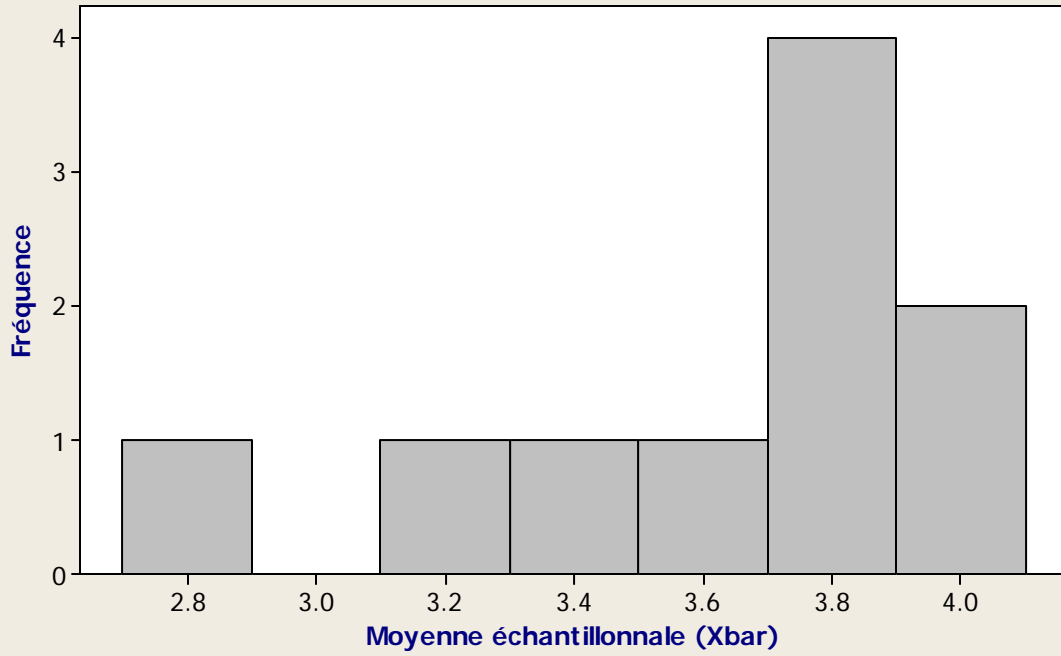
Histogramme pour 1000 échantillons avec $n = 10$



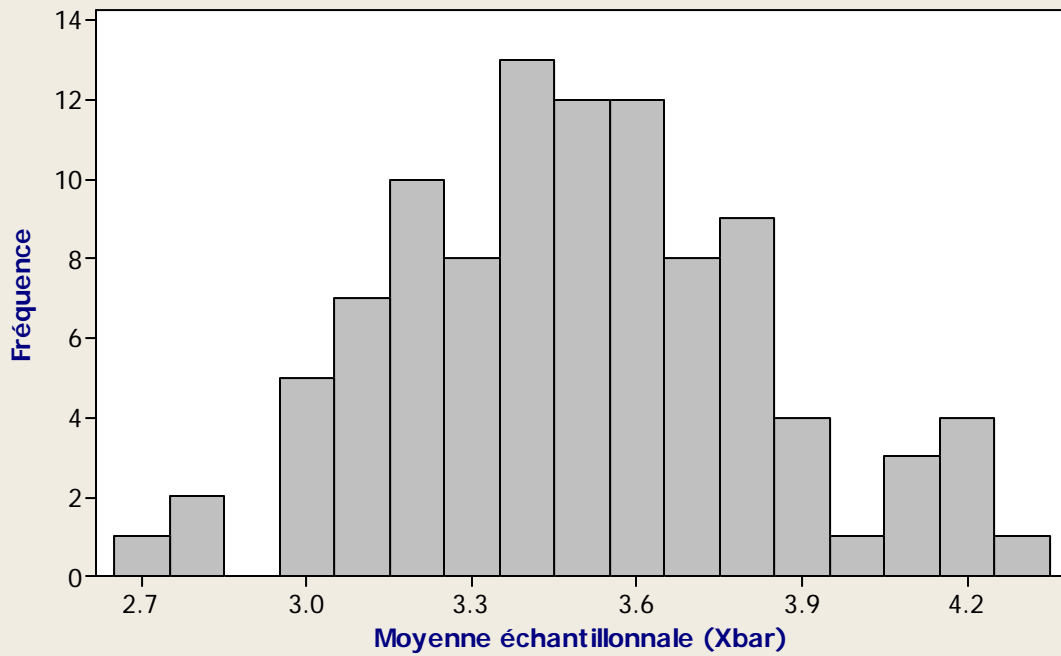
Histogramme pour 10000 échantillons avec $n = 10$



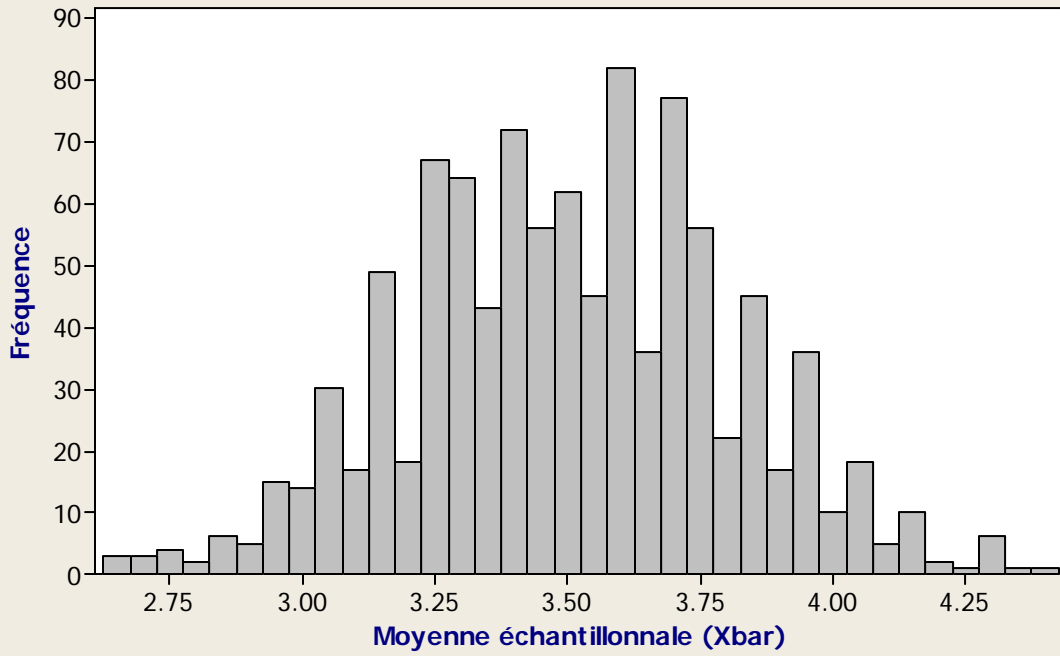
Histogramme pour 10 échantillons avec $n = 30$



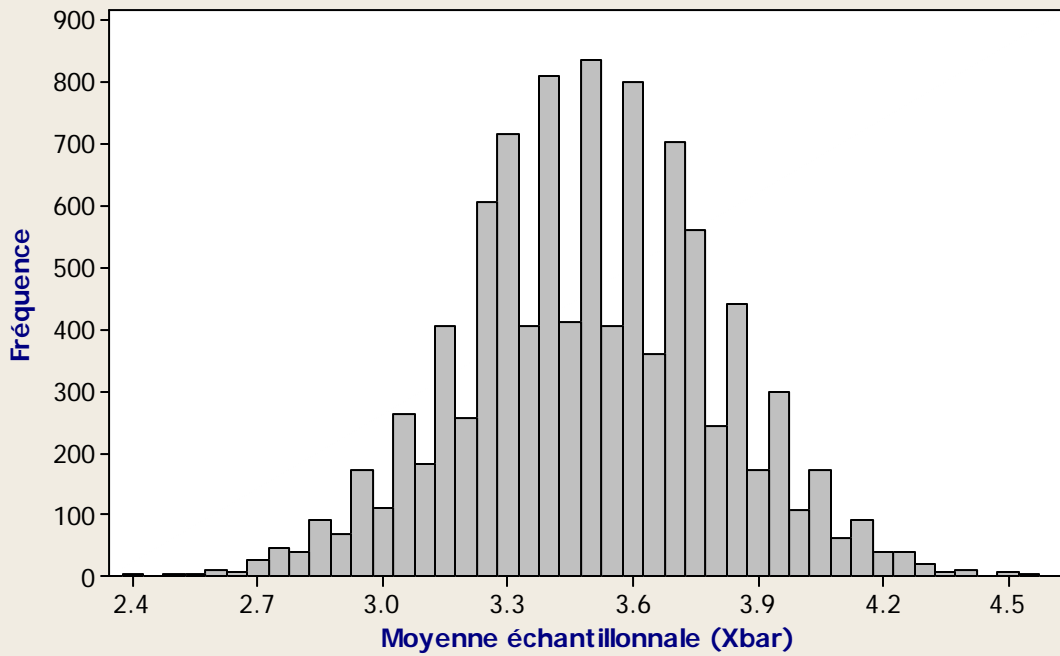
Histogramme pour 100 échantillons avec $n = 30$



Histogramme pour 1000 échantillons avec $n = 30$



Histogramme pour 10000 échantillons avec $n = 30$



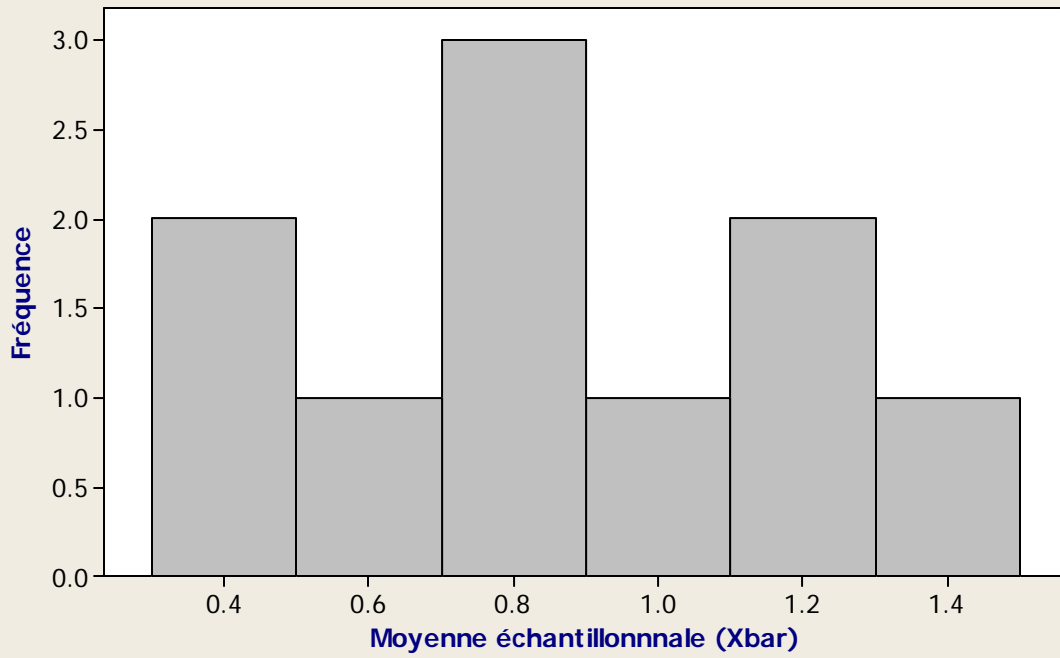
:

Exemple 4.5 Considérons la distribution exponentielle. Si X a une distribution exponentielle, alors sa fonction de densité est de la forme

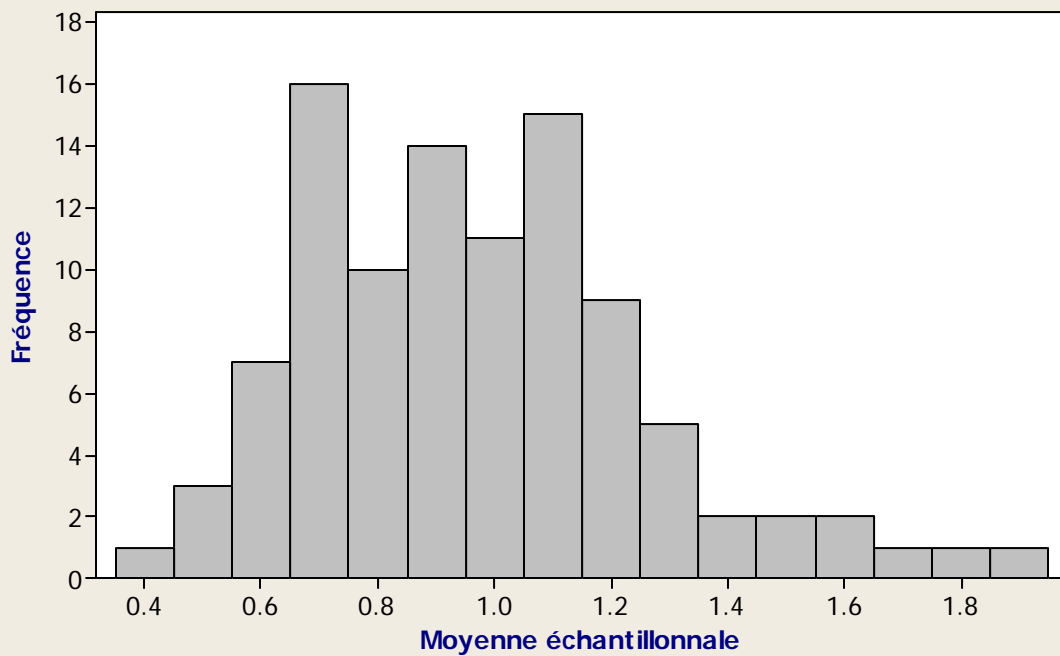
Encore une fois, les figures suivants montrent la distribution de \bar{X} avec R échantillons de taille $n = 10, n = 30$ et $n = 100$. Encore une fois les valeurs de R sont : 10, 100, 1000 et 10 000.

Que remarque t-on?

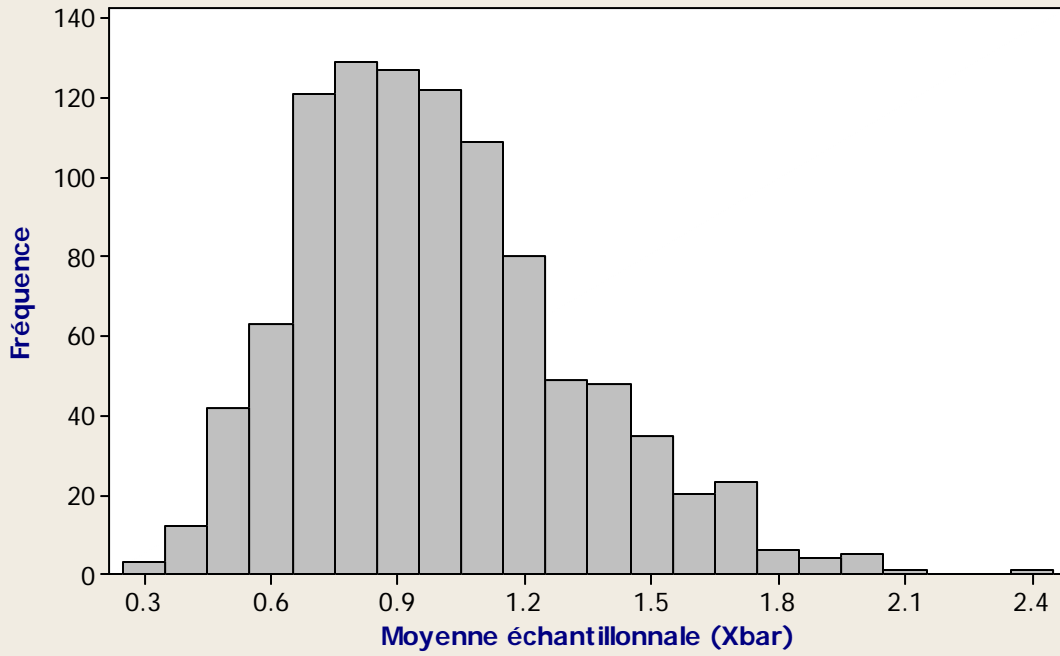
Histogramme pour 10 échantillons avec $n = 10$



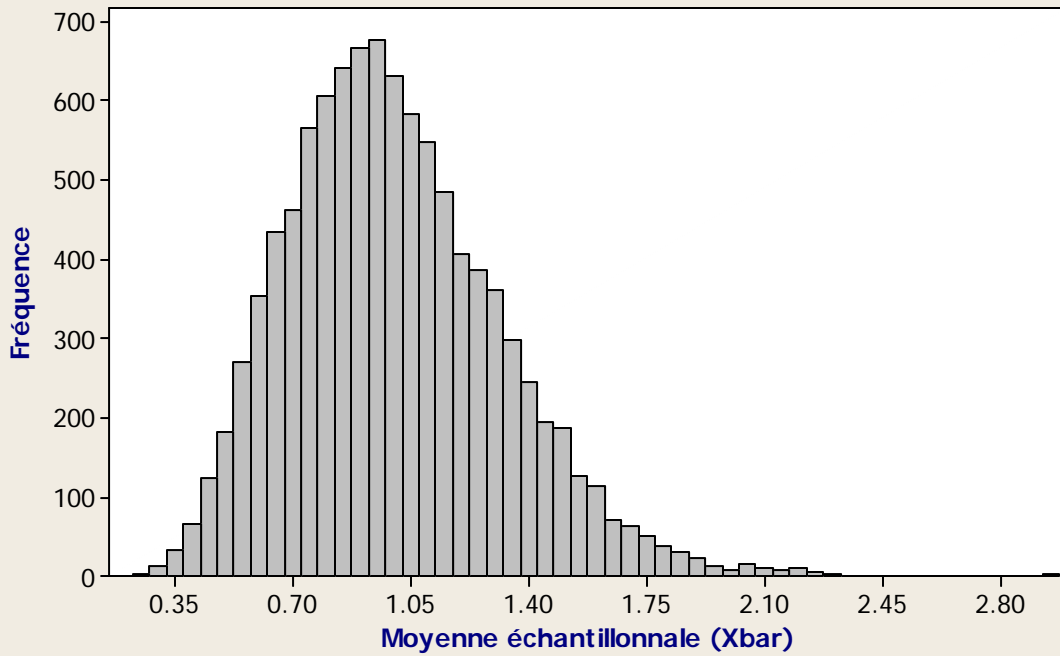
Histogramme pour 100 échantillons avec $n = 10$



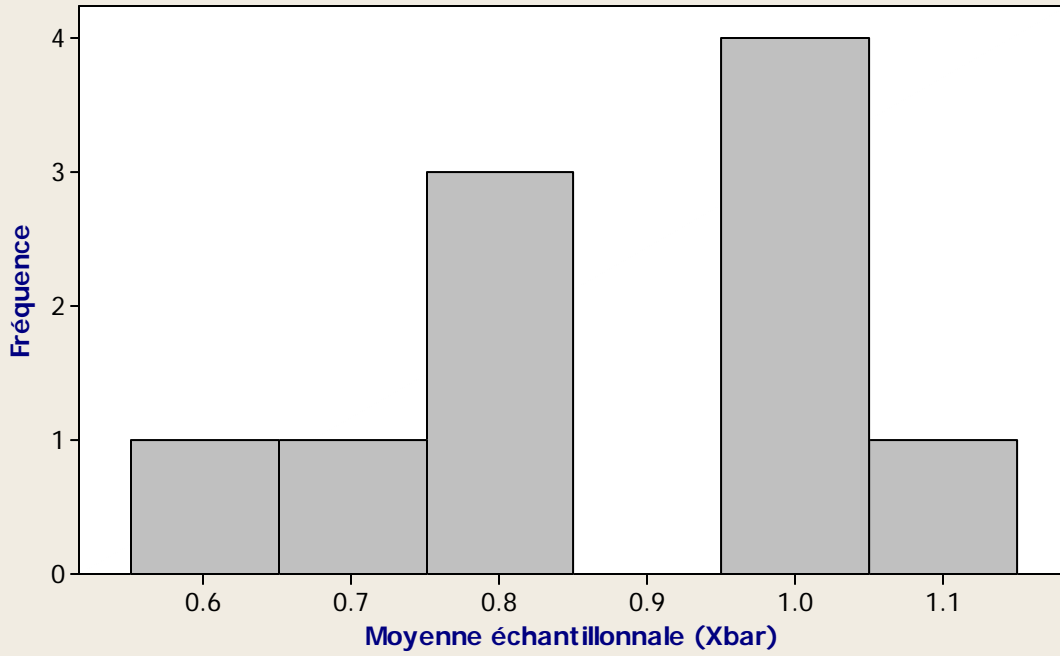
Histogramme pour 1000 échantillons avec $n = 10$



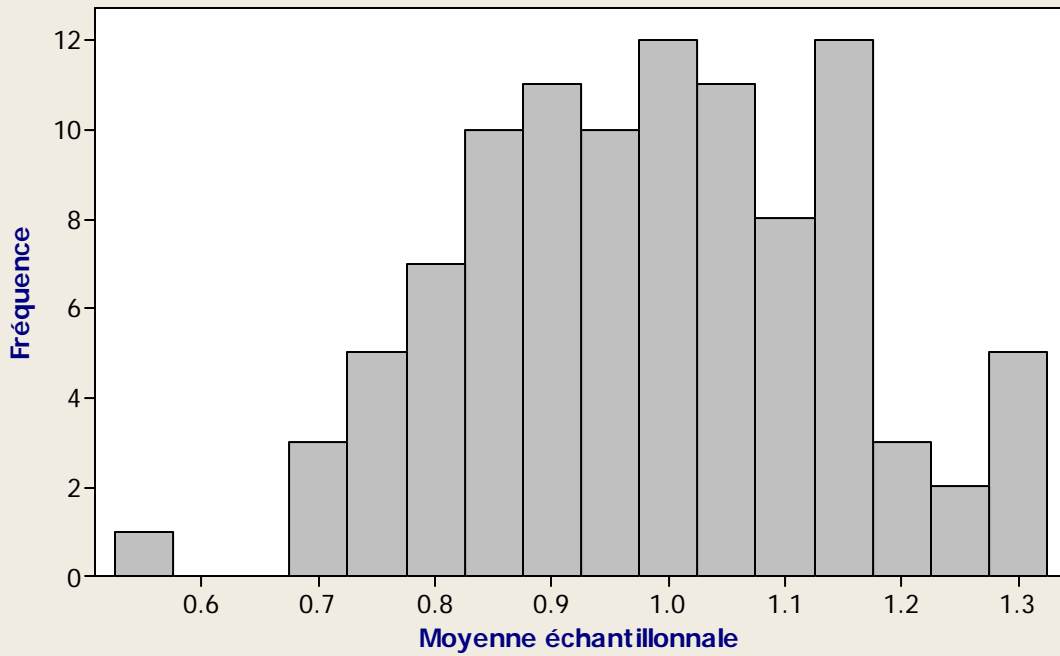
Histogramme pour 10000 échantillons avec $n = 10$



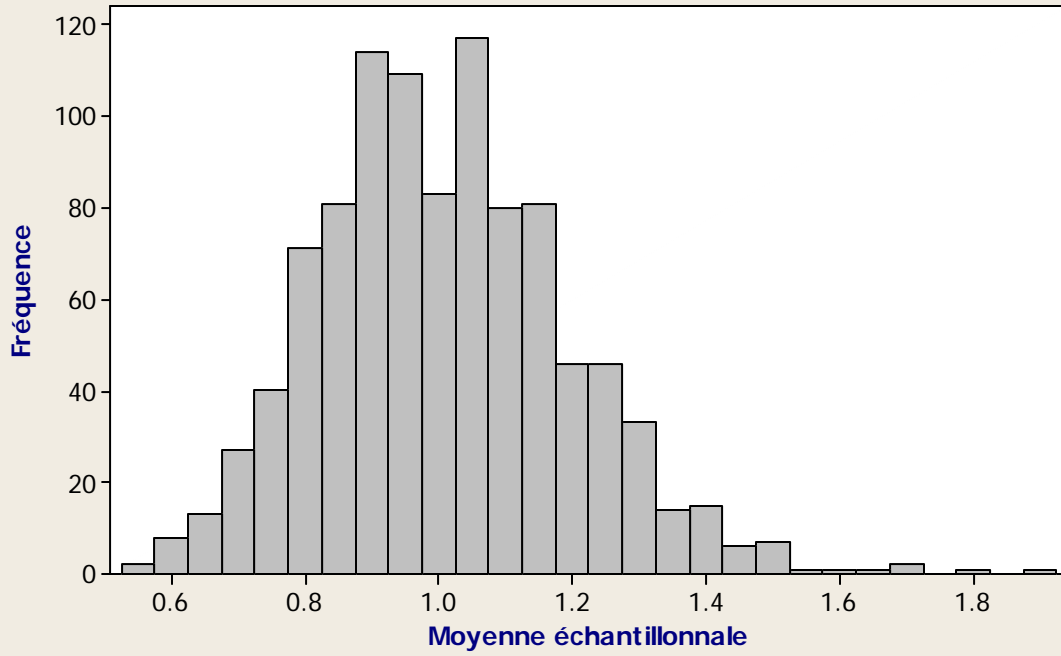
Histogramme pour 10 échantillons avec $n = 30$



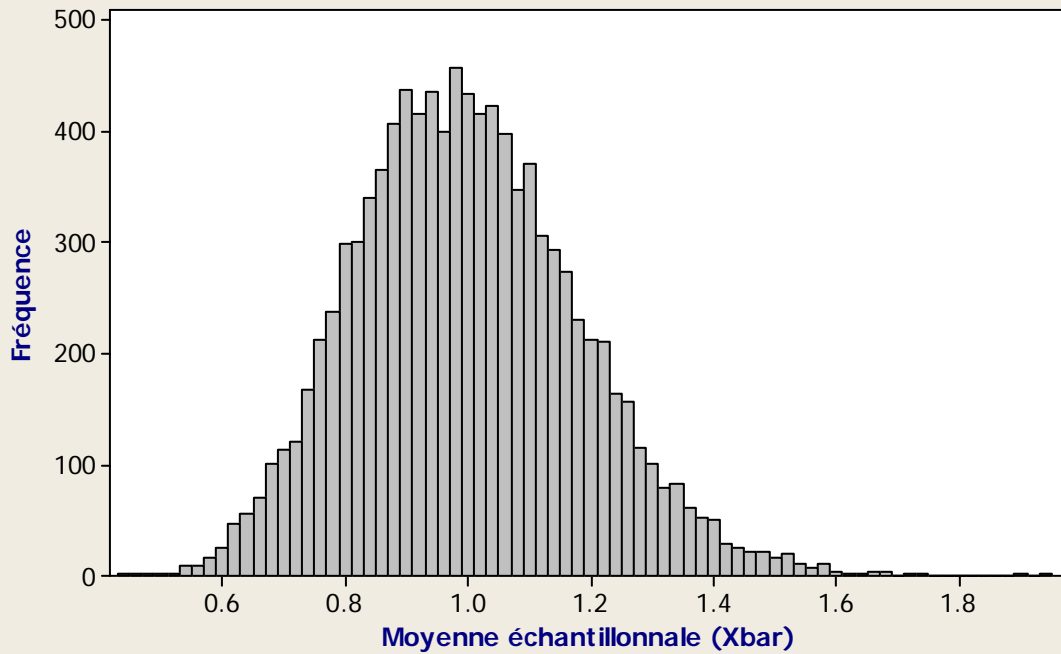
Histogramme pour 100 échantillons avec $n = 30$



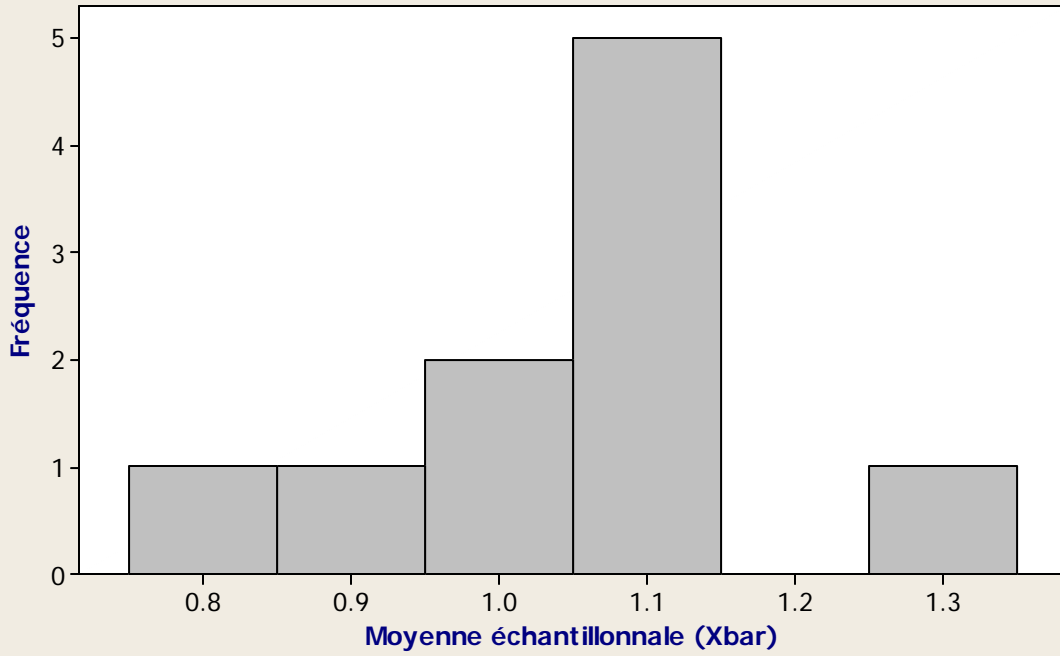
Historgramme pour 100 0 échantillons avec n = 30



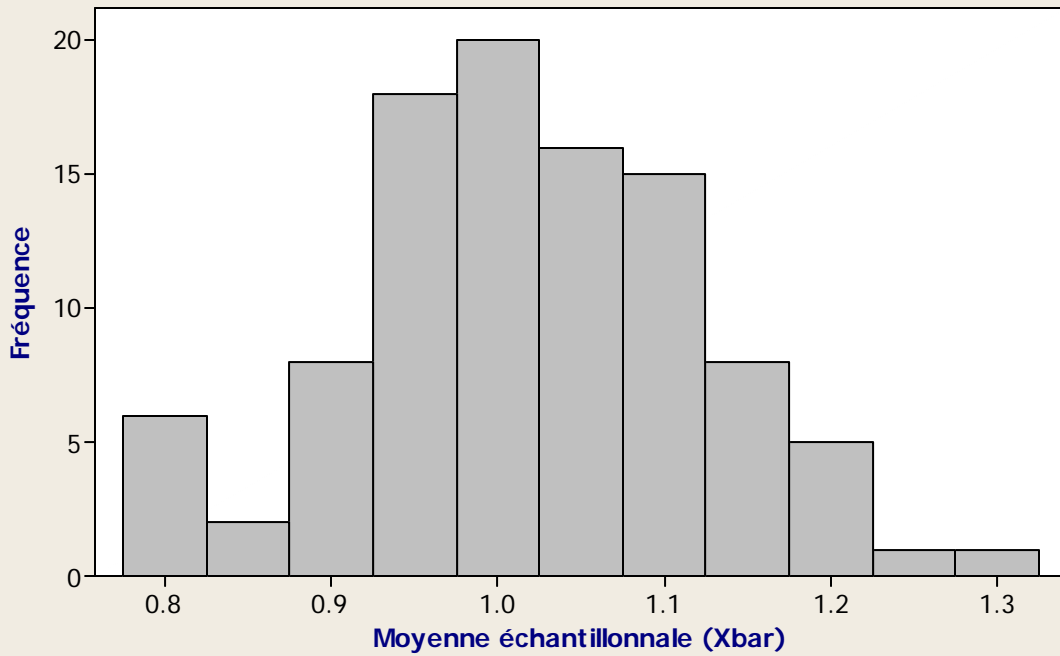
Historgramme pour 10000 échantillons avec n = 30



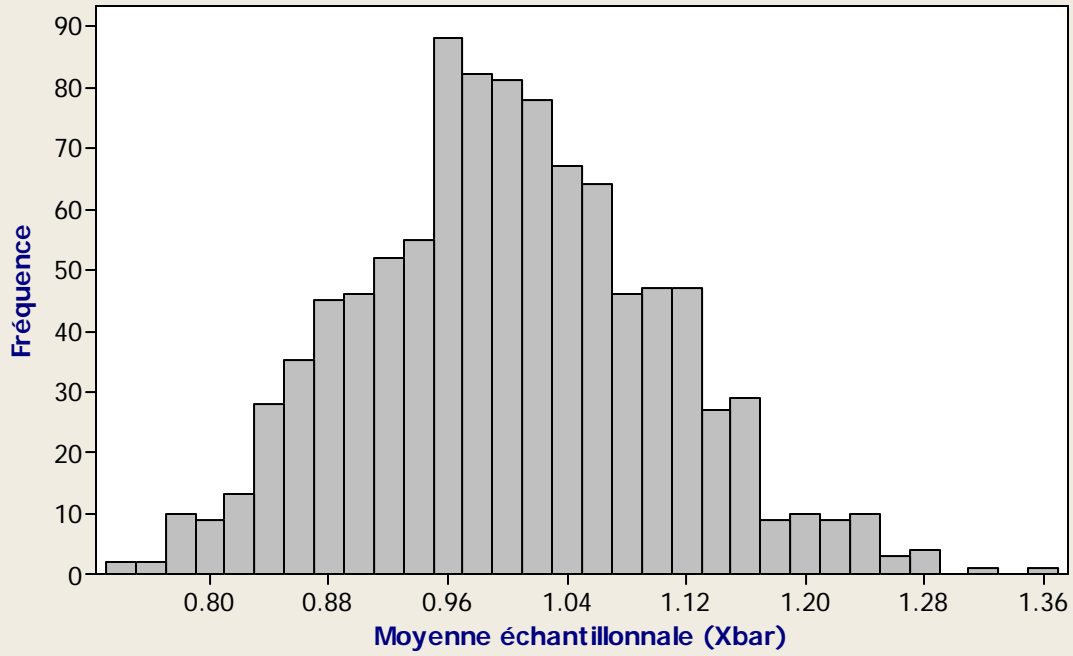
Histogramme pour 10 échantillons avec n = 100



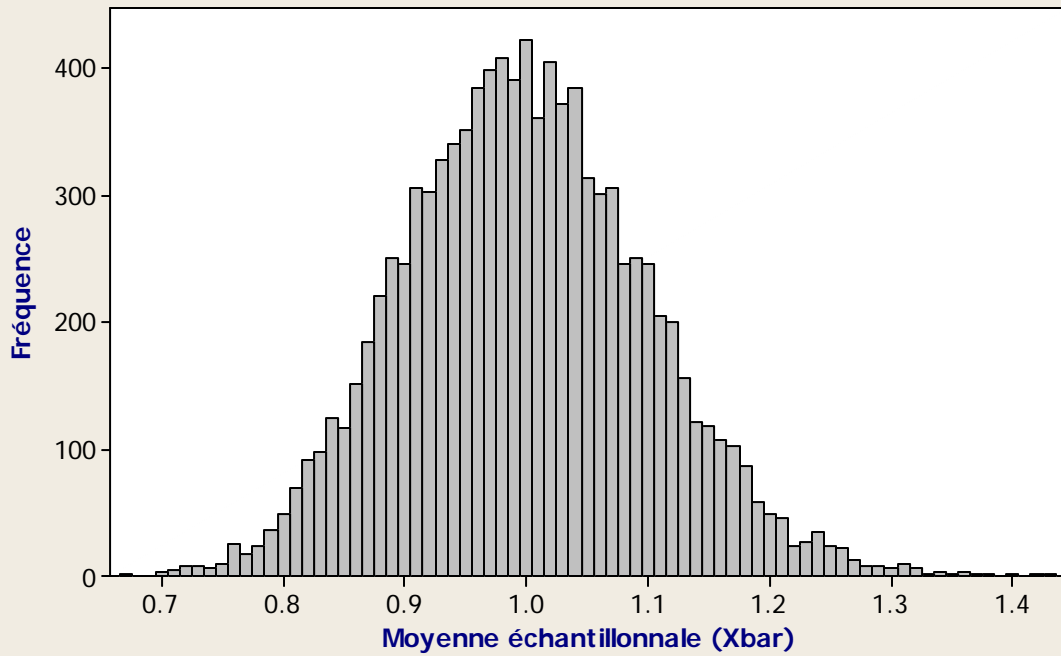
Histogramme pour 100 échantillons avec n = 100



Histogramme pour 1000 échantillons avec $n = 100$



Histogramme pour 10000 échantillons avec $n = 100$



Théorème 4.4 *Théorème limite central*

Soient X_1, \dots, X_n , une suite de variables aléatoires indépendantes de même loi, de moyenne μ et de variance σ^2 . Alors la variable \bar{X} tend en loi vers une normale de moyenne μ et de variance $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$.

En pratique, ceci signifie que si n est assez grand, alors on peut supposer que

$$T = \sum_{i=1}^n X_i \sim \mathfrak{N}(n\mu; \sigma_T^2)$$

où $\sigma_T = \sqrt{n}\sigma$.

Remarque On ne peut pas être définitif ni très précis quant à ce qu'on entend par « n grand ». En pratique, on se donne comme limite le nombre 30: **si $n \geq 30$, on dit que n est «grand»** et donc que \bar{X} est approximativement de loi $\mathfrak{N}(\mu; \sigma^2/n)$. Mais il est évident que l'approximation ne sera pas toujours bonne pour $n \geq 30$, ni nécessairement mauvaise pour $n < 30$.

Exemple 4.6 L'épaisseur des pièces de 25 ¢ est une variable de moyenne 1,625 *mm* et d'écart-type 0,16 *mm*. Une succursale de banque qui reçoit régulièrement des dépôts sous forme de rouleaux de 40 pièces de 25 ¢ mesure la longueur de chaque rouleau déposé et rejette les rouleaux de moins de 64 *mm* de longueur. Quelle est la probabilité de rejeter un rouleau de 40 pièces?

Solution :

Récapitulatif

- (1) Si les données proviennent d'une population normale, alors la distribution de leur moyenne \bar{X} (ou la distribution de leur somme) est normalement distribuée quelque soit la taille de l'échantillon
- (2) Si les données ne proviennent pas d'une population normale, alors la distribution de leur moyenne \bar{X} (ou la distribution de leur somme) est normalement distribuée si la taille de l'échantillon est suffisamment grande. C'est le TCL. Habituellement, une taille $n \geq 30$ est jugée suffisante dans beaucoup de situations mais il ne faut appliquer cette règle aveuglément. Si les données proviennent d'une population dont la distribution est très éloignée de la distribution normale, alors il faut s'attendre à avoir besoin d'une taille d'échantillon plus grande pour que le TCL soit valide.
- (3) La distribution de \bar{X} (qui est une statistique comme on le verra au chapitre 5) est appelée la **distribution d'échantillonnage**. La distribution d'échantillonnage d'une statistique est la fonction de masse de cette statistique.

- (4) Le TCL jouera un rôle important dans le reste du cours car il nous permettra de construire des intervalles de confiance et des tests d'hypothèse.

4.5 Approximation normale de la loi binomiale

Une variable de loi binomiale peut être approchée par une variable de loi normale lorsque n est grand. **Ceci est une application immédiate du théorème limite central.**

En effet, une variable binomiale peut être vue comme une somme $X = \sum_{i=1}^n X_i$, où $X_i = 1$ si le i^{e} essai est un succès et $X_i = 0$ si le i^{e} essai est un échec. X représente donc bien le nombre de succès en n essais. On a $X_i = 1$ avec probabilité p et $X_i = 0$ avec probabilité $q = 1 - p$.

Les variables aléatoires X_i sont indépendantes puisque les épreuves sont indépendantes. De plus, leur moyenne μ et variance σ^2 sont données par $\mu = E(X_i) = p$, $\sigma^2 = \text{Var}(X_i) = pq$.

La variable X peut donc être approchée par une

$$\mathfrak{N}(np ; npq).$$

En plus d'exiger que n soit grand il faut que p ne soit ni trop grand ni trop petit car dans ces cas la loi binomiale est fortement asymétrique. En général, plus p est proche des extrémités 0 ou de 1, plus il faut que n soit grand. Une façon de combiner ces conditions est d'exiger que np et nq ne soient pas trop petits.

Théorème 4.5 *Approximation normale de la binomiale*

Soit $X \sim \mathfrak{B}(n ; p)$ et supposons que $n \geq 30$, $np \geq 5$, et $nq \geq 5$. Alors, il est approximativement vrai que

$$\frac{X - np}{\sqrt{npq}} \sim \mathfrak{N}(0 ; 1),$$

ou encore, que $X \sim \mathfrak{N}(np ; npq)$.

Remarque : Les Figures 4.4-4.7 illustrent le fait que lorsque n croît, l'approximation de la loi binomiale par la loi normale devient meilleure.

Figure 4.4 Loi binomiale vs. loi normale avec $n = 5$

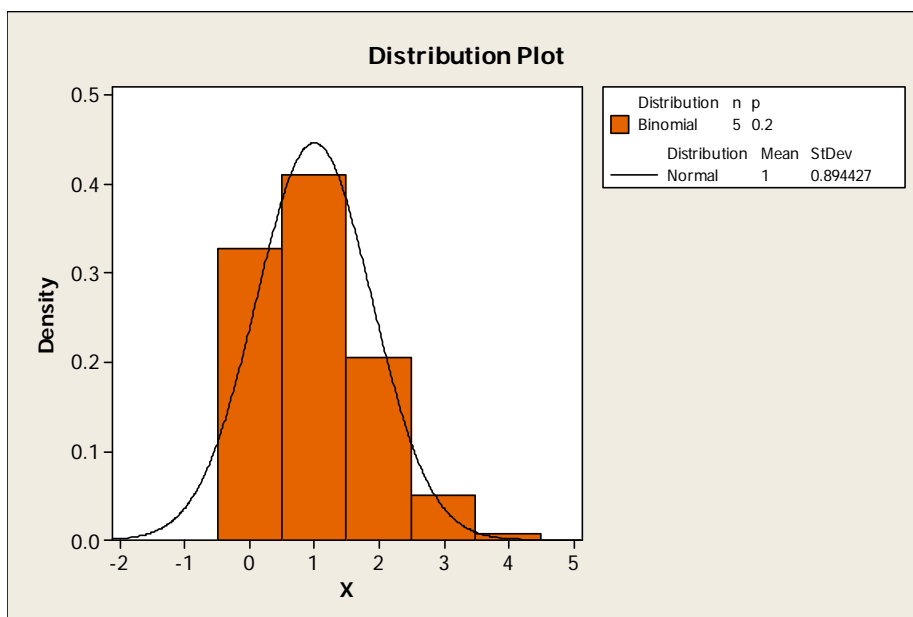


Figure 4.5 Loi binomiale vs. loi normale avec $n = 10$

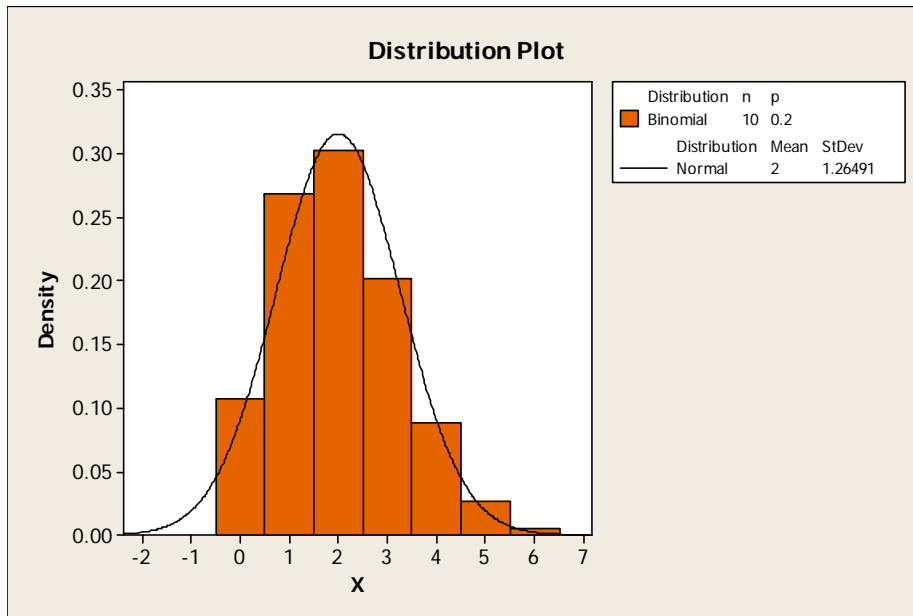


Figure 4.6 Loi binomiale vs. loi normale avec $n = 40$

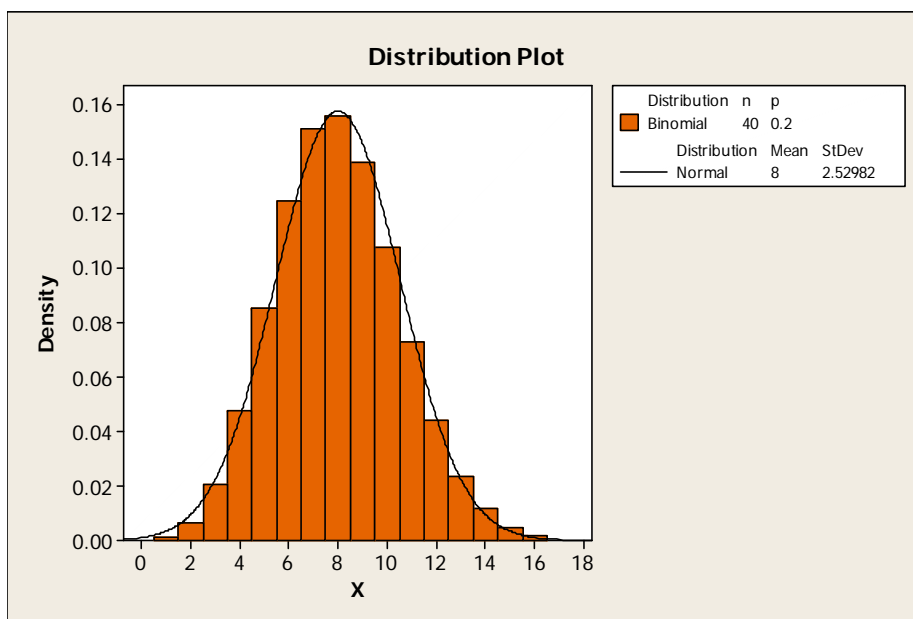
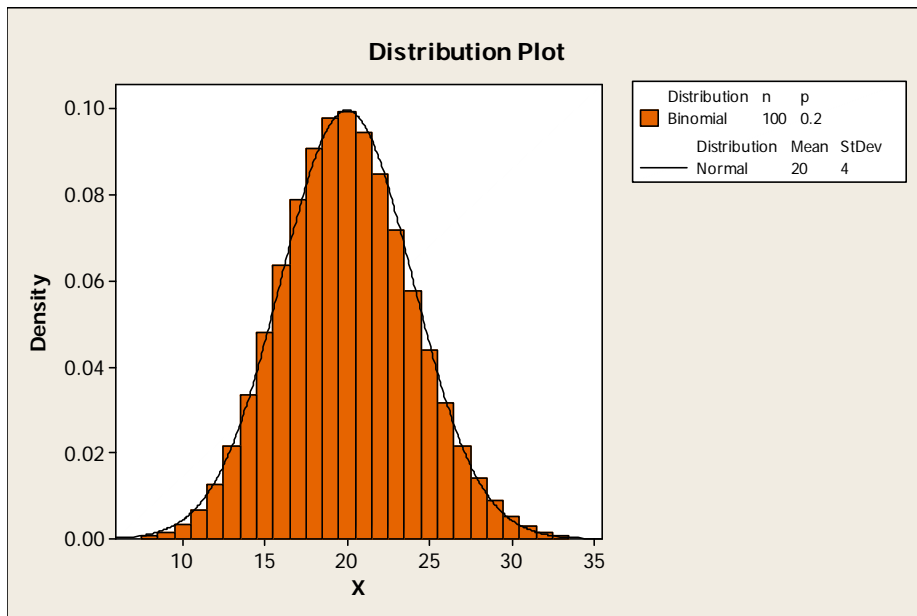


Figure 4.7 Loi binomiale vs. loi normale avec $n = 100$



Exemple 4.7 Le chroniqueur d'un journal crée un émoi dans un ménage après avoir informé ses lecteurs de la durée moyenne d'une grossesse : 266 jours. Dans ce ménage, la femme avait accouché 312 jours après le départ de son mari, un marin, et celui-ci parle déjà de divorce et engage un avocat. L'avocat se renseigne : il apprend qu'effectivement, la durée d'une grossesse est distribuée selon une loi normale de moyenne de 266 jours, avec un écart-type de 16 jours.

- a) L'avocat s'adresse à vous et vous demande de faire un calcul de probabilité pour montrer que la femme a été infidèle. Faites ce calcul et expliquez à l'avocat comment il doit s'exprimer à la cour.
- b) C'est au tour de l'avocat de la femme de vous consulter maintenant. Il vous dit : « je veux bien croire que 312 jours, c'est beaucoup, mais il y a quand même eu 5000 femmes enceintes dans la ville durant cette période, et il ne faudrait pas trop s'étonner que l'une d'elle ait eu une grossesse de durée excessive. Le contraire m'aurait surpris. Votre mandat : montrez qu'en effet ce qui a été observé n'a rien d'étonnant.

Solution :

4.6 Loi du khi-deux

La **loi khi-deux** est une autre loi continue importante. La fonction de densité est donnée par

$$f(x) = \frac{1}{\Gamma(v/2)2^{v/2}} x^{v/2-1} e^{-x/2}, x > 0$$

v est un entier positif appelé **nombre de degrés de**

liberté et $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-t} dt$.

On peut montrer (mais on ne le fera pas!) que

$$E[X] = v \text{ et } Var[X] = 2v.$$

On écrit $X \sim \chi_v^2$ pour signifier que X suit une loi khi-deux à v degrés de liberté.

L'importance de cette loi provient de son lien avec des variables normales centrées réduites: une variable khi-deux est une somme de carrés de variables $\mathfrak{N}(0; 1)$. Nous commençons par le cas d'une seule variable $\mathfrak{N}(0; 1)$.

Théorème 4.6 Soit $Z \sim \mathfrak{N}(0 ; 1)$. Alors $Y = Z^2 \sim \chi_1^2$.

Nous généraliserons ce théorème en montrant qu'une variable χ_v^2 est une somme de carrés de v variables aléatoires $\mathfrak{N}(0 ; 1)$ indépendantes. Mais nous énonçons d'abord ceci:

Proposition 4.1 Soit X_1, \dots, X_n , n variables aléatoires indépendantes, $X_i \sim \chi_{v_i}^2$. Alors la somme

$$X = X_1 + \dots + X_n \sim \chi_v^2$$

où $v = \sum_{i=1}^n v_i$.

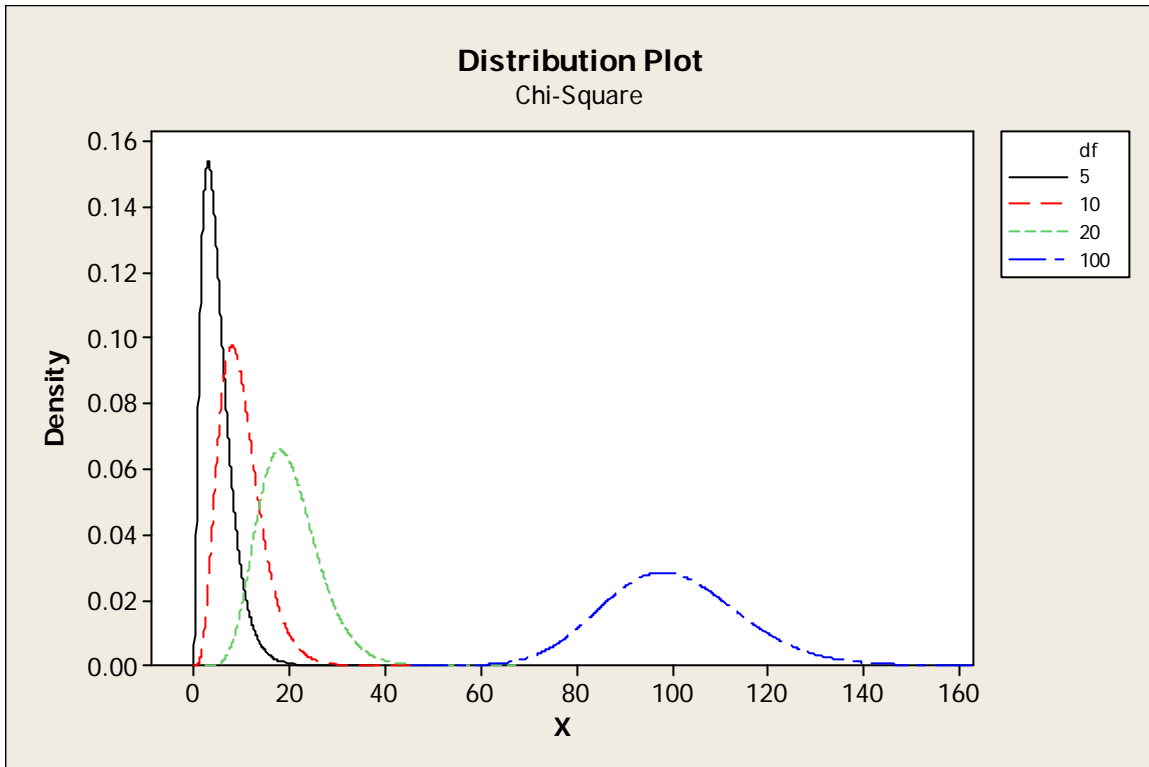
Théorème 4.7 Soit Z_1, \dots, Z_n n variables aléatoires indépendantes, chacune de loi $\mathfrak{N}(0 ; 1)$ et soit

$$X = Z_1^2 + \dots + Z_n^2.$$

Alors $X \sim \chi_n^2$.

Démonstration Les Z_i^2 sont indépendantes, chacune de loi χ_1^2 . Par la Proposition 4.1, leur somme suit une loi χ_n^2 .

Figure 4.8 Fonctions densité de probabilité de la loi du χ^2_ν pour quelques valeurs de ν .



4.7 Loi de Student

La **loi de Student** est une autre loi continue importante.

Définition Soit W et V deux variables aléatoires indépendantes où $W \sim \mathfrak{N}(0 ; 1)$ et $V \sim \chi_v^2$. Alors la variable

$$T = \frac{W}{\sqrt{V / v}}$$

suit une loi appelée **loi de Student à v degrés de liberté**.

On écrit « $T \sim t_v$ » pour signifier que T suit une loi de Student à v degrés de liberté.

On peut démontrer (mais on ne le fera pas!) que si $T \sim t_v$, alors

$$E(T) = 0 \text{ et } \text{Var}(T) = \frac{v}{v-2}, \quad v > 2.$$

La fonction de densité de Student est unimodale et symétrique par rapport à l'origine. Sa forme est semblable à celle de la loi $\mathcal{N}(0 ; 1)$, sauf qu'elle décroît moins rapidement à mesure que la variable s'éloigne de l'origine. Cependant, lorsque le nombre de degrés de liberté croît l'apparence des deux lois devient de plus en plus similaire. Ceci est illustré par les Figures 4.9-4.11.

Figure 4.9 Loi normale vs. loi de Student à 1 d.l.

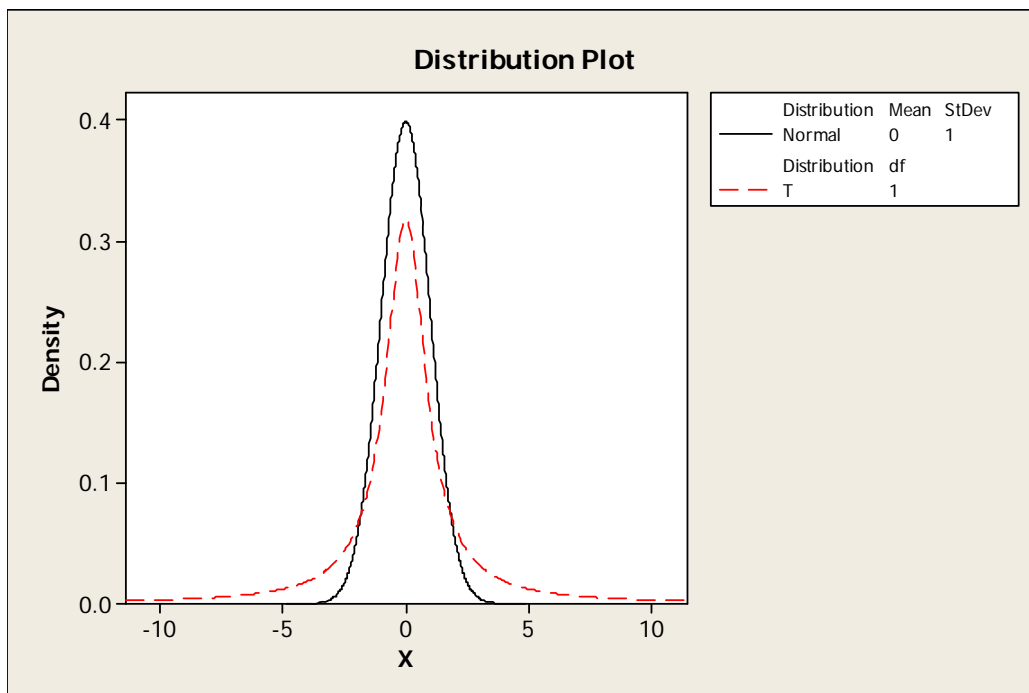


Figure 4.10 Loi normale vs. loi de Student à 10 d.l.

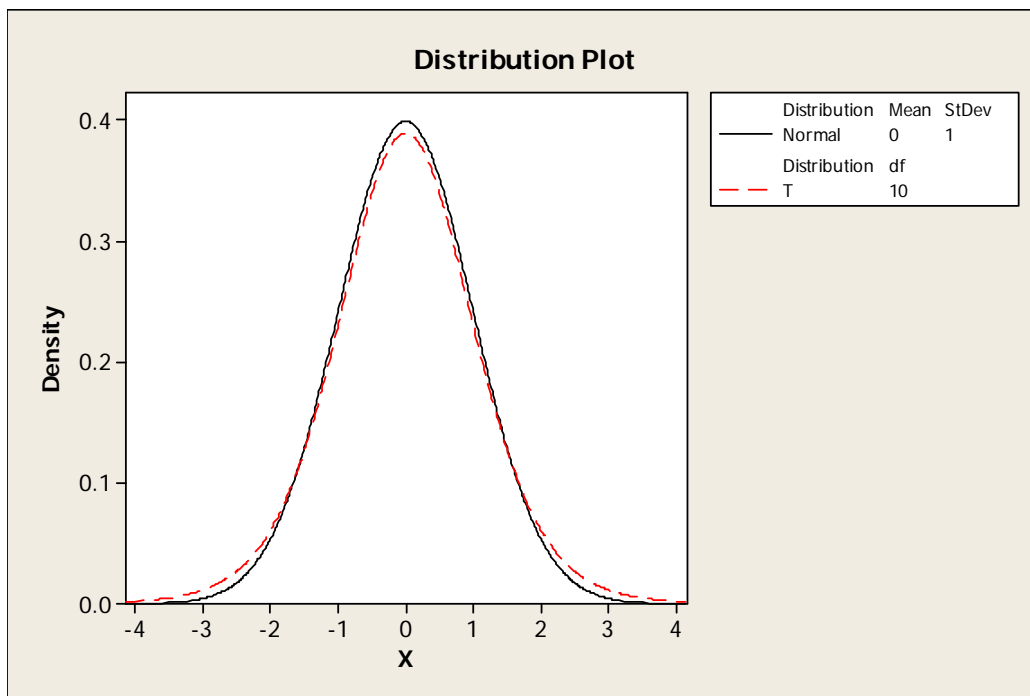
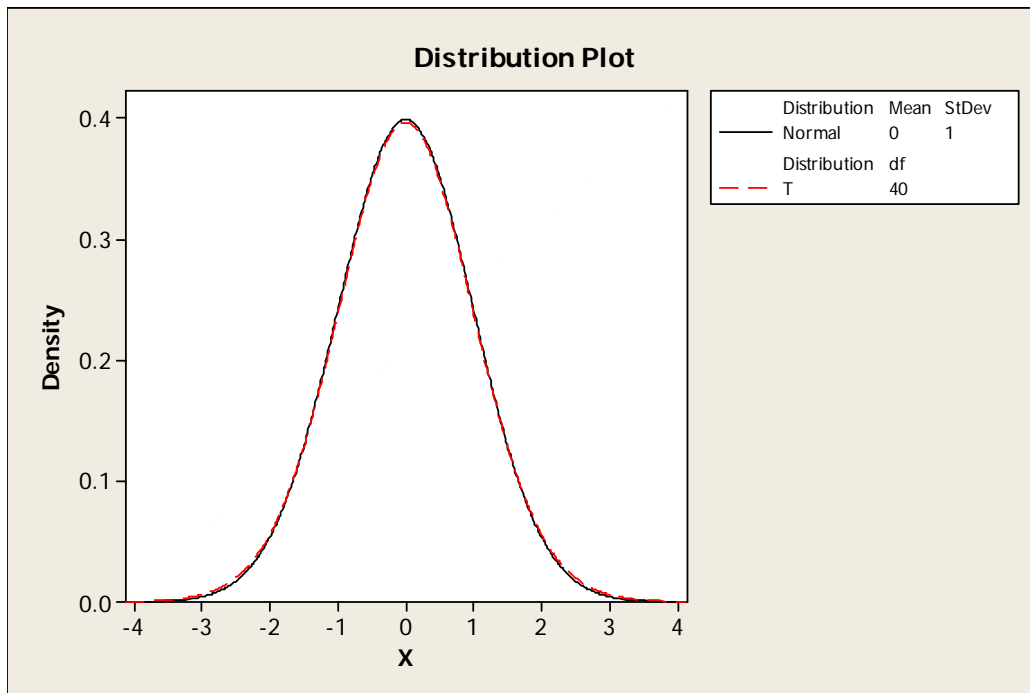


Figure 4.11 Loi normale vs. loi de Student à 40 d.l.



Chapitre 5

Estimation ponctuelle

5.1 Population et échantillon

Une partie de la statistique consiste à estimer ce qu'on appelle les **paramètres** d'une *population*.

Exemple 5.1 Voici quelques exemples de populations et de paramètres:

<i>Population</i>	<i>Paramètre</i>
L'ensemble des salariés québécois	La moyenne μ des revenus des salariés québécois
Un lot de boulons	L'écart-type σ des longueurs des boulons
L'ensemble des étudiants de l'UdM	La proportion p des étudiants qui demeurent chez leurs parents
L'ensemble des ménages d'une petite ville	Le nombre moyen de postes de radio par personne

Considérons le premier exemple, une population de salariés, et supposons qu'on veuille estimer la moyenne μ de leurs revenus. Puisqu'il serait trop coûteux d'interroger tous les salariés de la population, on se contente d'un échantillon de n personnes. On se servira alors de la moyenne des n personnes de l'échantillon pour estimer μ . Intuitivement, l'idée d'utiliser la moyenne de l'échantillon pour estimer la moyenne de la population semble parfaitement raisonnable, même banale. Mais ceci ne nous empêche pas de chercher une façon plus formelle et objective de justifier cette procédure, car les problèmes d'estimation ne sont pas toujours aussi évidents. La question posée dans l'exemple suivant n'a pas de réponse évidente.

Exemple 5.2 On tire un échantillon de 10 ménages afin d'estimer le nombre de téléphones par personne dans la population. Supposons qu'on obtienne les données suivantes.

<i>Ménage</i>	1	2	3	4	5	6	7	8	9	10
<i>Nombre de personnes</i>	5	4	6	8	3	5	2	3	5	2
<i>Nombre de téléphones</i>	2	3	3	2	4	4	2	3	2	3

Comment estimer le paramètre « nombre de téléphones par personne »?

Une façon consiste à diviser le nombre total de téléphones par le nombre total de personnes, soit:

$$\frac{2 + 3 + 3 + 2 + 4 + 4 + 2 + 3 + 2 + 3}{5 + 4 + 6 + 8 + 3 + 5 + 2 + 3 + 5 + 2} = \frac{28}{43} = 0,65.$$

Mais on pourrait également songer à calculer le nombre de téléphones par personne dans chaque ménage, et en calculer ensuite la moyenne:

$$\frac{1}{10} \left(\frac{2}{5} + \frac{3}{4} + \frac{3}{6} + \frac{2}{8} + \frac{4}{3} + \frac{4}{5} + \frac{2}{2} + \frac{3}{3} + \frac{2}{5} + \frac{3}{2} \right) = 0,79$$

Laquelle des deux façon est meilleure?

C'est la première méthode qui est préférable. Mais pour en arriver là, il faut d'abord établir des critères, et avant cela, définir le contexte de façon formelle.

Définition Une suite de n variables aléatoires X_1, \dots, X_n est appelée **échantillon aléatoire simple** si X_1, \dots, X_n sont indépendantes et suivent une même loi.

La population sera identifiée à la fonction de répartition F qui dépendra d'un **paramètre** θ . Nous dénoterons donc la population par $F(x | \theta)$.

5.2 Statistiques et estimateurs

Soit X_1, \dots, X_n un échantillon aléatoire simple provenant d'une population $F(x | \theta)$. Toute variable aléatoire $T(X_1, \dots, X_n)$ fonction de X_1, \dots, X_n est appelée une **statistique**.

Exemple 5.3 Voici quelques exemples de statistiques calculées à partir d'un échantillon aléatoire simple X_1, \dots, X_n de n salaires.

- $T_1(X_1, \dots, X_n) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$: le salaire moyen.
- $T_2(X_1, \dots, X_n) = S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$: l'écart-type des salaires.
- $T_3(X_1, \dots, X_n) = \text{Méd}(X_1, \dots, X_n)$: la médiane des n données X_1, \dots, X_n .
- $T_4(X_1, \dots, X_n) = \hat{p}$: la proportion des salaires inférieurs à 40 000\$.

Le traitement théorique d'un problème d'inférence portant sur une population $F(x | \theta)$ consiste à choisir une statistique appropriée (par exemple, \bar{X} , S , p , ..., etc.) et associer à chaque valeur de la statistique choisie une « décision » à propos du paramètre. La « décision » peut prendre différentes formes, trois desquelles seront traitées dans ce cours:

- 1) **Estimation ponctuelle**: On peut décider que le paramètre a telle ou telle valeur.
- 2) **Estimation par intervalle**: On peut décider que le paramètre se trouve vraisemblablement dans tel ou tel intervalle.
- 3) **Test d'hypothèses**: On peut décider que la valeur du paramètre est ou n'est pas égale à un nombre fixé d'avance.

L'estimation ponctuelle consiste à trouver un estimateur d'un paramètre inconnu θ , c'est-à-dire, une statistique dont les valeurs auraient tendance, en un certain sens, à s'approcher du paramètre.

Par exemple, la moyenne arithmétique $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ est généralement utilisée comme estimateur de l'espérance mathématique μ des variables X_1, \dots, X_n , et la variance échantillonnale $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ est utilisée comme estimation de leur variance σ^2 .

L'estimation par intervalle de confiance consiste à déterminer deux bornes, LI et LS , toutes deux fonctions des observations, et affirmer que le paramètre se situe entre ces deux bornes. Une telle affirmation peut, bien sûr, être erronée, mais les bornes sont déterminées de façon que la probabilité d'erreur soit faible.

Un **test d'hypothèse** consiste à déterminer une règle pour décider quand une hypothèse H_0 concernant un paramètre doit être rejetée. Par exemple,

« *Rejeter l'hypothèse H_0 que $\mu = 10$ si $\bar{X} > 14,3$* »

est une règle, ou un test statistique.

Dans ce chapitre, nous traiterons du problème d'estimation ponctuelle. La notion d'intervalle de confiance sera discutée au chapitre 6, et celle de test d'hypothèse au chapitre 7.

Estimateurs

Toute statistique $T = T(X_1, \dots, X_n)$ utilisée pour faire une estimation d'un paramètre θ est appelée **estimateur** de θ et notée $\hat{\theta} = T(X_1, \dots, X_n)$.

Dans certains cas, le choix d'un estimateur est naturel et intuitif: nous estimons la moyenne μ d'une population par la moyenne \bar{X} de l'échantillon; ou bien nous estimons une probabilité de succès par la proportion de succès dans l'échantillon. Mais il nous faut des critères objectifs pour choisir un estimateur, car parfois 1) plusieurs estimateurs semblent aussi «naturels» l'un que l'autre; 2) aucun estimateur ne se présente à l'esprit comme particulièrement naturel; et 3) certains estimateurs peuvent sembler naturels alors que d'autres sont en fait meilleurs.

5.3 Estimateurs sans biais

Qu'est-ce qui fait qu'un estimateur est préférable à un autre? L'une des qualités généralement souhaitées d'un estimateur est celui d'être sans biais:

Définition Un estimateur $\hat{\theta}$ est dit **sans biais** pour θ si $E(\hat{\theta}) = \theta$ quelle que soit la valeur de θ .

Cette propriété est souhaitable parce qu'elle signifie que l'estimateur n'a tendance ni à sous-estimer ni à surestimer le paramètre: « **en moyenne** » il vise juste.

Estimateur sans biais d'une moyenne

Théorème 5.1 Soit $\{X_1, \dots, X_n\}$ un échantillon aléatoire provenant d'une population de moyenne μ . Alors $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur sans biais de μ .

Démonstration :

Estimateur sans biais d'une variance

Le théorème suivant présente un estimateur sans biais de la variance σ^2 d'une population.

Théorème 5.2 Soit $\{X_1, \dots, X_n\}$ un échantillon aléatoire provenant d'une population de moyenne μ et de variance σ^2 . Alors $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ est un estimateur sans biais de σ^2 .

Démonstration:

Estimateur sans biais d'une proportion p

On prélève un échantillon de taille n d'une grande population de pièces fabriquées, afin d'estimer la proportion p de pièces défectueuses dans la population. Si X est le nombre de pièces défectueuses dans l'échantillon, il est naturel de considérer comme estimateur de p la proportion échantillonnale

$$\hat{p} = \frac{X}{n}.$$

Théorème 5.3 Soit p la proportion des individus d'une population qui appartiennent à une certaine classe \mathcal{C} . Soit X le nombre d'individus qui appartiennent à la classe \mathcal{C} dans un échantillon de taille n . Alors

$$\hat{p} = \frac{X}{n} \text{ est un estimateur sans biais de } p.$$

Démonstration:

5.4 Variance d'un estimateur

Le fait qu'un estimateur soit sans biais, quoique rassurant, ne garantit pas nécessairement une bonne précision. Un estimateur sans biais prend **en moyenne** la valeur juste; mais ceci ne l'empêche pas de s'en écarter de beaucoup. La moyenne \bar{X} d'un échantillon aléatoire simple X_1, \dots, X_n n'est pas le seul estimateur sans biais de la moyenne μ : chacune des observations X_i , par exemple, est sans biais. Il est intuitivement clair que \bar{X} est préférable à un estimateur basé sur une seule des observations. On conçoit qu'il a une plus forte tendance à rester près de μ . C'est là une autre caractéristique souhaitable dans un estimateur: nous voulons qu'il ait tendance à rester près du paramètre. Autrement dit, nous voulons qu'un **estimateur ait une petite variance**.

Variance de \bar{X} et de \hat{p}

Nous savons que

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

et que

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

Ce qu'il faut remarquer ici, c'est que le dénominateur est n : plus n augmente, plus la variance est petite. Ceci correspond à ce que l'on sait déjà par intuition: une estimation est d'autant meilleure que l'échantillon est grand.

Variance de S^2

Si on suppose que la population est normale, on peut démontrer (mais nous le ferons pas!) le résultat suivant:

$$\boxed{\text{Var}(S^2) = 2\sigma^4/(n-1).}$$

Exemple 5.4 Soit X_1, \dots, X_n un échantillon aléatoire simple d'une population de moyenne μ . Montrez que les estimateurs $T_1 = X_1$, $T_2 = 2X_1 - X_2$ et $T_3 = 2\sum_{i=1}^n \frac{i}{n(n+1)} X_i$ sont tous sans biais pour μ .

Calculez leur variance, et dites pourquoi \bar{X} est préférable à ces trois.

Solution :

Exemple 5.5 Soit X_1, \dots, X_n un échantillon aléatoire tiré d'une population de moyenne μ et de variance σ^2 . Considérons l'ensemble des estimateurs de la forme $\hat{\mu} = \sum_{i=1}^n a_i X_i$ où a_1, \dots, a_n sont des constantes.

- a) Quelles conditions les a_i devraient-elles satisfaire pour que $\hat{\mu}$ soit sans biais?
- b) Parmi tous les estimateurs sans biais de la forme $\hat{\mu} = \sum_{i=1}^n a_i X_i$, montrer que celui qui a la plus petite variance est \bar{X} .

Suggestion: Déduisez de l'inégalité $\sum_{i=1}^n \left(a_i - \frac{1}{n} \right)^2 \geq 0$ que

$$\sum_{i=1}^n a_i^2 \geq 1/n.$$

Solution :

Chapitre 6

Intervalles de confiance pour de grands échantillons

6.1 Introduction

Pour un échantillon donné, un estimateur prend une valeur unique, notre estimation du paramètre. Si un échantillon de boîtes de conserves vous donne un poids moyen $\bar{x} = 200$ g, vous présenterez cette valeur comme estimation de la moyenne de la population. Mais il y a peu de chance que cette valeur coïncide exactement avec μ . L'énoncé « la moyenne de la *population* est de 200 » est plutôt audacieux et presque certainement faux. Une affirmation moins téméraire, comme « la moyenne de la population se trouve entre 195 et 205 » a de meilleures chances d'être vraie.

C'est ce qu'on appelle une **estimation par intervalle de confiance**: on entoure la valeur observée d'un certain intervalle (a, b) et on affirme « μ se trouve dans (a, b) ».

La probabilité que cette proposition soit vraie est d'autant plus grande que l'intervalle est large. On choisira donc cet intervalle de telle sorte que la probabilité de dire vrai soit assez bonne.

Dans ce chapitre, on considérera le cas de grandes tailles d'échantillons n . Le cas de n petit sera traité au chapitre 8.

6.2 Intervalle de confiance pour μ

Considérons un échantillon aléatoire X_1, \dots, X_n . Puisque n est grand, on a par le TLC que $\bar{X} \sim \mathfrak{N}(\mu ; \sigma^2/n)$ et

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim \mathfrak{N}(0 ; 1)$$

où

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Soit $z_{\alpha/2}$ un nombre provenant d'une loi normale de moyenne 0 et de variance 1 tel que

$$P(N(0,1) \geq z_{\alpha/2}) = \alpha / 2.$$

Du fait que

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

nous obtenons, en inversant les inégalités,

$$P(\bar{X} - z_{\alpha/2} \sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2} \sigma_{\bar{X}}) = 1 - \alpha$$

L'intervalle

$$[\bar{X} - z_{\alpha/2} \sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2} \sigma_{\bar{X}}]$$

est appelé **intervalle de confiance à $100(1 - \alpha)\%$** : la probabilité que l'intervalle $[\bar{X} - z_{\alpha/2} \sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2} \sigma_{\bar{X}}]$ recouvre la moyenne μ est $1 - \alpha$.

$100(1 - \alpha)\%$ est appelé le niveau de confiance de l'intervalle.

Pour un intervalle de niveau 90%, on a $z_{\alpha/2} = 1.645$; pour un intervalle de niveau 95%, on a $z_{\alpha/2} = 1.96$ et pour un intervalle de niveau 99%, on a $z_{\alpha/2} = 2.58$. Donc, à mesure que le niveau de confiance croît, la largeur d'un intervalle de confiance croît également. La largeur d'un intervalle de confiance, L , est définie comme la différence de la borne supérieure et de la borne inférieure. On a

$$L = 2z_{\alpha/2} \sigma_{\bar{X}}.$$

La demi-largeur de l'intervalle, $z_{\alpha/2} \sigma_{\bar{X}} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, est appelée **marge d'erreur**.

Remarques sur la marge d'erreur:

La marge d'erreur diminue lorsque n augmente. En fait, en prenant 4 fois plus d'observations, on diminue la marge d'erreur de moitié.

- (1) La marge d'erreur est grande si σ est grand pour un n donné.
- (2) On voudrait avoir la plus petite marge d'erreur et le plus grand niveau de confiance. Malheureusement, ce sont deux objectifs contradictoires. Plus le niveau de confiance est grand, plus la marge d'erreur est grande.
- (3) Si on veut une marge d'erreur m pour un niveau de confiance donné, la taille d'échantillon requise est donnée par

$$n = \left(\frac{z_{\alpha/2} \sigma}{m} \right)^2.$$

La formule $[\bar{X} - z_{\alpha/2} \sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2} \sigma_{\bar{X}}]$ ne peut être calculée en pratique, puisqu'elle exige qu'on connaisse σ , ce qui n'arrive que dans des situations très exceptionnelles. Il serait naturel alors de remplacer σ par un estimateur S dans l'expression ci-dessus. Puisque

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

est un estimateur sans biais de σ^2 , nous estimerons $\sigma_{\bar{X}}^2$

par $\hat{\sigma}_{\bar{X}}^2 = \frac{S^2}{n}$ et $\sigma_{\bar{X}}$ par $\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}$.

Puisque n est grand, la statistique

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}_{\bar{X}}}$$

suit à peu près une loi $\mathfrak{N}(0 ; 1)$.

On peut donc continuer à employer la procédure décrite dans cette section, avec pour seule modification le remplacement de $\sigma_{\bar{X}}$ par $\hat{\sigma}_{\bar{X}}$ dans la formule $Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$.

Exemple 6.1 Estimation d'une moyenne

D'une grande population de comptes de banque, on prélève un échantillon de taille $n = 30$ afin d'estimer la valeur moyenne d'un compte ainsi que le montant total des comptes. Voici les résultats, en dollars:

240,82	232,50	740,8	860,32	224,10	7,15	324,12	240,12	190,08	182,75
160,21	148,22	132,	119,25	113,85	108,30	107,10	101,19	99,21	93,12
88,13	80,15	78,13	72,15	67,13	65,14	41,10	32,17	10,02	9,15

a) Estimer la moyenne μ de la population et l'écart-type de l'estimateur.

b) Déterminer un intervalle de confiance à 95% pour la moyenne μ .

6.3 Estimation d'une proportion p

Considérons une population dont une proportion p des membres appartient à une certaine classe \mathcal{C} . Supposons que dans un échantillon de taille n , on trouve X unités appartenant à la classe \mathcal{C} .

Si les tirages sont faits avec remise, ou si la population est grande, alors $X \sim \mathcal{B}(n ; p)$. Pourvu que n ne soit pas trop petit, la distribution de X s'approche d'une $\mathcal{N}(np ; npq)$ (Théorème 4.5). On peut déterminer un intervalle de confiance pour p à partir de son

estimateur $\hat{p} = \frac{X}{n}$. On a

$$\hat{p} \sim \mathcal{N}(p; \sigma_{\hat{p}}^2)$$

$$\text{où } \sigma_{\hat{p}}^2 = \frac{pq}{n}.$$

Alors, on peut affirmer que

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sigma_{\hat{p}}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

Un intervalle de confiance de niveau $1 - \alpha$ est donc donné par

$$\hat{p} - z_{\alpha/2} \sigma_{\hat{p}} \leq p \leq \hat{p} + z_{\alpha/2} \sigma_{\hat{p}}$$

Cependant, $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$ est fonction de p et est donc inconnu. Une solution approximative, presque toujours adéquate, consiste à estimer $\sigma_{\hat{p}}$ par : $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$, ce qui mène à

$$\hat{p} - z_{\alpha/2} \hat{\sigma}_{\hat{p}} \leq p \leq \hat{p} + z_{\alpha/2} \hat{\sigma}_{\hat{p}}.$$

Exemple 6.2 *Estimation d'une proportion*

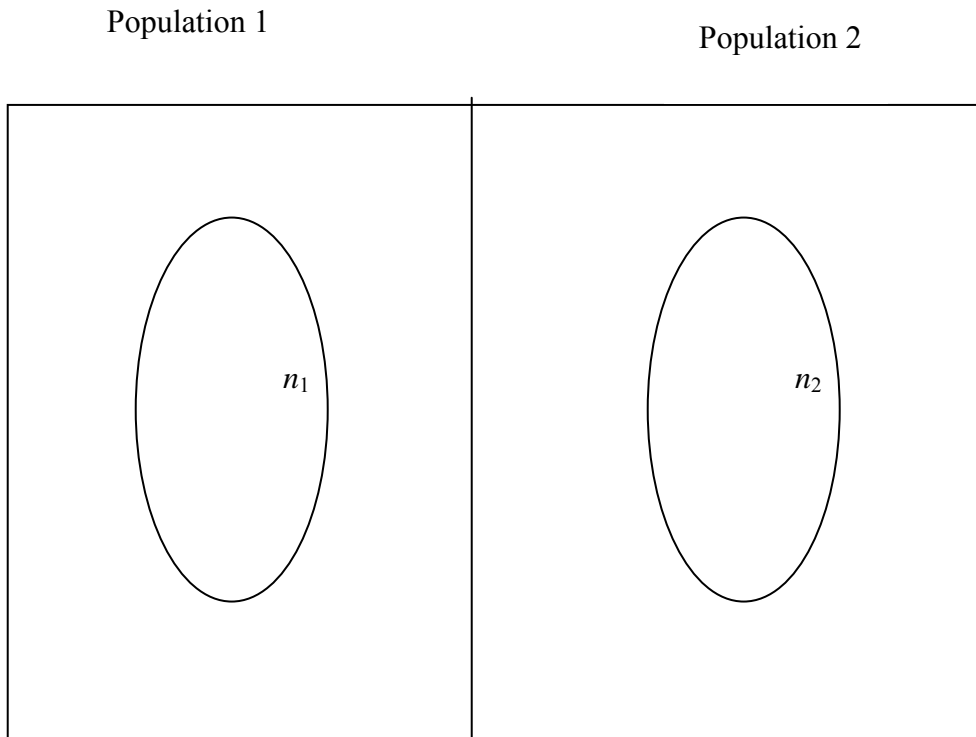
Lors d'un sondage auprès de 500 personnes et portant sur leurs opinions politiques, 180 personnes se sont déclarées favorables au parti A. Estimer la proportion p des gens favorables au parti A au moyen d'un intervalle de confiance de niveau 90%.

Solution :

6.4 Estimation d'une différence de moyenne

En pratique, il est fréquent d'avoir à estimer deux groupes appartenant à une population. Par exemple, on peut vouloir comparer le salaire annuel des hommes et celui des femmes au Canada. Considérons une population qui est divisée en deux-sous populations : Population 1 et Population 2. On tire un échantillon aléatoire de taille n_1 dans la Population 1 et un échantillon aléatoire de taille n_2 dans la Population 2 (voir Figure 6.1). Soient $X_{11}, X_{12}, \dots, X_{1n_1}$ et $X_{21}, X_{22}, \dots, X_{2n_2}$ les échantillons observés. On supposera que **les deux échantillons sont indépendants**. Soient μ_1 et σ_1^2 la moyenne et la variance dans la Population 1 et soient μ_2 and σ_2^2 la moyenne et la variance dans la population 2. On cherche à construire un intervalle de confiance pour la différence, $\mu_1 - \mu_2$. On supposera que les tailles n_1 et n_2 sont grande (chacune ≥ 30).

Figure 6.1 Comparaison de deux échantillons indépendants



Un intervalle de confiance à $100(1 - \alpha)\%$ est donné par

$$\left(\bar{X}_1 - \bar{X}_2\right) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\bar{X}_1 - \bar{X}_2\right) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Les étapes pour la construction de l'intervalle ci-dessus sont laissées en exercice!

Bien sûr, les variances σ_1^2 et σ_2^2 ne sont pas connues en pratique. On les estimera alors par S_1^2 et S_2^2 . Un intervalle de confiance à $100(1 - \alpha)\%$ est donné par

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Exemple 6.3: La durée de vie de deux types de pneus est comparée au moyen d'essais routiers. Un échantillon de $n_1 = 100$ pneus de type 1 et $n_2 = 100$ pneus de type 2 ont servi aux tests. La durée de vie d'un pneu est définie par le nombre de km effectué par le pneu avant qu'il ne devienne inutilisable. Les résultats sont exhibés dans le tableau suivant :

Pneus de type 1	$\bar{X}_1 = 26400$ km; $S_1^2 = 1\,440\,000$
Pneus de type 2	$\bar{X}_2 = 25100$ km; $S_2^2 = 1\,960\,000$

Construisez un intervalle de confiance pour $\mu_1 - \mu_2$, la différence moyenne de la durée de vie des deux types de pneus. Que concluez-vous?

Solution :

6.5 Estimation d'une différence de proportions

Au lieu d'estimer la différence de deux moyennes, on peut vouloir estimer la différence de deux proportions. Considérons une population qui est divisée en deux-sous populations : Population 1 et Population 2. On tire un échantillon aléatoire de taille n_1 dans la Population 1 et un échantillon aléatoire de taille n_2 dans la Population 2 (Voir Figure 6.1). On supposera que **les deux échantillons sont indépendants**. Soient p_1 la proportion d'individus dans la population 1 qui possède une certaine caractéristique et p_2 la proportion d'individus dans la population 2 qui possède ladite caractéristique. On cherche à construire un intervalle de confiance pour la différence, $p_1 - p_2$. On supposera que les tailles n_1 et n_2 sont grande (chacune ≥ 30). Soient \hat{p}_1 et \hat{p}_2 les proportions observés dans les échantillons tirés de la population 1 et de la population 2, respectivement, d'individus qui possèdent la caractéristique d'intérêt.

Un intervalle de confiance à $100(1 - \alpha)\%$ est donné par

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Les étapes pour la construction de l'intervalle ci-dessus sont laissées en exercice!

Bien sûr, les proportions p_1 et p_2 ne sont pas connues en pratique. On acceptera de les estimer par \hat{p}_1 et \hat{p}_2 . Un intervalle de confiance à $100(1 - \alpha)\%$ est donné par

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \leq p_1 - p_2 \leq (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Exemple 6.4: Une enquête a été effectuée dans les écoles secondaires dans une certaine région du Québec. L'échantillon de taille 200 était composé de $n_1 = 100$ filles et $n_2 = 100$ garçons. Parmi les 100 filles, 46 ont avoué consommer de l'alcool sur une base régulière alors que 58 des 100 garçons interrogés ont avoué consommer de l'alcool sur une base régulière. Construire un intervalle de confiance de niveau 90% pour la différence $p_F - p_G$, où p_F et p_G désignent les proportions dans la population de filles et de garçons qui consomment de l'alcool sur une base régulière, respectivement. Que concluez-vous?

Solution :

Chapitre 7

Tests d'hypothèses pour de grands échantillons

7.1 Introduction

Une application importante de la théorie des probabilités consiste à évaluer la vraisemblance d'hypothèses scientifiques à partir des résultats d'une expérience. Ceci nous amène à introduire une procédure importante dans le domaine de la statistique : **le test d'hypothèse**.

Exemple 7.1 Vous lancez une pièce de monnaie 100 fois et obtenez le résultat *FACE* 90 fois. L'hypothèse que la pièce est équilibrée est-elle vraisemblable ?

La réponse intuitive est que non, l'hypothèse que la pièce est équilibrée n'est pas vraisemblable. Pourquoi? Parce que nous avons observé 90 *FACE*, 40 de plus que le nombre auquel on s'attend avec une pièce équilibrée; et que s'il est probable que le nombre de faces s'écarte un petit peu de 50, il est très peu probable que l'écart soit aussi grand que 40. Donc l'hypothèse que la pièce est équilibrée n'est pas plausible. Nous la rejetons.

Exemple 7.2 Un professeur qui prétend pouvoir distinguer l'écriture d'un garçon de celle d'une fille offre de mettre sa prétention à l'épreuve à l'aide de 32 copies d'examen dont 16 sont écrites par des filles et 16 par des garçons. On forme, de façon aléatoire, 16 couples de copies, où dans chaque couple une copie appartient à une fille, l'autre à un garçon. Le professeur décide lequel des deux membres de chaque couple appartient à une fille. Sur les 16 essais, il réussit 14 fois. Est-ce que ceci prouve qu'il a une certaine capacité de distinguer les écritures?

Nous sommes en présence d'une expérience composée de 16 épreuves ($n = 16$), et le nombre de succès X suit une loi binomiale de paramètres $n = 16$ et p . La valeur de p nous est inconnue : c'est justement sur p que porte la question. Si le professeur est absolument incapable de distinguer l'écriture féminine de l'écriture masculine, alors ses réponses sont purement aléatoires et $p = 1/2$; si, par contre, il saisit quelque peu la différence, alors $p > 1/2$. Nous formulons donc notre question de la façon suivante :

Supposons pour l'instant que $p = 1/2$. Alors $X \sim \mathcal{B}(16; 1/2)$ et $E[X] = 8$. Ceci veut dire qu'on s'attend à avoir 8 succès sur 16. Or il y en a eu bien plus. Un nombre de succès *aussi grand* est-il probable lorsque $p = 1/2$? La probabilité d'un nombre de succès aussi grand que 14 lorsque $p = 1/2$ est 0,0021. Donc si $p = 1/2$, il est très *peu* probable d'avoir 14 succès ou plus. On dit alors qu'un tel nombre de succès ne se serait probablement pas réalisé si p ne valait vraiment que $1/2$. On se permet donc de conclure que $p > 1/2$, c'est-à-dire que le professeur a une certaine capacité de distinguer les deux écritures.

L'analyse que nous venons de faire est appelée **test d'hypothèse**. L'hypothèse à tester est que $p = 1/2$, et le test a mené à son rejet. Elle est rejetée parce que si elle était vraie un nombre de succès *aussi grand* que celui qui a été observé aurait été fort peu probable (une probabilité d'environ 0,2 %).

7.2 Développement formel

Reprenons l'exemple 7.2 et supposons que l'expérience n'ait pas encore été faite. Le but de l'expérience est de tester une hypothèse que nous appelons **l'hypothèse nulle** et désignons par H_0 , à savoir

$$\text{Hypothèse nulle} : H_0 : p = \frac{1}{2}.$$

Nous formulons une autre hypothèse, appelée **l'alternative** et désignée par H_1 , soit

$$\text{Alternative} : H_1 : p > \frac{1}{2}.$$

L'une et l'une seule des deux hypothèses doit être vraie, car on exclut d'emblée la possibilité que $p < 1/2$. Nous déciderons laquelle des deux est vraie après avoir fait l'expérience et observé X .

Mais nous pouvons, avant même de faire l'expérience, fixer l'ensemble des valeurs de X pour lesquelles *on rejettera* H_0 . Une chose est évidente : nous ne rejetterons H_0 que si X est trop grand, c'est-à-dire, si et seulement si

$$X \geq C$$

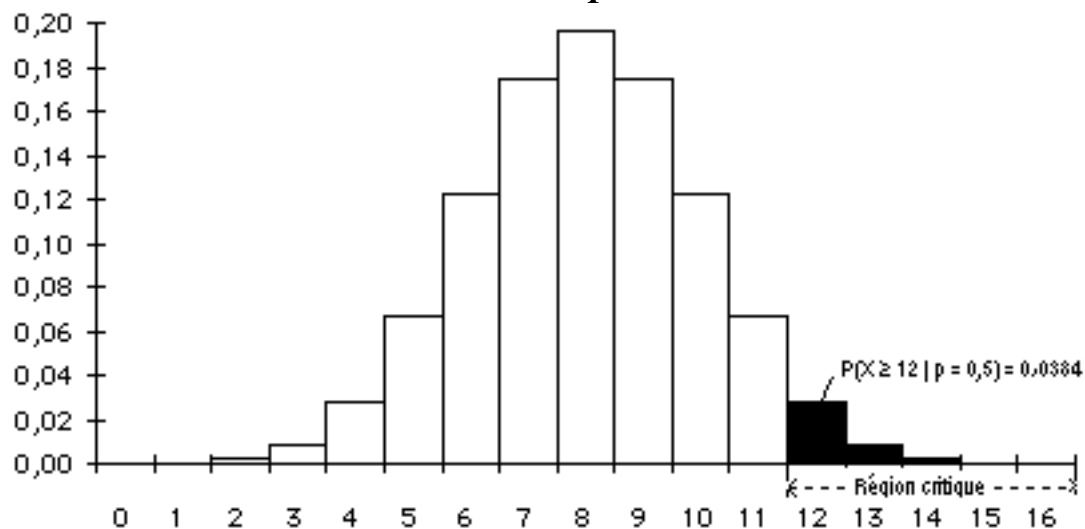
où C est un nombre à déterminer.

L'ensemble des valeurs $X \geq C$ est appelé **région critique**.

Région critique : l'ensemble des valeurs de X pour lesquelles on rejettera H_0 .

Il reste à déterminer la valeur de C . Considérons les conséquences d'un choix quelconque, disons $C = 10$: on fera l'expérience et on rejettera H_0 si $X \geq 10$. Il est possible que H_0 soit vraie et que néanmoins $X \geq 10$; auquel cas on rejetterait H_0 à tort. C'est une erreur qu'on ne peut être sûr d'éviter ; le risque de la commettre est toujours présent. Mais la région critique peut être choisie de façon à réduire ce risque à un niveau acceptable. Que vaut cette probabilité lorsque la région critique est $\{X \geq 10\}$?

Figure 7.1 Région critique pour tester $H_0 : p = 1/2$ contre $H_1 : p > 1/2$



$$P\{\text{rejeter } H_0 \mid H_0 \text{ est vraie}\} = P\{X \geq 10 \mid p = 1/2\} = 0,2272$$

Cette probabilité n'est pas négligeable : si H_0 est vraie, la probabilité est de 22,72 % qu'on la rejette quand même. Dans la plupart des applications on trouverait ce risque d'erreur inacceptable. Le contexte nous fera normalement souhaiter une probabilité plus faible que celle-ci.

Taille de la région critique

La probabilité de rejeter H_0 lorsqu'elle est vraie est appelée **taille de la région critique**.

Taille de la région critique : Probabilité de rejeter H_0 lorsque H_0 est vraie

On souhaite que cette taille soit petite. En général, on fixe un seuil, désigné par α , et on choisit la région critique de telle sorte que sa taille ne soit pas supérieure à α . Considérons quelques régions critiques et déterminons leur taille :

<i>Région critique</i>	<i>Taille</i>
$X \geq 11$	0,1051
$X \geq 12$	0,0384
$X \geq 13$	0,0106
$X \geq 14$	0,0021
$X \geq 15$	0,0002
$X \geq 16$	0,000015

Posons $\alpha = 0,05$. Nous devons déterminer une région critique de taille inférieure ou égale à $0,05$. Le tableau montre que la région critique que nous devons choisir est $X \geq 12$. On dit alors qu'on a un test à 5% . La taille de la région critique est de $0,0384$.

Exemple 7.3 Un procédé de fabrication de boulons est considéré satisfaisant si le pourcentage de boulons défectueux est de 1% . Un inspecteur prélève un échantillon de 200 boulons pour savoir si le procédé est satisfaisant. Soit X le nombre de boulons défectueux qu'on trouvera dans l'échantillon.

Alors $X \sim \mathfrak{B}(200 ; p)$. L'hypothèse nulle est

$$H_0 : p = 0,01$$

L'alternative est

$$H_1 : p > 0,01,$$

car le but de l'inspection est de déceler des lots pour lesquels $p > 0,01$. Posons $\alpha = 5\%$.

On rejettera le lot si le nombre de boulons défectueux est trop grand : la région critique sera de la forme $X \geq C$ où C est un entier qu'on choisit de telle sorte que

$$P\{X \geq C \mid p = 0,01\} \leq 0,05$$

Dans le tableau suivant nous calculons la probabilité

$P\{X \geq C \mid p = 0,01\}$ pour quelques valeurs de C .

C	0	1	2	3	4	5	6
$P\{X \geq C \mid p = 0,01\}$	1	0,8660	0,5954	0,3233	0,1420	0,0517	0,0160

La région critique de niveau 0,05 est donc $\{X \geq 6\}$. Elle est de taille à peu près égale à 0,0160.

Types d'erreur

Nous avons mentionné l'erreur qui consiste à rejeter H_0 quand H_0 est vraie. Cette erreur est appelée **erreur de première espèce**.

Erreur de première espèce : Rejeter H_0 lorsque H_0 est vraie

L'autre erreur possible, l'erreur de seconde espèce, consiste à accepter H_0 quand H_0 est fausse.

Erreur de seconde espèce: accepter H_0 quand H_0 est fausse

Nous avons donc quatre situations possibles. Elles sont schématisées dans le tableau suivant :

		Réalité	
		H ₀ vraie	H ₀ fausse
Décision	On rejette H ₀	Erreur de première espèce	Bonne décision
	On accepte H ₀	Bonne décision	Erreur de seconde espèce

La taille d'une région critique est donc la probabilité d'une erreur de première espèce. **Un test d'hypothèse est conçu de telle façon que la probabilité d'une erreur de première espèce ne soit pas supérieure à α , α fixé à l'avance.**

7.3 Les cinq composantes d'un test d'hypothèse

- (1) L'hypothèse nulle H_0
- (2) L'hypothèse alternative H_A
- (3) La statistique du test et sa p -valeur
- (4) La région critique
- (5) La conclusion

Revenons sur ces 5 composantes plus en détail :

(1) et (2) L'hypothèse alternative H_A est généralement celle que le chercheur essaie de démontrer alors que l'hypothèse nulle H_0 est l'hypothèse contraire. Cette dernière est généralement l'hypothèse 'neutre'.

Par exemple, un organisme pour la défense des femmes cherche à démontrer que les femmes gagnent moins que les hommes pour un même emploi. L'hypothèse alternative sera donc : « les femmes gagnent moins que les hommes pour un même emploi » alors que l'hypothèse nulle sera : « le salaire des hommes et des femmes est identique pour un même emploi ».

L'hypothèse nulle est de la forme *paramètre = constante*. Il y a toujours un signe « = ». Par exemple, dans l'exemple 7.3, on avait $p = 0,01$.

On distingue 3 types d'hypothèses alternatives H_A :

- Celles du type $p < 0,01$. On dira alors du test que c'est un **test unilatéral à gauche**.
- Celles du type $p > 0,01$. On dira alors du test que c'est un **test unilatéral à droite**.
- Celles du type $p \neq 0,01$. On dira alors du test que c'est un **test bilatéral**.

Comment choisir une des trois hypothèses alternatives? C'est ce que l'on cherche à démontrer qui le dictera.

(3) **La statistique du test** est un nombre représentant en quelque sorte une mesure de la distance (standardisée) entre ce que l'on observe dans l'échantillon et l'hypothèse nulle. Si cette distance est grande, cela signifie que l'hypothèse nulle n'est probablement pas vraie auquel cas elle sera rejetée en faveur de l'hypothèse alternative. En revanche, si la distance est petite, alors cela signifie que l'hypothèse nulle est vraisemblable auquel cas on ne la rejettera pas.

Définition La *p-valeur* est la probabilité d'observer une distance aussi grande (i.e., plus grande ou égale à) que celle observée si l'hypothèse nulle est vraie.

On rejettera donc H_0 lorsque la *p-valeur* sera petite. En effet, une petite *p-valeur* signifie que si l'hypothèse nulle est vraie, il est très improbable d'observer une distance entre ce que l'on a recueilli dans l'échantillon et l'hypothèse nulle, aussi grande que celle obtenue.

La statistique du test ainsi que sa *p-valeur* sont deux mesures qui nous permettront de prendre une décision quant au rejet ou au non-rejet de l'hypothèse nulle. Une des deux mesures suffit pour prendre une décision.

(4) La région critique est l'ensemble des valeurs de la statistique du test pour lesquelles on rejettera l'hypothèse nulle. La taille de la région critique est égale à α , qui est fixée à l'avance. Habituellement, les valeurs de α sont égales à 1%, 5% ou 10%.

(5) La conclusion d'un test doit toujours être clairement exprimée dans le contexte de l'expérience. Si l'on rejette l'hypothèse nulle, on dira que **les résultats sont statistiquement significatifs**. Dans le cas contraire, on dira qu'ils ne sont pas statistiquement significatifs.

7.4 Tests d'hypothèses pour μ

Supposons que d'une population, on prélève un échantillon **de grande taille** n afin de déterminer si oui ou non, la moyenne μ est égale à une constante donnée μ_0 . Par exemple, le poids réel du contenu d'une boîte de conserves suit une certaine distribution (pas nécessairement normale), et un inspecteur du gouvernement veut savoir si oui ou non, le poids moyen de toutes les boîtes est égal au poids affiché de $\mu_0 = 400$ g.

Le modèle est le suivant. Nous observons un échantillon aléatoire simple, c'est-à-dire, n variables aléatoires indépendantes et identiquement distribuées X_1, X_2, \dots, X_n , de moyenne μ et de variance σ^2 .

Nous voudrions confronter une hypothèse nulle

$$H_0 : \mu = \mu_0$$

à l'alternative

$$H_1 : \mu < \mu_0$$

Il est naturel de baser le test sur la statistique $\bar{X} = (1/n) \sum_{i=1}^n X_i$, puisque \bar{X} est un estimateur de μ ; et il est également naturel de prendre pour région critique une région de la forme

$$\bar{X} \leq C$$

puisque ce sont les *petites* valeurs de \bar{X} qui devraient mener à la conclusion que $\mu < \mu_0$ — et au rejet de H_0 .

Pour que le test soit de niveau α , il faut choisir C de telle sorte que

$$P\{\bar{X} \leq C \mid H_0\} \leq \alpha$$

Puisque la taille n de l'échantillon est grande, le TLC nous permet d'affirmer que $\bar{X} \sim \mathfrak{N}(\mu_0; \sigma_{\bar{X}}^2)$ approximativement lorsque H_0 est vraie.

Quand rejette-t-on H_0 ?

Afin de prendre une décision, on peut utiliser **3 méthodes équivalentes** : (i) la méthode par la région critique; (ii) la méthode par la valeur critique et (iii) la méthode par la p -valeur.

(i) **Méthode par la région critique**: Le nombre C doit donc satisfaire

$$\begin{aligned}
 P(\bar{X} \leq C) \leq \alpha &\Leftrightarrow P\left(\frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} \leq \frac{C - \mu_0}{\sigma_{\bar{X}}}\right) \leq \alpha \\
 &\Leftrightarrow \frac{C - \mu_0}{\sigma_{\bar{X}}} \leq -z_\alpha \\
 &\Leftrightarrow C \leq \mu_0 - z_\alpha \sigma_{\bar{X}}
 \end{aligned}$$

Si on prend la plus grande valeur de C qui satisfait cette condition, nous obtenons la règle suivante :

$$\text{On rejette } H_0 \text{ si } \bar{X} \leq \mu_0 - z_\alpha \sigma_{\bar{X}}$$

Cette région représente la région critique.

(ii) **Méthode par le point critique:** Nous allons, dans les tests qui suivent, trouver plus commode d'exprimer la région critique en fonction de la variable centrée-réduite

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}. \text{ On dira donc plutôt :}$$

$$\text{On rejette } H_0 \text{ si } Z \leq -z_\alpha$$

(iii) **Méthode par la p -valeur:** on calcule la p -valeur qui est donnée par

$$p\text{-valeur} = P(N(0;1) \leq z_0 \mid \mu = \mu_0),$$

où z_0 est la valeur observée de $\frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}}$. Quand rejettera-t-on H_0 ? **Lorsque la p -valeur est plus petite ou égale à α !**

Exemple 7.4 D'un grand lot de boîtes de conserves, on décide de prélever un échantillon de 40 boîtes de conserves afin de s'assurer que le lot est acceptable, c'est-à-dire, que le poids moyen μ du lot est bien de 400g tel qu'affiché. L'hypothèse nulle est

$$H_0 : \mu = 400.$$

Considérons l'alternative

$$H_1 : \mu < 400.$$

On pèse les 40 boîtes qui affichent un poids moyen $\bar{X} = 396,5$. Supposons que l'écart-type de la population est connu : $\sigma = 2,5$. Posons $\alpha = 0,05$. On a alors $z_\alpha = 1,645$. On a $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 2,5/\sqrt{40} = 0,3952$.

Doit-on rejeter H_0 ? On considère les 3 méthodes.

(i) Méthode par la région critique :

En termes de \bar{X} , cette règle devient : on rejette H_0 si

$$\bar{X} \leq 400 - (1,645 \times 0,3952) = 399,34.$$

Puisque $\bar{X} = 396.5$, on rejette l'hypothèse nulle. On conclut que le fabricant des boites de conserve ne dit pas la vérité et que sa machine est réglée afin de produire des boites ayant un poids inférieur à 400g.

(ii) Méthode par le point critique : On calcule

$$\frac{396,5 - 400}{0,3952} = -8,85 < -1,645.$$

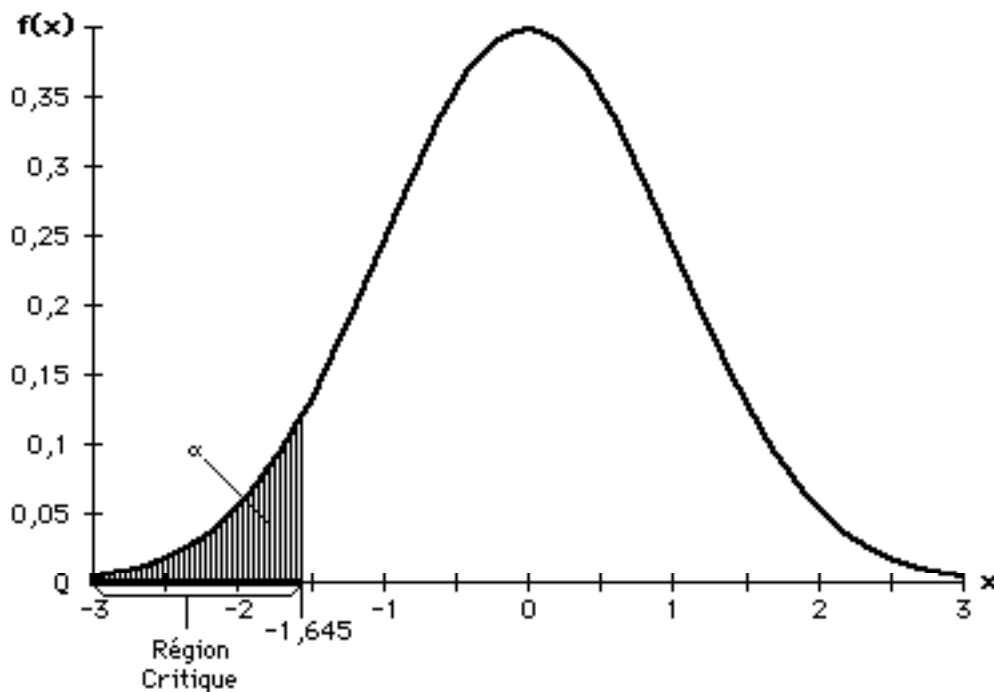
On rejette H_0 car $-8,85 \leq -1.645$. Cette région critique est illustrée dans la figure 7.2.

(iii) Méthode par la p -valeur : on calcule

$$p\text{-valeur} = P(N(0;1) \leq -8,85 | \mu = 400) \\ \approx 0.$$

La p -valeur étant plus petite que 0.05, on rejette l'hypothèse nul.

Figure 7.2 - Région critique pour l'exemple 7.4



Dans l'exemple 7.4, on a utilisé les 3 approches pouvant servir à prendre une décision. L'approche par la p -valeur est populaire puisque les logiciels de statistique fournissent tous les p -valeurs.

Le tableau suivant résume la procédure dans le cas de tests unilatéraux ou bilatéraux.

<i>Hypothèses</i>	<i>Région critique</i>	<i>p-valeur</i>
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$\frac{\bar{X} - \mu_0}{\sigma_{\bar{x}}} < -z_\alpha$	$P(N(0;1) \leq z_0 \mid \mu = \mu_0)$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$\frac{\bar{X} - \mu_0}{\sigma_{\bar{x}}} > z_\alpha$	$P(N(0;1) \geq z_0 \mid \mu = \mu_0)$
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$\left \frac{\bar{X} - \mu_0}{\sigma_{\bar{x}}} \right > z_{\alpha/2}$	$P(N(0;1) \geq z_0 \mid \mu = \mu_0)$

Remarque : Tous ces critères supposent que l'écart-type σ est connu, ce qui est rarement le cas. **Puisque la taille de l'échantillon n est supposée grande, on remplacera σ par S et les procédures décrites ci-dessus demeurent valides.**

Fonction de puissance

Dans un test d'hypothèse, la seule erreur qui est contrôlée est l'erreur de première espèce. Par construction, la probabilité d'une erreur de première espèce est inférieure ou égale à α :

$$P(\text{erreur de première espèce}) = P(\text{rejeter } H_0 \mid H_0 \text{ vraie}) \leq \alpha.$$

Mais que peut-on dire de la probabilité de rejeter H_0 lorsque H_0 est fausse ?

Posons la question dans le cadre de l'exemple 7.4. Puisqu'on rejette H_0 lorsque $\bar{X} < 398,55$, on a

$$P(\text{rejeter } H_0 \mid H_0 \text{ est fausse}) = P(\bar{X} \leq 398,55 \mid \mu < 400).$$

Il est impossible de donner à cette probabilité une valeur unique : elle est fonction de μ . Dénotons cette fonction par ϕ : $\phi(\mu) = P(\text{rejeter } H_0 \mid \mu)$.

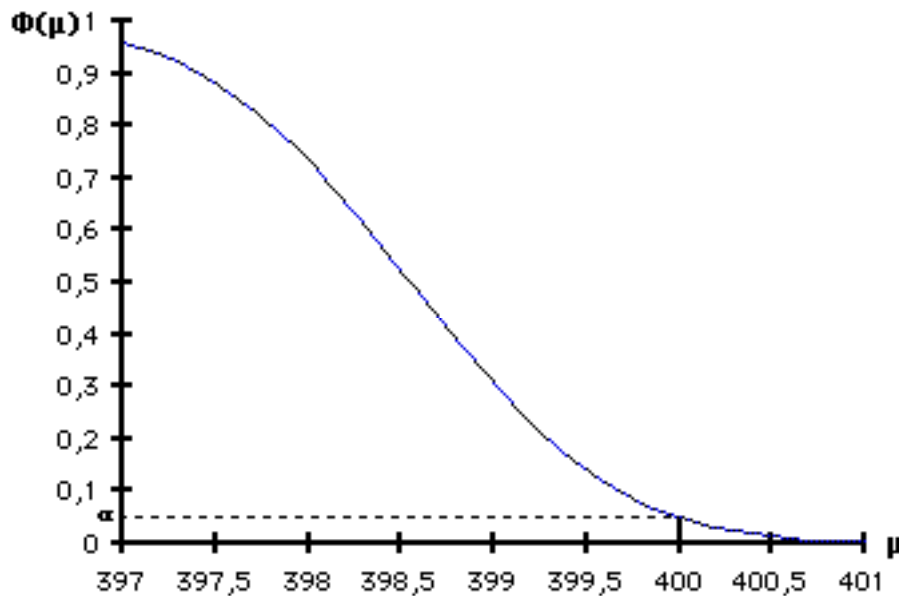
Cette fonction — la fonction qui donne, pour chaque valeur du paramètre, la probabilité de rejeter H_0 — est appelée **fonction de puissance**. Nous pouvons l'exprimer à l'aide de Φ , la fonction de répartition d'une variable de loi $\mathcal{N}(0,1)$. Dans l'exemple 7.4,

$$\begin{aligned}
 \phi(\mu) &= P(\bar{X} \leq 398,55 \mid \mu) = P\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq \frac{398,55 - \mu}{\sigma_{\bar{X}}} \mid \mu\right) \\
 &= P\left(Z \leq \frac{398,55 - \mu}{0,3952} \mid \mu\right) \\
 &= \Phi\left(\frac{398,55 - \mu}{0,3952}\right) \\
 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(398,55 - \mu)/0,8839} e^{-t^2 / 2} dt
 \end{aligned}$$

Plusieurs logiciels permettent de calculer ces probabilités, et de tracer la courbe. La figure 7.3, qui présente un graphique de la fonction de puissance, a été tracée par le logiciel Excel.

Figure 7.3 *Fonction de puissance pour le test de l'exemple 7.4*

$$\phi(\mu) = P(\text{rejeter } H_0 \mid \mu)$$



Remarques à propos de la fonction de puissance (Figure 7.3)

- La valeur de ϕ au point $\mu = 400$ est 0,05, la probabilité d'une erreur de première espèce.
- Lorsque $\mu < 400$, la valeur de ϕ est supérieure à 0,05, ce qui est normal : lorsque $\mu < 400$, H_0 est fausse, et il faut bien que la probabilité de la rejeter soit élevée.
- Lorsque μ n'est que légèrement inférieure à 400, la valeur de ϕ , bien que légèrement supérieure à 0,05, est faible. Ce qui signifie que lorsque H_0 n'est que « un petit peu fausse », il est peu probable qu'on la rejette.
- Inversement, lorsque μ est bien plus petit que 400, la probabilité de rejet est forte : un écart est d'autant plus facile à détecter qu'il est important.
- La probabilité d'une erreur de seconde espèce est représentée graphiquement par la distance entre 1 et $\phi(\mu)$.
- En principe, les valeurs de $\phi(\mu)$ pour $\mu > 400$ sont sans intérêt, puisque nous avons d'emblée décidé d'exclure ces valeurs.

Exemple 7.5 Dans l'exemple 7.4 nous avons pris pour alternative l'hypothèse

$$H_1 : \mu < 400.$$

Ceci signifiait que nous ne voulions rejeter H_0 que si le contenu des boîtes était insuffisant. Ce serait, par exemple, l'attitude d'un inspecteur gouvernemental dont le seul souci serait de protéger le consommateur.

Mais le fabricant normalement serait intéressé à détecter tout écart à la norme, qu'il soit en trop ou en moins. On voudra donc rejeter H_0 si $\mu > 400$ aussi bien que si $\mu < 400$. On signifie ceci en posant les hypothèses de cette façon :

$$H_0 : \mu = 400,$$

$$H_1 : \mu \neq 400.$$

On rejette H_0 lorsque \bar{X} s'éloigne trop de 400. La région critique est donc de la forme

$$\frac{|\bar{X} - 400|}{\sigma_{\bar{X}}} > C,$$

où C doit être choisi de telle sorte que la probabilité de rejeter H_0 à tort soit au plus α .

On rejette H_0 si

$$\bar{X} > 400 + z_{\alpha/2} \sigma_{\bar{X}}$$

ou si

$$\bar{X} \leq 400 - z_{\alpha/2} \sigma_{\bar{X}}$$

Par conséquent, $C = z_{\alpha/2}$, et la région critique est

$$\left| \frac{\bar{X} - 400}{\sigma_{\bar{X}}} \right| > z_{\alpha/2}$$

Rappelons que $\sigma = 2,5$ et donc que $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 2,5/\sqrt{8} = 0,3952$. Posons $\alpha = 0,05$. On a alors $z_{\alpha/2} = 1,96$, et on rejettera H_0 si

$$\left| \frac{\bar{X} - 400}{0,3952} \right| > 1,96$$

En termes de \bar{X} , cette règle devient : on rejette H_0 si $\bar{X} \geq 400,77$ ou si $\bar{X} \leq 399,22$.

La fonction de puissance de ce test est

$$\phi(\mu) = P(\bar{X} \geq 400,77) + P(\bar{X} \leq 399,22)$$

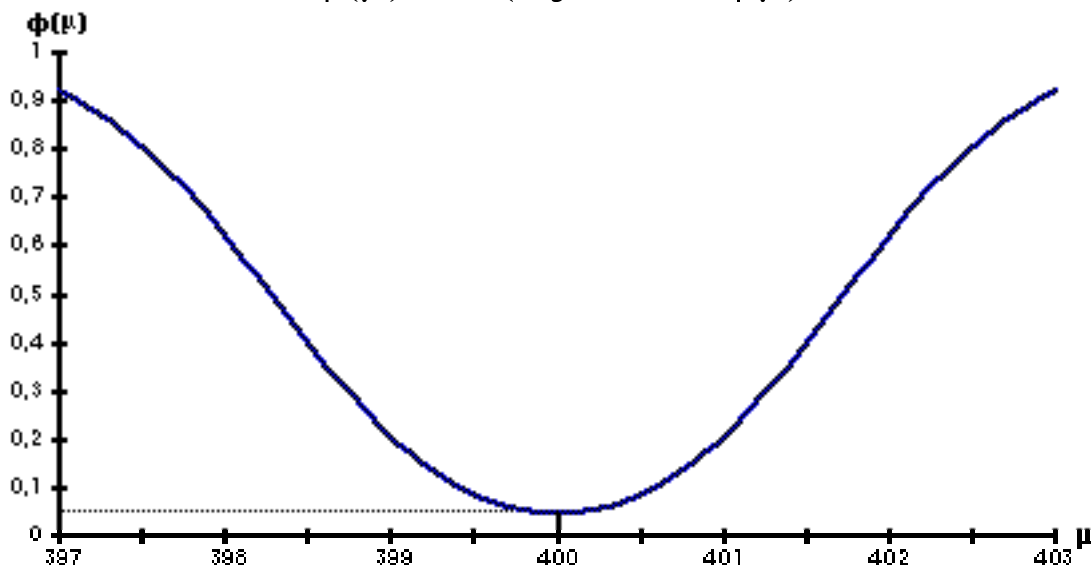
$$= P\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \geq \frac{400,77 - \mu}{\sigma_{\bar{X}}}\right) + P\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq \frac{399,22 - \mu}{\sigma_{\bar{X}}}\right)$$

$$= 1 - \Phi\left(\frac{400,77 - \mu}{0,3952}\right) + \Phi\left(\frac{399,22 - \mu}{0,3952}\right)$$

La figure 7.4 présente le graphique de cette fonction de puissance.

Figure 7.4 - *Fonction de puissance pour le test de l'exemple 7.3.2*

$$\phi(\mu) = P(\text{rejeter } H_0 \mid \mu)$$



7.5 Test sur la différence de deux moyennes

Plusieurs enquêtes et expériences scientifiques ont pour but de déterminer s'il y a une différence entre les moyennes de deux populations.

Le modèle mathématique est le suivant : Considérons une population qui est divisée en deux-sous populations : Population 1 et Population 2. On tire un échantillon aléatoire de taille n_1 dans la Population 1 et un échantillon aléatoire de taille n_2 dans la Population 2 (voir Figure 6.1). Soient $X_{11}, X_{12}, \dots, X_{1n_1}$ et $X_{21}, X_{22}, \dots, X_{2n_2}$ les échantillons observés. On supposera que **les deux échantillons sont indépendants**. Soient μ_1 et σ_1^2 la moyenne et la variance dans la Population 1 et soient μ_2 and σ_2^2 la moyenne et la variance dans la population 2. On cherche à construire un intervalle de confiance pour la différence, $\mu_1 - \mu_2$. On supposera que les tailles n_1 et n_2 sont grande (chacune ≥ 30).

Par le TLC, on a donc

$$\begin{aligned}\bar{X}_1 &\sim \mathfrak{N}(\mu_1; \sigma_1^2/n_1), \quad i = 1, 2, \dots, n_1 \\ \bar{X}_2 &\sim \mathfrak{N}(\mu_2; \sigma_2^2/n_2), \quad j = 1, 2, \dots, n_2\end{aligned}$$

Et, grâce à l'indépendance des 2 échantillons, on a

$$\bar{X}_1 - \bar{X}_2 \sim \mathfrak{N}(\mu_1 - \mu_2; \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

Supposons que μ_1 et μ_2 sont inconnues et qu'on veuille tester l'hypothèse que, pour un nombre δ donné,

$$H_0 : \mu_1 - \mu_2 = \delta$$

contre l'une des alternatives

$$H_1 : \mu_1 - \mu_2 \neq \delta$$

$$H_1 : \mu_1 - \mu_2 > \delta$$

$$H_1 : \mu_1 - \mu_2 < \delta$$

Notons que le cas $\delta = 0$ est le plus courant.

Variations connues

Intuitivement, nous devrions baser notre test sur l'écart

$$\bar{X}_1 - \bar{X}_2 - \delta$$

La statistique du test est donnée par

$$Z_{\bar{X}_1 - \bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

On rejettera H_0 lorsque

(i) $Z_{\bar{X}_1 - \bar{X}_2} > z_\alpha$ si l'hypothèse alternative est de la forme :

$$H_1: \mu_1 - \mu_2 > \delta.$$

(ii) $Z_{\bar{X}_1 - \bar{X}_2} < -z_\alpha$ si l'hypothèse alternative est de la

$$\text{forme : } H_1: \mu_1 - \mu_2 < \delta.$$

(iii) $|Z_{\bar{X}_1 - \bar{X}_2}| > z_{\alpha/2}$ si l'hypothèse alternative est de la

$$\text{forme : } H_1: \mu_1 - \mu_2 \neq \delta.$$

On peut également utiliser l'approche par la p -valeur pour prendre une décision. Dans ce cas, on rejettera H_0 lorsque

(i) $P(N(0;1) \geq z_0 \mid \mu_1 = \mu_2) \leq \alpha$ si l'hypothèse alternative est de la forme : $H_1: \mu_1 - \mu_2 > \delta$.

(ii) $P(N(0;1) \leq z_0 \mid \mu_1 = \mu_2) \leq \alpha$ si l'hypothèse alternative est de la forme : $H_1: \mu_1 - \mu_2 < \delta$.

(iii) $P(|N(0;1)| \geq |z_0| \mid \mu_1 = \mu_2) \leq \alpha$ si l'hypothèse alternative est de la forme : $H_1: \mu_1 - \mu_2 \neq \delta$.

Variances inconnues

En pratique, bien sûr, σ_1^2 et σ_2^2 ne sont pas connues et les critères proposés ne sont pas utilisables. Puisque n_1 et n_2 sont grands, les tests ci-dessus peuvent quand même être employés ; il suffit de remplacer σ_1^2 et σ_2^2 par leurs estimateurs respectifs

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^n (X_{1i} - \bar{X}_1)^2 \quad \text{et} \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^n (X_{2i} - \bar{X}_2)^2 .$$

On utilise alors

$$Z_{\bar{X}_1 - \bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2 - \delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Exemple 7.6 Un diététicien a développé un nouveau régime alimentaire faible en lipides, en glucides et en cholestérol. Bien que le régime visait initialement les individus atteints de maladies cardio-vasculaires, le diététicien souhaite étudier l'effet du régime sur des personnes souffrant d'obésité. Deux échantillons de personnes obèses de taille 100 ont été sélectionnés. Le premier groupe est soumis au nouveau régime développé par le diététicien alors que le deuxième groupe est soumis à un régime régulier qui comprend approximativement la même quantité de nourriture que le premier sauf qu'il est plus riche en lipides, en glucides et en cholestérol. Pour chaque individu soumis à l'un des deux régimes, on a recueilli le poids perdu (ou gagné) au bout de 3 semaines. Effectuez un test d'hypothèse pour déterminer si le nouveau régime a un effet bénéfique. Utilisez $\alpha = 5\%$.

<i>Nouveau régime</i>	$\bar{X}_1 = 9.31; \quad S_1 = 4.668$
<i>Régime régulier</i>	$\bar{X}_2 = 7.40; \quad S_2 = 4.035$

Solution :

7.6 Test sur données appariées

Supposons que l'on cherche à comparer deux crèmes solaires A et B. Il y a au moins deux manières d'effectuer l'expérience :

(i) On sélectionne un échantillon d'individus de taille n_1 à qui on administrera la crème solaire A et on sélectionne un deuxième échantillon de taille n_2 , indépendant du premier, à qui on administrera la crème solaire B. Comme on l'a fait en section 7.5, on peut effectuer un test d'hypothèse du type $H_0 : \mu_1 - \mu_2 = 0$ contre l'alternative $H_1 : \mu_1 - \mu_2 \neq 0$, où μ_1 est une mesure du dommage moyen qu'a subie la peau soumise aux rayons du soleil pour la crème solaire A et μ_2 est une mesure du dommage moyen qu'a subie la peau soumise aux rayons du soleil pour la crème solaire B.

Le problème dans ce cas est que le test pourrait ne pas être équitable pour l'une des deux crèmes solaires si, par exemple, les individus à qui on administre la crème A ont la peau plus foncée que ceux à qui on administre la crème B, sachant que les peaux foncées sont moins sensibles aux effets du soleil.

(ii) Une meilleure manière d'effectuer cette expérience est de sélectionner un seul échantillon d'individus. A chaque individu, on administrera les deux crèmes. Par exemple, la crème A sur le coté droit du corps et la crème B sur le coté gauche. Pour chaque individu, on recueillera deux données (une paire d'observations). La première étant une mesure du dommage qu'a subie la peau soumise aux rayons du soleil pour la crème solaire A et la deuxième étant une mesure du dommage qu'a subie la peau soumise aux rayons du soleil pour la crème solaire B. On pourra alors tester s'il y a une différence entre les deux crèmes solaires, **Mais attention, on ne peut appliquer les techniques présentées à la section 7.5 car les deux ensembles de données ne proviennent pas d'échantillons indépendants!**

On examinera les différences entre chaque paire d'observations, raison pour laquelle on appelle ce type de tests « **test pour données appariées** ». On supposera dans cette section que la taille d'échantillon est grande (> 30).

Soit X_1, X_2, \dots, X_n et Y_1, Y_2, \dots, Y_n deux séries d'observations disponibles pour n individus. Pour l'individu i , on observe donc la paire (X_i, Y_i) , $i = 1, 2, \dots, n$. On travaillera avec les différence $D = Y_i - X_i$.

Par exemple, dans l'exemple précédent, X_i représente le dommage qu'a subie la peau soumise aux rayons du soleil pour la crème solaire B et Y_i représente le dommage qu'a subie la peau soumise aux rayons du soleil pour la crème solaire A.

On cherche à tester $H_0 : \mu_D = 0$ contre l'alternative $H_1 : \mu_D \neq 0$. **On est donc ramené à un test pour une moyenne tel que décrit dans la section 7.4.**

Exemple 7.7 Les expériences sur la perception extrasensorielle sont souvent faites à l'aide des "Cartes de Zener". C'est un jeu de 5 cartes distinctes. Récemment, plusieurs recherches ont été faites pour déterminer si l'hypnose n'aurait pas pour effet de faciliter la perception extrasensorielle. Nous décrivons ici une expérience faite avec 15 étudiants. On a demandé à chacun d'eux de deviner l'identité de 200 cartes de Zener. Aux premiers 100 essais, l'étudiant était dans un état normal. Aux 100 essais suivants il était hypnotisé. A chaque essai le "message" était envoyé par

un même « émetteur ». L'émetteur était lui-même hypnotisé lorsque l'étudiant l'était. On prend note du nombre de réponses correctes parmi 100. Supposez que la taille $n = 15$ de l'échantillon est assez grande pour que s'appliquent les théorèmes limites. Voici les résultats :

<i>Étudiant</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
<i>État normal</i>	18	19	16	21	16	20	20	14	11	22	19	29	16	27	15
<i>Sous hypnose</i>	25	20	26	26	20	23	14	18	18	20	22	27	19	27	21

Peut-on conclure que l'hypnose facilite la perception extrasensorielle ?

7.7 Tests sur une proportion

L'exemple 7.2 traite du test d'hypothèse concernant une proportion. La procédure utilisée, basée sur la loi binomiale est plutôt onéreuse lorsque les observations sont nombreuses. Lorsque n est grand, nous pouvons nous servir de la loi normale comme approximation de la distribution de $\hat{p} = X / n$. Considérons l'hypothèse

$$H_0 : p = p_0$$

et l'alternative

$$H_1 : p \neq p_0.$$

La région critique peut s'exprimer en fonction de la statistique centrée-réduite

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0) / n}},$$

qui suit à peu près une loi $\mathfrak{N}(0 ; 1)$ sous H_0 .

On rejette donc H_0 quand

(i) $Z > z_\alpha$ si l'hypothèse alternative est de la forme :

$$H_1 : p > p_0.$$

(ii) $Z < -z_\alpha$ si l'hypothèse alternative est de la forme :

$$H_1 : p < p_0.$$

(iii) $|Z| > z_{\alpha/2}$ si l'hypothèse alternative est de la forme :

$$H_1 : p \neq p_0.$$

On peut également utiliser l'approche par la p -valeur pour prendre une décision.

Exemple 7.8 Une enquête auprès d'utilisateurs de drogues injectables, a montré que parmi les 423 personnes interrogées, 18 étaient infectées par le virus du VIH. Des chercheurs veulent savoir si on peut conclure que moins de 5% des utilisateurs de drogues injectables dans la population sont infectés par le virus du VIH. Utilisez $\alpha = 5\%$.

Solution :

7.8 Test sur la différence entre deux proportions

Considérons une population qui est divisée en deux-sous populations : Population 1 et Population 2. On tire un échantillon aléatoire de taille n_1 dans la Population 1 et un échantillon aléatoire de taille n_2 dans la Population 2 (Voir Figure 6.1). On supposera que **les deux échantillons sont indépendants**. Soient p_1 la proportion d'individus dans la population 1 qui possède une certaine caractéristique et p_2 ladite proportion d'individus dans la population 2 qui possède la caractéristique. On supposera que les tailles n_1 et n_2 sont grandes (chacune ≥ 30). Soient \hat{p}_1 et \hat{p}_2 les proportions observés dans les échantillons tirés de la population 1 et de la population 2, respectivement, d'individus qui possèdent la caractéristique d'intérêt.

On veut tester l'hypothèse

$$H_0 : p_1 = p_2$$

Le test devrait normalement être basé sur la statistique

$$\frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1 - \hat{p}_2}}.$$

Mais puisque l'écart-type

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

dépend des inconnues p_1 et p_2 , nous devons remplacer $\sigma_{\hat{p}_1 - \hat{p}_2}$ par

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

La statistique

$$Z_{\hat{p}_1 - \hat{p}_2} = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_{\hat{p}_1 - \hat{p}_2}}$$

suit à peu près une loi $\mathcal{N}(0 ; 1)$ sous H_0 (Pourquoi?).

Donc on rejette H_0 quand

(i) $Z_{\hat{p}_1 - \hat{p}_2} > z_\alpha$ si l'hypothèse alternative est de la forme :

$H_1 : p_1 > p_2$.

(ii) $Z_{\hat{p}_1 - \hat{p}_2} < -z_\alpha$ si l'hypothèse alternative est de la forme : $H_1 : p_1 < p_2$.

(iii) $|Z_{\hat{p}_1 - \hat{p}_2}| > z_{\alpha/2}$ si l'hypothèse alternative est de la forme : $H_1 : p_1 \neq p_2$.

On peut également prendre une décision en utilisant l'approche par la p -valeur.

Exemple 7.9 Durant la dernière décennie, les campagnes anti-tabac, financées par les gouvernements provinciaux et fédéraux, se sont multipliées. Supposons que la Société Canadienne du Cancer a tiré un échantillon de 1500 individus en 1997 et qu'en 2007, elle tire de nouveau un échantillon d'individus afin de déterminer si la proportion de fumeurs au Canada a diminué. Soient X_1 le nombre de fumeurs dans l'échantillon en 1997 et X_2 le nombre de fumeurs dans l'échantillon en 2007. Les résultats des deux enquêtes sont exhibés ci-dessous. Les données indiquent-elles que la proportion de fumeurs au Canada a diminué durant cette période de 10 ans? Utilisez $\alpha = 5\%$.

1997	$n_1 = 1500;$ $X_1 = 555$
2007	$n_2 = 1750;$ $X_2 = 578$

Solution :

Chapitre 8

Inférence pour de petits échantillons

8.1 Introduction

Aux chapitres 6 et 7, nous avons introduit les notions d'intervalles de confiance et de tests d'hypothèses pour de grands échantillons, ce qui nous a permis de faire appel au théorème limite central pour déterminer la distribution d'une statistique. En pratique, on est parfois confronté à des échantillons de petite taille. En effet, le coût de collecte peut, dans certaines situations, être élevé. Par exemple, dans le domaine pharmaceutique, on soumet un échantillon d'individus à des tests médicaux qui peuvent s'avérer très coûteux. Dans ce cas, les échantillons sont habituellement de petite taille afin de respecter les budgets de l'étude. Malheureusement, on ne peut plus faire appel aux théorèmes limites dans le cas de petits échantillons. On se contentera de supposer que **les observations proviennent de populations normales (ce qui n'était pas nécessaire dans un contexte de grands échantillons)**.

8.2 Inférence pour μ

Supposons encore que nous disposons d'un échantillon X_1, \dots, X_n d'une population $\mathfrak{N}(\mu ; \sigma^2)$ d'écart-type σ inconnu.

Si n est grand, nous avons vu à la section 7.4 que la statistique

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}_{\bar{X}}} = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

suit à peu près une loi normale de moyenne 0 et de variance 1. Qu'en est-il si n n'est pas grand?

Théorème 8.1 Si \bar{X} et S^2 sont la moyenne et la variance d'un échantillon aléatoire de taille n tiré d'une population $\mathfrak{N}(\mu ; \sigma^2)$, alors

1) Les variables aléatoires \bar{X} et S^2 sont indépendantes.

$$2) \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Nous ne démontrerons pas ces résultats. À partir des résultats 1) et 2), il est aisé de montrer que

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}_{\bar{X}}} \sim t_{n-1},$$

car on peut écrire T comme

$$\frac{\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right)}{\sqrt{S^2 / \sigma^2}}.$$

Le numérateur est une $\mathfrak{N}(0 ; 1)$ et le dénominateur est

bien de la forme $\sqrt{\chi_{n-1}^2 / (n-1)}$ (voir section 4.7).

On désigne par $t_{v;a}$ le point qui a une probabilité a d'être excédé par une variable T de Student à v degrés de liberté, c'est-à-dire,

$$P\{T > t_{v;a}\} = a$$

Un intervalle de confiance pour μ à $100(1- \alpha)\%$ est donné par

$$\boxed{\bar{X} - t_{n-1;\alpha/2} \hat{\sigma}_{\bar{X}} \leq \mu \leq \bar{X} + t_{n-1;\alpha/2} \hat{\sigma}_{\bar{X}}}$$

Supposons maintenant que nous voulons tester l'hypothèse $H_0 : \mu = \mu_0$ contre l'une des alternatives $H_1 : \mu < \mu_0$, $H_1 : \mu > \mu_0$, $H_1 : \mu \neq \mu_0$.

Encore une fois, la statistique

$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}_{\bar{X}}}$$

est de loi de Student à $(n - 1)$ degrés de liberté lorsque H_0 est vraie. Il suffit donc de remplacer le point critique z_α (voir section 7.4) par $t_{n-1;\alpha}$. Le tableau suivant résume la procédure dans le cas de tests unilatéraux ou bilatéraux.

<i>Hypothèses</i>	<i>Région critique</i>	<i>p-valeur</i>
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$\frac{\bar{X} - \mu_0}{\hat{\sigma}_{\bar{X}}} < -t_{n-1;\alpha}$	$P(T_{n-1} \leq t_0 \mid \mu = \mu_0)$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$\frac{\bar{X} - \mu_0}{\hat{\sigma}_{\bar{X}}} > t_{n-1;\alpha}$	$P(T_{n-1} \geq t_0 \mid \mu = \mu_0)$
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$\left \frac{\bar{X} - \mu_0}{\hat{\sigma}_{\bar{X}}} \right > t_{n-1;\alpha/2}$	$P(T_{n-1} \geq t_0 \mid \mu = \mu_0)$

***Dans le tableau t_0 désigne la statistique du test.**

Exemple 8.1 Des expériences passées ont permis de déterminer que le temps moyen de sommeil des gens est de 7,7 heures. Une compagnie pharmaceutique, voulant tester la valeur d'un nouveau somnifère, a effectué des expériences. Un échantillon de taille 10, où le somnifère a été utilisé, a donné les résultats suivants :

7,8 8,3 7,2 9,1 8,4 6,8 7,3 7,7 8,9 9,2

Le nouveau somnifère a-t-il un effet sur la durée du sommeil? Utilisez $\alpha = 5\%$.

Solution :

8.3 Inférence pour deux moyennes

Plusieurs enquêtes et plusieurs expériences scientifiques ont pour but de déterminer s'il y a une différence entre les moyennes de deux populations.

Le modèle mathématique est le suivant : on dispose de deux échantillons indépendants, $X_{11}, X_{12}, \dots, X_{1n_1}$ et $X_{21}, X_{22}, \dots, X_{2n_2}$. Nous supposons que ces observations sont indépendantes et, de plus, nous faisons les suppositions suivantes :

$$X_{1i} \sim \mathfrak{N}(\mu_1; \sigma_1^2), \quad i = 1, 2, \dots, n_1$$

$$X_{2j} \sim \mathfrak{N}(\mu_2; \sigma_2^2), \quad j = 1, 2, \dots, n_2$$

Soit \bar{X}_1 et \bar{X}_2 les moyennes de deux échantillons de taille n_1 et n_2 respectivement, tirés de deux populations de moyennes μ_1 et μ_2 et de variance σ_1^2 et σ_2^2 respectivement.

Lorsque n_1 et n_2 sont grands, la statistique

$$Z_{\bar{X}_1 - \bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

suit à peu près une loi normale de moyenne 0 et de variance 1 (voir section 7.5). Qu'en est-il si n_1 et n_2 sont petits?

Malheureusement, la statistique $Z_{\bar{X}_1 - \bar{X}_2}$ ne suit pas une loi de Student. Il nous faudra faire une hypothèse supplémentaire : celle qui consiste à supposer les variances des deux populations sont égales; c'est-à-dire, $\sigma_1^2 = \sigma_2^2$.

Si le contexte nous permet de supposer que **les populations sont normales et qu'elles ont la même variance** σ^2 , on peut déterminer des intervalles de confiance ou construire des tests en faisant appel à une loi de Student appropriée. Si $\sigma_1^2 = \sigma_2^2 = \sigma^2$, les statistiques S_1^2 et S_2^2 sont toutes deux des estimateurs sans biais du même paramètre σ^2 et pour trouver un estimateur commun, il faudra prendre une moyenne pondérée de S_1^2 et S_2^2 .

On peut montrer que

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

est un estimateur sans biais de σ^2 et qu'il est optimal (la démonstration est laissée en exercice).

Sachant que $\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi_{n_1 - 1}^2$ et $\frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi_{n_2 - 1}^2$,

on peut montrer (la démonstration est laissée en exercice) que

$$\frac{(n_1 + n_2 - 2)S^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi_{n_1 + n_2 - 2}^2.$$

Finalement, la statistique

$$T_{\bar{X}_1 - \bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit une loi de Student avec $n_1 + n_2 - 2$ degrés de libertés.

La démonstration est laissée en exercice.

Ce qui précède nous permet maintenant de construire des intervalles de confiance ou des tests d'hypothèse.

Un intervalle de confiance à $100(1 - \alpha)\%$ pour $\mu_1 - \mu_2$ est donné par

$$(\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2; \alpha/2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2; \alpha/2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Supposons que μ_1 et μ_2 sont inconnues et qu'on veuille tester l'hypothèse que, pour un nombre δ donné,

$$H_0 : \mu_1 - \mu_2 = \delta$$

contre l'une des alternatives

$$H_1 : \mu_1 - \mu_2 \neq \delta$$

$$H_1 : \mu_1 - \mu_2 > \delta$$

$$H_1 : \mu_1 - \mu_2 < \delta$$

Comme au chapitre 7, notons que le cas $\delta = 0$ est le plus courant.

On rejettera H_0 lorsque

(i) $T_{\bar{X}_1 - \bar{X}_2} > t_{n_1+n_2-2; \alpha}$ si l'hypothèse alternative est de la forme : $H_1: \mu_1 - \mu_2 > \delta$.

(ii) $T_{\bar{X}_1 - \bar{X}_2} < -t_{n_1+n_2-2; \alpha}$ si l'hypothèse alternative est de la forme : $H_1: \mu_1 - \mu_2 < \delta$.

(iii) $|T_{\bar{X}_1 - \bar{X}_2}| > t_{n_1+n_2-2; \alpha/2}$ si l'hypothèse alternative est de la forme : $H_1: \mu_1 - \mu_2 \neq \delta$.

On peut également utiliser l'approche par la p -valeur pour prendre une décision.

Remarque : Le test précédent suppose que les deux échantillons sont indépendants. Dans le cas de données appariées (voir section 7.6), on travaillera sur les différences, ce qui nous ramènera à un test pour une moyenne tel que décrit dans la section 8.2.

Exemple 8.2 Un jardinier amateur veut savoir si l'engrais qu'il utilise est vraiment efficace. Pour ce faire, il a privé d'engrais 2 de ses plants de tomates, choisis a hasard en début de saison, et n'a donné de l'engrais qu'aux 6 autres plants. Les plants sans engrais ont fourni respectivement 12,3 et 13,6 kg de tomates. Pour les plants traités à l'engrais, les résultats, en kg, ont été :

14,1 12,8 15,1 13,7 13,4 15,4

L'engrais a-t-il un effet sur la production de tomates?
Utilisez $\alpha = 10\%$.

Solution :

Chapitre 9

Tests du khi-deux

Nous discutons ici deux tests basés sur la loi du khi-deux: **le test d'ajustement et le test d'indépendance**. Ce sont des techniques utilisées souvent avec des données qualitatives: le test d'ajustement est employé pour **une seule variable qualitative**, alors que le test d'indépendance s'applique à la distribution conjointe de **deux variables qualitatives**.

9.1 Tests d'ajustement

Supposons qu'on prélève des données afin de déterminer si le jour de la semaine a une influence sur les suicides. L'hypothèse nulle est que les suicides ne sont pas affectés par le jour de la semaine :

H_0 : Les suicides sont également fréquents tous les jours de la semaine

On prélève un échantillon de 780 suicides, puis on les classe en 4 catégories, selon le moment de la semaine où le suicide a eu lieu : le début de la semaine (lundi), le milieu de la semaine (mardi à jeudi), la fin de la semaine (vendredi), et le week-end (samedi-dimanche). Supposons qu'on obtienne la distribution suivante :

Tableau 9.1 : Effectifs observés du nombre de suicides

<i>Jour</i>	<i>Lundi</i>	<i>Mardi-jeudi</i>	<i>Vendredi</i>	<i>Samedi-dimanche</i>	<i>Total</i>
<i>Effectif observé</i>	110	320	100	250	780

Cette distribution est appelée **distribution observée**.

Si l'hypothèse nulle est vraie, les 780 suicides devraient se répartir de façon proportionnelle au nombre de jours: $1/7$ le lundi; $3/7$ les mardi-jeudi; $1/7$ le vendredi; et $2/7$ les samedi-dimanche. Cette répartition, appelée **distribution théorique**, est présentée dans le tableau 9.2.

Tableau 9.2 : Effectifs théoriques du nombre de suicides

<i>Jour</i>	<i>Lundi</i>	<i>Mardi-jeudi</i>	<i>Vendredi</i>	<i>Samedi-dimanche</i>	<i>Total</i>
<i>Effectif théorique</i>	111,4	334,3	111,4	222,9	780

L'écart entre les deux tableaux nous permettra de rejeter H_0 ou pas. Si l'écart est important, on rejette H_0 .

Écart entre les effectifs observés et théoriques

L'écart entre les deux tableaux est mesuré par une quantité dénotée par χ^2 et définie par

$$\chi^2 = \sum_i \frac{(O_i - T_i)^2}{T_i},$$

où les O_i sont les effectifs observés et les T_i sont les effectifs théoriques.

Dans l'exemple précédent, on a

$$\begin{aligned} \chi^2 &= \frac{(110 - 111,4)^2}{111,4} + \frac{(320 - 334,3)^2}{334,3} + \frac{(100 - 111,4)^2}{111,4} \\ &\quad + \frac{(250 - 222,9)^2}{222,9} \\ &= 5,09. \end{aligned}$$

Une distance entre les distributions théoriques et observées de 5,09 est-elle assez grande pour nous permettre de rejeter H_0 ?

Région critique

La région critique est basée sur χ^2 . Ce sont les grandes valeurs de χ^2 qui devraient mener au rejet de H_0 car un grand χ^2 reflète une grande différence entre les deux tableaux. χ^2 est grand lorsque les effectifs observés s'écartent beaucoup des effectifs théoriques, c'est-à-dire, des effectifs auxquels on s'attend lorsque H_0 est vraie. La région critique sera donc de la forme

$$\chi^2 > C.$$

La constante C doit être choisie de telle sorte que la probabilité d'une erreur de première espèce soit à peu près égale à un certain nombre α . Il est coutume d'utiliser des valeurs de α égales à 1%, 5% ou 10%.

Il faut donc que

$$P(\chi^2 > C \mid H_0) = \alpha.$$

Pour déterminer C , il faut connaître la distribution de χ^2 sous H_0 .

Théorème 9.1 Lorsque H_0 est vraie, la statistique χ^2 suit à peu près une loi χ^2_v , où $v = (\text{nombre de cases}) - 1$.

Donc la région critique est $\chi^2 > \chi^2_{v;\alpha}$.

Dans l'exemple des suicides, $v = 3$ et si on prend $\alpha = 0,05$, le point critique est $\chi^2_{3;0,05} = 7,8147$. Puisque $\chi^2 = 5,08 \leq 7,8147$, on ne rejette pas H_0 : on ne peut pas affirmer que le taux de suicide varie selon le jour de la semaine. L'écart entre les deux tableaux (observés et théoriques) pourrait bien être dû au hasard tout seul.

Le modèle

Nous allons formuler maintenant le modèle sur lequel ce test est basé. Nous le traiterons dans le cadre de l'exemple. La généralisation se fera sans difficulté. Les 4 effectifs observés O_1, O_2, O_3, O_4 sont les valeurs d'un vecteur aléatoire $\mathbf{X} = (X_1, X_2, X_3, X_4)$ qui suit une loi multinomiale de paramètres n ($= 780$ dans l'exemple) et (p_1, p_2, p_3, p_4) :

Modèle: $\mathbf{X} = (X_1, X_2, X_3, X_4) \sim MN(n; p_1, p_2, p_3, p_4)$

L'hypothèse nulle est :

$$H_o : p_1 = \frac{1}{7}, \quad p_2 = \frac{3}{7}, \quad p_3 = \frac{1}{7}, \quad p_4 = \frac{2}{7}$$

Les effectifs théoriques T_1, T_2, T_3, T_4 sont les espérances des variables X_1, X_2, X_3, X_4 sous H_o :

$$T_1 = E(X_1 | H_o) = n \frac{1}{7} = 780 \frac{1}{7}; \quad T_2 = E(X_2 | H_o) = n \frac{3}{7} = 780 \frac{3}{7}$$

$$T_3 = E(X_3 | H_o) = n \frac{1}{7} = 780 \frac{1}{7}; \quad T_4 = E(X_4 | H_o) = n \frac{2}{7} = 780 \frac{2}{7}$$

En général, le problème est le suivant. Les observations constituent un vecteur de loi multinomiale :

$$\text{Modèle: } \mathbf{X} = (X_1, X_2, \dots, X_k) \sim MN(n; p_1, p_2, \dots, p_k)$$

L'hypothèse nulle est de la forme

$$H_0: p_1 = p_{10}, \quad p_2 = p_{20}, \quad \dots, \quad p_k = p_{k0}$$

où $p_{10}, p_{20}, \dots, p_{k0}$ sont des nombres positifs tels que

$\sum_{i=1}^k p_{i0} = 1$. La statistique χ^2 peut s'écrire

$$\chi^2 = \sum_{i=1}^k \frac{(X_i - np_{i0})^2}{np_{i0}}$$

9.2 Tests d'indépendance

Ici aussi nous présenterons un cas particulier avant de décrire la procédure formellement. Considérons les données suivantes sur deux variables, la « scolarité » et « l'attitude face à l'avortement ».

Tableau 9.3 : Distribution conjointe (fréquences) des variables « scolarité » et « attitude face à l'avortement ».

		<i>Attitude face à l'avortement</i>			
		<i>Pour</i>	<i>Mixte</i>	<i>Contre</i>	
<i>Scolarité</i>	≤ 8	31	23	56	110
	9 — 12	171	89	177	437
	> 12	116	39	74	229
<i>Total</i>		318	151	307	776

Ces données ont été recueillies afin de déterminer s'il y a une relation entre les deux variables. Le tableau précédent présente **la distribution conjointe** observée des variables « scolarité » et « attitude face à l'avortement ». Une distribution conjointe peut également s'exprimer en fréquences relatives comme dans le tableau 9.4.

Tableau 9.4 : Distribution conjointe (fréquences relatives) des variables « scolarité » et « attitude face à l’avortement ».

		<i>Attitude face à l’avortement</i>			
		<i>Pour</i>	<i>Mixte</i>	<i>Contre</i>	
<i>Scolarité</i>	≤ 8	0,039	0,029	0,072	0,141
	9 — 12	0,220	0,114	0,228	0,563
	> 12	0,149	0,050	0,095	0,296
<i>Total</i>		0,409	0,196	0,396	1,000

Le tableau 9.3 (ou 9.4) exhibe la distribution conjointe des variables « scolarité » et « attitude face à l’avortement » et contient donc plus d’information que contiendraient deux tableaux donnant l’un la distribution de la variable « scolarité » et l’autre, celle de la variable « attitude face à l’avortement ».

Ces deux distributions, appelées **distributions marginales**, se retrouvent intégralement dans les marges du tableau 9.3 (ou 9.4).

La distribution marginale de la variable « scolarité » est exhibée dans le tableau 9.5.

Tableau 9.5 : Distribution marginale de la variable « scolarité » .

<i>Scolarité</i>	≤ 8	9 — 12	> 12	<i>Total</i>
<i>Fréquence relative</i>	0,141	0,563	0,296	1,000

La distribution marginale de la variable « attitude face à l'avortement » est exhibée dans le tableau 9.6.

Tableau 9.6 : Distribution marginale de la variable « attitude face à l'avortement » .

<i>Scolarité</i>	<i>Pour</i>	<i>Mixte</i>	<i>Contre</i>	<i>Total</i>
<i>Fréquence relative</i>	0,409	0,196	0,396	1,000

La distribution conjointe de deux variables contient toute l'information nécessaire à l'étude de la relation entre les variables. Mais elle ne met pas cette relation clairement en évidence. La notion de relation ou de dépendance entre deux variables s'exprime en termes de distributions conditionnelles. Une distribution conditionnelle de la variable « attitude face à l'avortement » est la distribution de cette variable confinée à une tranche de la population, cette tranche étant définie par une valeur de la variable « scolarité ».

Par exemple, la distribution conditionnelle de la variable « attitude face à l'avortement » étant donné que la variable scolarité = 9-12 est la distribution de la variable « attitude face à l'avortement » limitée à l'ensemble des individus qui ont une scolarité d'une durée 9 à 12 ans.

Le tableau 9.7 donne la distribution conditionnelle de l'attitude étant donné chaque niveau de scolarité :

Tableau 9.7 : Distribution conditionnelle de la variable « attitude face à l'avortement » étant donné la variable « scolarité » .

		<i>Attitude face à l'avortement</i>			
		<i>Pour</i>	<i>Mixte</i>	<i>Contre</i>	
<i>Scolarité</i>	≤ 8	0,28	0,21	0,51	1
	9 — 12	0,39	0,20	0,41	1
	> 12	0,51	0,17	0,32	1

Une certaine dépendance se manifeste clairement dans ces distributions conditionnelles. La question est de savoir si cette dépendance, évidente au niveau de l'échantillon, existe aussi au niveau de la population.

L'hypothèse nulle est formulée comme suit :

H_0 : Les variables « scolarité » et « attitude face à l'avortement » sont indépendantes;

L'hypothèse alternative est :

H_1 : non H_0 .

Les observations dans les 9 cases du tableau 9.3 sont une réalisation de 9 variables aléatoires

$$\begin{array}{ccc} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{array}$$

qui suivent conjointement une loi multinomiale:

Modèle:

$$\mathbf{X} = (X_{11}, X_{12}, X_{13}, X_{21}, X_{22}, X_{23}, X_{31}, X_{32}, X_{33}) \sim \mathcal{M}(n; p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33})$$

Le tableau 9.8 présente les probabilités p_{11} , p_{12} , p_{13} , p_{21} , p_{22} , p_{23} , p_{31} , p_{32} , p_{33} . Le tableau présente également certaines fonctions de ces probabilités, soit $p_{1.}$, $p_{2.}$, $p_{3.}$, $p_{.1}$, $p_{.2}$, $p_{.3}$, les sommes des probabilités des lignes et des colonnes.

Tableau 9.8 : Vraie distribution conjointe des variables des variables « scolarité » et « attitude face à l'avortement ».

		<i>Attitude face à l'avortement</i>			
		<i>Pour</i>	<i>Mixte</i>	<i>Contre</i>	
<i>Scolarité</i>	≤ 8	p_{11}	p_{12}	p_{13}	$p_{1.}$
	9 — 12	p_{21}	p_{22}	p_{23}	$p_{2.}$
	> 12	p_{31}	p_{32}	p_{33}	$p_{3.}$
<i>Total</i>		$p_{.1}$	$p_{.2}$	$p_{.3}$	1

En fonction de ces paramètres, l'hypothèse nulle est

$$H_0: p_{ij} = p_{i.} \times p_{.j} \text{ pour tout } i, j$$

Puisque, sous H_0 , $E(X_{ij}) = np_{i.}p_{.j}$, la statistique χ^2 devrait normalement être

$$\chi^2 = \sum_i \sum_j \frac{(X_{ij} - np_{i.}p_{.j})^2}{np_{i.}p_{.j}}$$

Mais les $p_{i.}$ et les $p_{.j}$ ne sont pas connus; elles devront donc être estimées. Les estimateurs $\hat{p}_{i.}$ et $\hat{p}_{.j}$ sont donnés par

$$\hat{p}_{i.} = \frac{1}{n} \sum_j X_{ij} \quad \hat{p}_{.j} = \frac{1}{n} \sum_i X_{ij}$$

Les estimations des p_i et des p_j ainsi que les effectifs théoriques sont présentés dans le tableau 9.9.

Tableau 9.9 : Effectifs théoriques

		<i>Attitude face à l'avortement</i>			
		<i>Pour</i>	<i>Mixte</i>	<i>Contre</i>	
<i>Scolarité</i>	≤ 8	$776 \hat{p}_{1.\hat{p}.1}$	$776 \hat{p}_{1.\hat{p}.2}$	$776 \hat{p}_{1.\hat{p}.3}$	$\hat{p}_{1.}$ =110/776
	9 — 12	$776 \hat{p}_{2.\hat{p}.1}$	$776 \hat{p}_{2.\hat{p}.2}$	$776 \hat{p}_{2.\hat{p}.3}$	$\hat{p}_{2.}$ =437/776
	> 12	$776 \hat{p}_{2.\hat{p}.1}$	$776 \hat{p}_{3.\hat{p}.2}$	$776 \hat{p}_{3.\hat{p}.3}$	$\hat{p}_{3.}$ =229/776
<i>Total</i>		$\hat{p}_{.1}$ =318/776	$\hat{p}_{.2}$ =151/776	$\hat{p}_{.3}$ =307/776	1

Les calculs donnent :

Tableau 9.10 : Effectifs théoriques

		<i>Attitude face à l'avortement</i>			
		<i>Pour</i>	<i>Mixte</i>	<i>Contre</i>	
<i>Scolarité</i>	≤ 8	45,08	21,40	43,52	110
	9 — 12	179,08	85,03	172,89	437
	> 12	93,84	44,56	90,60	229
<i>Total</i>		318	151	307	776

La valeur de la statistique χ^2 est

$$\chi^2 = \sum_{i=1}^3 \sum_{j=1}^3 \frac{(X_{ij} - np_{i \cdot} p_{\cdot j})^2}{np_{i \cdot} p_{\cdot j}} = 17,7$$

Théorème 9.2 Lorsque H_0 est vraie, la statistique χ^2 suit à peu près une loi χ^2_v , où $v = (\ell-1)(c-1)$, ℓ étant le nombre de lignes, c le nombre de colonnes, du tableau.

La région critique est donc

$$\chi^2 > \chi^2_{(\ell-1)(c-1); \alpha}$$

Dans l'exemple, puisque $\chi^2 = 17,7 > \chi^2_{4;0,05} = 9,4877$, nous rejetons H_0 à 5%. Nous concluons qu'il y a vraiment une dépendance entre la scolarité et l'attitude face à l'avortement.

Autre modélisation

Dans l'exemple de cette section, le chercheur a choisi 776 sujets et les a ensuite classés selon la scolarité et l'attitude. Par conséquent, tous les effectifs observés étaient aléatoires, y compris les effectifs des marges. Mais il existe des situations où les effectifs des marges sont fixés. Dans ce cas, le modèle d'une multinomiale n'est pas valide. Par exemple, supposons qu'on veuille savoir si la durée d'une hospitalisation pour une certaine maladie dépend de l'hôpital; et que pour ce faire, on choisit un certain nombre fixe de patients dans chaque hôpital, et qu'on recueille les données suivantes :

Tableau 9.11 : Distribution conjointe des variables « Durée de l'hospitalisation » et « Hôpital ».

		<i>Durée de l'hospitalisation</i>			
		<i>1-2 jours</i>	<i>3-5 jours</i>	<i>plus de 5 jours</i>	
<i>Hôpital</i>	1	40	20	40	100
	2	50	35	65	150
	3	95	45	60	200

Visiblement, les effectifs de la marge de droite sont fixes: on a décidé, avant de prélever les données, qu'on prendrait respectivement 100, 155 et 200 cas dans les trois hôpitaux. Ce ne sont pas des variables aléatoires. Dans le tableau 9.12, on nomme les variables observées.

Tableau 9.12 : Distribution conjointe des variables « Durée de l'hospitalisation » et « Hôpital ».

		<i>Durée de l'hospitalisation</i>			
		<i>1-2 jours</i>	<i>3-5 jours</i>	<i>plus de 5 jours</i>	
<i>Hôpital</i>	1	X_1	Y_1	Z_1	n_1
	2	X_2	Y_2	Z_2	n_2
	3	X_3	Y_3	Z_3	n_3

Les 9 variables aléatoires dans le tableau ne suivent pas une loi multinomiale. Le modèle est plutôt

$$\begin{aligned} (X_1, Y_1, Z_1) &\sim MN(n_1 ; p_1, r_1, s_1), \\ (X_2, Y_2, Z_2) &\sim MN(n_2 ; p_2, r_2, s_2), \\ (X_3, Y_3, Z_3) &\sim MN(n_3 ; p_3, r_3, s_3). \end{aligned}$$

L'hypothèse nulle est

$$H_o : (p_1, r_1, s_1) = (p_2, r_2, s_2) = (p_2, r_2, s_2)$$

Malgré la différence entre ce modèle (trois multinomiales) et celui de l'exemple précédent (une seule multinomiale), la *procédure reste la même* : on utilise exactement le même test dans les deux cas.

9.3 Dépendance et causalité

Lorsque l'on étudie la dépendance entre deux variables, on envisage presque toujours la possibilité d'un **lien de causalité** entre les variables. Par exemple, plusieurs études établissent l'existence d'un lien entre l'état de santé et la consommation de cigarettes. Si on montre, au moyen de techniques statistiques, qu'il existe un lien entre l'état de santé et la consommation de cigarettes, on résiste difficilement à la tentation de conclure que la cigarette cause la maladie. Bien que cette conclusion soit vraisemblable, on court le risque d'identifier trop hâtivement une variable à la cause et l'autre à l'effet. Il est fort possible que la dépendance entre deux variables soit due, non pas à l'effet de l'une sur l'autre, mais à l'effet simultané d'une troisième variable sur les deux premières. Nous illustrons ce phénomène par un exemple fictif mais révélateur.

Une expérience a été effectuée avec 300 rats atteints d'une certaine maladie. Soit X la pression artérielle et Y une variable qui identifie les conséquences éventuelles de la maladie. L'expérience donne les résultats suivants :

Tableau 9.13 : Effectifs observés

	Y : conséquences de la maladie		
X : Pression artérielle	Succombe	Survit	Total
Élevée	136	44	180
Normale	64	56	120
Total	200	100	300

On constate que le taux de mortalité est de 75,6% parmi les rats qui ont une pression élevée alors qu'il n'est que de 53,3% parmi ceux ayant une pression normale. **Peut-on conclure que la pression élevée est la cause de ce taux de mortalité?**

C'est possible mais avant de tirer cette conclusion, on décide une fois de plus d'examiner les données à la lumière de l'information que l'on possède sur l'âge des rats. Il y a 100 qu'on classifie comme « jeune » et 200 qu'on classifie comme « vieux ». On dresse les tableaux suivants :

Tableau 9.14 : Effectifs observés chez les jeunes rats

	<i>Y</i> : conséquences de la maladie		
<i>X</i> : Pression artérielle	Succombe	Survit	Total
Élevée	8	12	20
Normale	32	48	80
Total	40	60	100

Tableau 9.15 : Effectifs observés chez les vieux rats

	<i>Y</i> : conséquences de la maladie		
<i>X</i> : Pression artérielle	Succombe	Survit	Total
Élevée	128	32	160
Normale	32	8	40
Total	160	40	200

Parmi les jeunes rats, le taux de mortalité est de 40%, quelle que soit la pression. Parmi les vieux, le taux de mortalité est de 80% quelle que soit la pression. Donc, contrairement à la conclusion suggérée par le tableau 9.13, la pression n'agit pas de façon directe sur la mortalité. C'est apparemment l'âge qui agit en même temps sur la pression et sur la mortalité.

Chapitre 10

Régression linéaire simple

10.1 Introduction

Au chapitre 9, nous avons effectué des tests d'hypothèse pour savoir si deux variables catégorielles étaient indépendantes ou non. Dans ce chapitre, on s'intéresse à la dépendance entre deux variables continues. Au Chapitre 1, nous avons présenté la droite des moindres carrés et le coefficient de corrélation comme techniques descriptives. Ces techniques permettent de décrire la relation entre deux variables à l'aide d'une droite, et de mesurer la force de la dépendance linéaire dans un échantillon. À l'époque, on n'avait pas encore fait la distinction entre échantillon et population. Dans plusieurs applications, cependant, on voudra tirer des conclusions à propos d'une dépendance dans la population, à partir de celle observée dans l'échantillon.

Pour ce faire, nous devons adopter un modèle pour décrire la population. Nous allons présenter dans ce chapitre un modèle appelé **régression linéaire simple**.

Exemple 10.1 Le Tableau 10.1 présente, pour un ensemble de 18 individus, les valeurs de deux variables :

x : le poids d'un individu, en kg

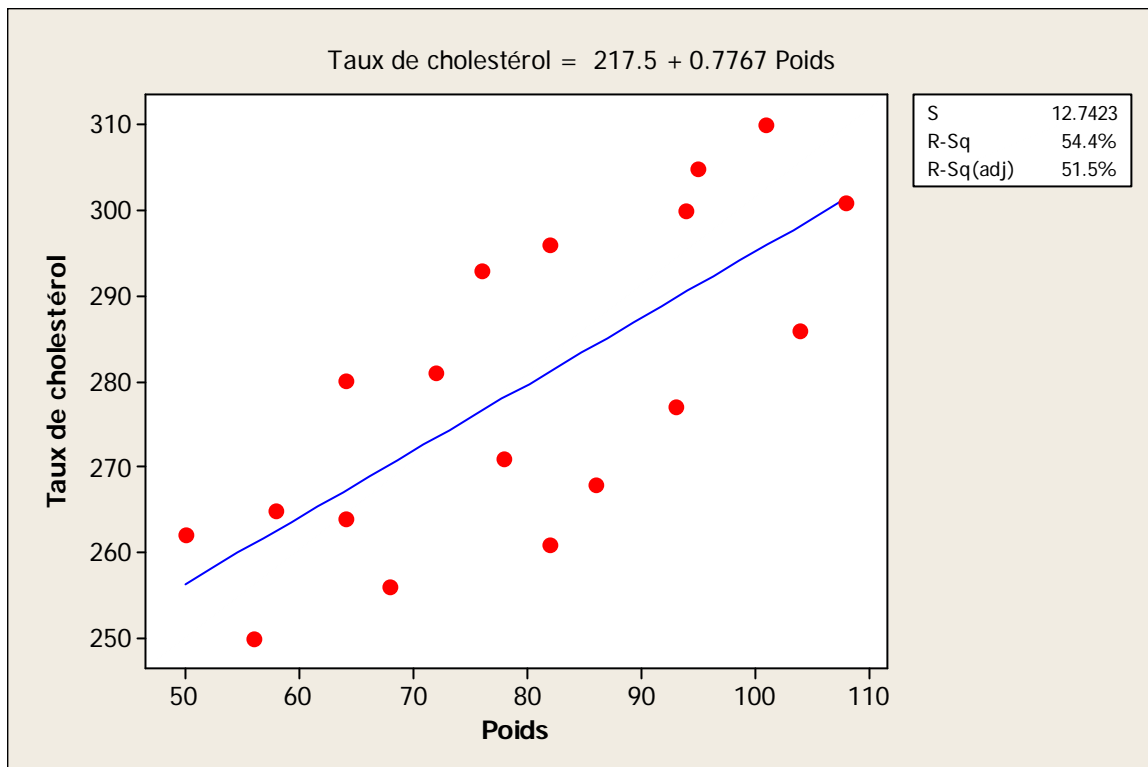
y : le taux de cholestérol, en mg par 100 ml

Tableau 10.1 Taux de cholestérol (y) et poids (x) de 18 individus

x	y	x	y
50	262	82	261
56	250	82	296
58	265	86	268
64	280	93	277
64	264	94	300
68	256	95	305
72	281	101	310
76	293	104	286
78	271	108	301

Ces données, avec la droite des moindres carrés sont représentées comme 18 points dans \mathbb{R}^2 dans la figure 10.1.

Figure 10.1 Relation entre le taux de cholestérol (y) et poids (x)



La figure 10.1 montre qu'il semble y avoir une relation entre le poids d'individu et son taux de cholestérol et que cette relation semble être linéaire.

Rappelons que la droite des moindres carrés est la droite

$$y = b_0 + b_1x$$

qui minimise

$$D = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

où $\hat{y}_i = b_0 + b_1x_i$. On obtient

$$b_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad b_0 = \bar{y} - b_1\bar{x}$$

Remarque : La notation a et b utilisée au chapitre 1 pour désigner les coefficients des moindres carrés est remplacée par la notation b_0 et b_1 .

Dans l'exemple 10.1, la droite des moindres carrés est donnée par

$$y = 217,5 + 0,7767x$$

10.2 Le modèle de régression linéaire simple

Afin de passer du pur descriptif à l'inférence, nous allons définir un modèle de population, c'est-à-dire, un ensemble de suppositions à propos de la population. Le modèle est appelé **modèle de régression simple**.

Pour chacune des n unités dans l'échantillon, on dispose d'une paire d'observations : (x_i, y_i) , $i = 1, \dots, n$. On suppose, dans ce modèle que les variables x et y sont liées selon

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (10.1)$$

- y est communément appelée **variable dépendante**.
C'est la variable que l'on cherche à modéliser, celle que l'on cherche à expliquer.

- x est la **variable indépendante** ou la **variable prédictrice**. C'est la variable utilisée pour prédire la variable dépendante y . Il est important de noter que les observations x_1, \dots, x_n de x sont considérées comme des constantes.
- β_0 représente l'ordonnée à l'origine et β_1 représente la pente de la droite. On utilise des lettres grecques pour représenter l'ordonnée à l'origine et la pente pour bien insister sur le fait que ce sont des **paramètres inconnus**. Leur valeur respective serait connue si on avait accès à toute la population, ce qui n'est jamais le cas en pratique. Il nous faudra donc les estimer.
- ε est une variable aléatoire qui est souvent appelée erreur aléatoire ou bruit aléatoire.

- Le modèle (10.1) peut être vu comme la somme de deux composantes : une composante déterministe (non-aléatoire), $\beta_0 + \beta_1 x_i$ et une composante aléatoire ε .

Il est important de comprendre qu'un modèle statistique est un ensemble d'hypothèses et le modèle de régression linéaire simple n'échappe pas à la règle.

Quelles sont les hypothèses sous-jacentes au modèle de régression linéaire simple?

- (1) La relation entre la variable dépendante y et la variable indépendante x est linéaire.
- (2) $E(\varepsilon_i) = 0$.

(3) $V(\varepsilon_i) = \sigma^2$. L'hypothèse que les ε_i ont toutes la même variance est appelée **hypothèse d'homoscédasticité**. Il est difficile de traiter un modèle qui n'inclue pas cette hypothèse. Notons que σ^2 est un paramètre inconnu qu'il nous faudra estimer.

(4) Les ε_i sont des variables aléatoires mutuellement indépendantes.

(5) La distribution des ε_i est normale. En combinant (2) et (3), on peut écrire $\varepsilon_i \sim \mathfrak{N}(0, \sigma^2)$.

En combinant (10.1) avec les hypothèse (2)-(5), on conclut que les y_i sont des variables aléatoires indépendantes normales de moyenne $E(y_i) = \beta_0 + \beta_1 x_i$ et de variance σ^2 . On a

$$y_i \sim \mathfrak{N}(\beta_0 + \beta_1 x_i; \sigma^2)$$

Remarque Voici une façon d'interpréter les suppositions du modèle dans le contexte de l'exemple 10.1. Pour chaque valeur de x , considérons l'ensemble des individus dont le poids est x_i . Les taux de cholestérol dans cette sous-population sont distribués selon une loi normale. Le taux de cholestérol pour ces individus dépend de x : il est égal à $\beta_0 + \beta_1 x$. La variance σ^2 est la dispersion des taux de cholestérol dans cette sous-population. On suppose que cette variance est la même pour toute sous-population constituée des individus ayant un même poids. C'est l'hypothèse d'homoscédasticité. Il est rare qu'elle soit vérifiée exactement en pratique, mais on ne s'attend pas à des effets très graves si les différences de variances ne sont pas très grandes.

10.3 Estimation des paramètres

Nous avons 3 paramètres à estimer : β_0 , β_1 , et σ^2 .

Il existe plusieurs critères possibles qui pourraient diriger la recherche d'estimateurs de β_0 et β_1 . Mais presque tous mènent aux mêmes estimateurs, et ceux-ci sont justement les quantités b_0 et b_1 définies plus haut. Ce sont des estimateurs sans biais de β_0 et β_1 , respectivement.

Proposition 10.1 : $E(b_0) = \beta_0$ et $E(b_1) = \beta_1$.

Démonstration :

Sous l'hypothèse que les y_i sont de loi normale, on peut démontrer la proposition suivante :

Proposition 10.2 : On a

$$(i) b_1 \sim \mathfrak{N}(\beta_1 ; \sigma_{b_1}^2), \text{ où } \sigma_{b_1}^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}.$$

$$(ii) b_0 \sim \mathfrak{N}(\beta_0 ; \sigma_{b_0}^2) \text{ où}$$

$$\sigma_{b_0}^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)$$

Il suit de (i) et de (ii) que

$$\frac{b_1 - \beta_1}{\sigma_{b_1}} \sim \mathfrak{N}(0 ; 1) \text{ et } \frac{b_0 - \beta_0}{\sigma_{b_0}} \sim \mathfrak{N}(0 ; 1).$$

Démonstration :

Il nous reste à estimer σ^2 et à estimer les variances de b_0 et b_1 données par la Proposition 10.2.

On acceptera sans démonstration qu'un estimateur sans biais de la variance σ^2 est donné par

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2}{n-2}$$

Remarque On peut justifier cet estimateur intuitivement. On sait que σ^2 est la variance des $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$, lesquelles sont de moyenne nulle. Un estimateur de σ^2 aurait donc été la moyenne des ε_i^2 , $\sum \varepsilon_i^2 / n$, si ce n'était que les ε_i^2 ne sont pas connus. Mais on peut toujours remplacer les ε_i par les estimateurs $\varepsilon = y_i - (b_0 + b_1 x_i)$, et c'est ce qu'on fait. Le dénominateur doit cependant changer car le nombre de degrés de liberté de la somme de carrés $\sum \hat{\varepsilon}_i^2 / n$ est $n-2$ et non plus n . Ce qui donne

l'estimateur $\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-2}$.

Une fois σ^2 estimée, nous pouvons facilement estimer les variances de b_0 et b_1 :

$$\hat{\sigma}_{b_1}^2 = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} \quad \text{et} \quad \hat{\sigma}_{b_0}^2 = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)$$

Dans l'exemple 10.1, on vérifiera que $\hat{\sigma}^2 = 162,36$,

$\hat{\sigma}_{b_1} = 14,36$ et $\hat{\sigma}_{b_0} = 0,7767$.

Distribution des statistiques de tests

Nous avons montré que $Z_1 = \frac{b_1 - \beta_1}{\sigma_{b_1}} \sim \mathfrak{N}(0; 1)$ et que

$Z_0 = \frac{b_0 - \beta_0}{\sigma_{b_0}} \sim \mathfrak{N}(0; 1)$. Lorsqu'on remplace les écarts-

types aux dénominateurs par leurs estimations, les variables qui en résultent suivent des lois de Student à $n - 2$ degrés de liberté:

$$T_1 = \frac{b_1 - \beta_1}{\hat{\sigma}_{b_1}} \sim t_{n-2}$$

et

$$T_0 = \frac{b_0 - \beta_0}{\hat{\sigma}_{b_0}} \sim t_{n-2}.$$

10.4 Inférence statistique; intervalles de confiance et tests d'hypothèse

En utilisant les résultats de la section 10.3, on peut construire des intervalles de confiance à $100(1-\alpha)\%$ pour β_0 et β_1 et on obtient :

$$b_0 - t_{n-2;\alpha/2} \hat{\sigma}_{b_0} \leq \beta_0 \leq b_0 + t_{n-2;\alpha/2} \hat{\sigma}_{b_0}$$

et

$$b_1 - t_{n-2;\alpha/2} \hat{\sigma}_{b_1} \leq \beta_1 \leq b_1 + t_{n-2;\alpha/2} \hat{\sigma}_{b_1}$$

où $t_{n-2;\alpha/2}$ est le point critique correspondant à une loi de student à $n-2$ degrés de liberté.

Dans l'exemple 10.1, les intervalles de confiance à 95% pour β_0 et β_1 sont respectivement donnés par :

$$217,7 \pm 2,12 \times 14,46 = (187,04; 248,35)$$

et

$$0,7717 \pm 2,12 \times 0,1779 = (0,394; 1,148)$$

On conclut que l'ordonnée à l'origine ne passe pas par 0 et que la pente de la droite est non-nulle.

On peut également tester des hypothèses concernant les paramètres β_0 et β_1 . On sera presque toujours intéressé à tester l'hypothèse que $\beta_1 = 0$. En effet, si on rejette l'hypothèse que $\beta_1 = 0$ et on conclut que $\beta_1 \neq 0$, cela veut dire que la pente de la droite est significativement différente de 0 et il y a donc une relation entre les variables y et x . Dans le contexte de l'exemple 10.1, le modèle sera donc utile si l'on cherche à prédire le taux de cholestérol d'un individu au moyen de son poids. Si on ne rejette pas l'hypothèse que $\beta_1 = 0$, alors on ne peut conclure qu'il y a une dépendance entre y et x et le modèle est à toutes fins pratiques inutile car il ne permettra pas de prédire le taux de cholestérol d'un individu au moyen de son poids.

Dans certaines situations, on peut être intéressé à tester l'hypothèse $\beta_0 = 0$. Autrement dit, on peut vouloir tester si la relation passe ou pas par l'origine.

Les tests d'hypothèses pour β_1 et β_0 se font de la même façon que les tests pour une moyenne μ .

Si on cherche à tester

$$H_0 : \beta_1 = c$$

vs

$$H_1 : \beta_1 \neq c$$

où c est une constante donnée (le plus souvent $c = 0$), alors on rejettera H_0 si

$$\left| \frac{b_1 - c}{\hat{\sigma}_{b_1}} \right| > t_{n-2; \alpha/2}$$

De même, si on cherche à tester

$$H_0 : \beta_0 = c$$

vs

$$H_1 : \beta_0 \neq c$$

alors on rejettera H_0 si

$$\left| \frac{b_0 - a}{\hat{\sigma}_{b_0}} \right| > t_{n-2; \alpha/2}.$$

Dans l'exemple 10.1, les chercheurs affirment que le poids et le taux de cholestérol sont liés. On cherche donc à tester

$$H_0 : \beta_1 = 0$$

vs

$$H_1 : \beta_1 \neq 0$$

L'hypothèse nulle revient à tester que le poids et le taux de cholestérol sont deux variables indépendantes.

Après calculs, on trouve que $\frac{b_1 - c}{\hat{\sigma}_{b_1}} = 4,37$ et puisque

$t_{16;0.025} = 2,12$, on rejette H_0 et on conclut que le poids et le taux de cholestérol sont liés.

10.5 Intervalles de confiance pour $E(y)$ et limites de prédiction

En pratique, on peut vouloir estimer $E(y) \equiv \mu_y = \beta_0 + \beta_1 x$ qui représente la moyenne des y qui correspondent à une valeur donnée de x , disons x^* . On estimera naturellement μ_y par et

$$\hat{\mu}_y = b_0 + b_1 x^*$$

La variance de cet estimateur est donnée par

$$\sigma_{\hat{\mu}_y}^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

Cette variance est estimée par

$$\hat{\sigma}_{\hat{\mu}_x}^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

Un intervalle de confiance à $100(1-\alpha)\%$ pour μ_y est donné par

$$\hat{\mu}_y - t_{n-2;\alpha/2} \hat{\sigma}_{\hat{\mu}_y} \leq \mu_y \leq \hat{\mu}_y + t_{n-2;\alpha/2} \hat{\sigma}_{\hat{\mu}_y}$$

Dans l'exemple 10.1, on peut vouloir construire un intervalle de confiance pour estimer le taux de cholestérol moyen pour les individus dont le poids est de $x^* = 64$ kg. Après calculs, on obtient l'intervalle

$$(258,54;275,82)$$

Limites de prédiction

Notez bien que l'intervalle ci-dessus est un intervalle de confiance pour la *moyenne* des y qui correspondent à une valeur donnée x^* . On peut affirmer, avec $100(1-\alpha)\%$ de confiance, que cette moyenne satisfait les inégalités suivantes:

$$\hat{\mu}_y - t_{n-2;\alpha/2} \hat{\sigma}_{\hat{\mu}_y} \leq \mu_y \leq \hat{\mu}_y + t_{n-2;\alpha/2} \hat{\sigma}_{\hat{\mu}_y}$$

Mais on ne prétend pas que la probabilité est $1-\alpha$ que **le prochain y** qui correspond à x^* se situera entre ces deux bornes. Pour déterminer des bornes dans lesquelles une valeur future de y se trouvera avec une probabilité de $1 - \alpha$, nous procédons de la façon suivante. Si y_{x^*} est la future observation qui correspond à la valeur x^* , notre prédiction de y_{x^*} sera identique à notre estimation $\hat{\mu}_y$ de la moyenne au point x^* . L'écart $y_{x^*} - \mu_y$ satisfait

$$E(y_{x^*} - \mu_y) = 0$$

et

$$Var(y_{x^*} - \hat{\mu}_y) = Var(y_{x^*}) + Var(\hat{\mu}_y).$$

La variance $Var(y_{x^*}) = \sigma^2$ est estimée par $\hat{\sigma}^2$ et $Var(\hat{\mu}_y)$ est estimée par la formule donnée plus haut.

Donc

$$\begin{aligned}\hat{\sigma}_{y_{x^*}-\hat{\mu}_y}^2 &= \hat{\sigma}^2 + \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\ &= \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)\end{aligned}$$

Les limites de prédiction à $100(1-\alpha)\%$ sont

$$\hat{\mu}_y - t_{n-2;\alpha/2} \hat{\sigma}_{y_{x^*}-\hat{\mu}_y} \leq y_{x^*} \leq \hat{\mu}_y + t_{n-2;\alpha/2} \hat{\sigma}_{y_{x^*}-\hat{\mu}_y}$$

Là on peut affirmer avec $100(1-\alpha)\%$ de sécurité que la prochaine observation se situera entre les deux bornes.

Dans l'exemple 10.1, supposons que le prochain individu entrant dans le bureau d'un médecin pèse 64 kg. Alors, après calculs, on peut affirmer avec 95% de sécurité que cette observation se situera entre les bornes

$$(238,82;295,54)$$

10.6 Coefficient de corrélation

Dans cette section, nous revenons sur la notion de coefficient de corrélation r vu au chapitre 1.

La somme des carrés $\sum_{i=1}^n (y_i - \bar{y})^2$, que nous appelons « somme des carrés *totale* » et dénotons par SCT est une mesure de la dispersion totale des y , indépendamment des x . Cette somme de carrés peut être décomposée en deux parties.

La première, $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, appelée « somme des carrés expliquée » et notée *SCE*, est la partie de la dispersion des y qui est attribuable à la dispersion des x , donc « expliquée » par x .

La deuxième, $\sum_{i=1}^n (y_i - \hat{y}_i)^2$, appelée « somme des carrés résiduelle » et notée *SCR*, est la partie de la dispersion totale des y que l'on ne peut *pas* attribuer aux variations des x . Nous avons donc :

$$\begin{array}{ccccc}
 \sum_{i=1}^n (y_i - \bar{y})^2 & = & \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 & + & \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\
 \text{SCT} & = & \text{SCE} & + & \text{SCR}
 \end{array}$$

Graphiquement, SCE est la somme des carrés des distances verticales entre les points sur la droite des moindres carrés $\hat{y} = b_0 + b_1x$ et les points sur la droite horizontale $y = \bar{y}$. Cette somme de carrés a tendance à être petite si la droite des moindres carrés s'approche d'une droite horizontale, c'est-à-dire, si les données ne témoignent pas d'une forte dépendance entre y et x . SCR est la somme des distances verticales entre les points du nuage et la droite des moindres carrés. Cette somme de carrés a tendance à être petite si les points sont rapprochés de la droite des moindres carrés, cas où la dépendance entre y et x est forte.

Remarques

1. SCR et $\hat{\sigma}^2$ sont liés par la relation suivante:

$$\hat{\sigma}^2 = SCR/(n-2)$$

Donc SCR petit signifie que les y_i ont tendance à être *peu* dispersés par rapport à leur moyenne

$$\beta_0 + \beta_1 x_i,$$

ce qui se manifeste dans l'échantillon par un nuage de points rapproché de la droite des moindres carrés. Nous avons aussi la relation suivante entre $\hat{\beta}_1$ et SCR :

$$SCR = \sum_{i=1}^n (y_i - \bar{y})^2 - b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2.$$

2. SCE et b_1 sont liés par la relation suivante :

$$SCE = b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2.$$

Donc SCE petit signifie que $|b_1|$ est petit, et par conséquent que la droite est près d'être horizontale.

Relation entre b_1 et le coefficient de corrélation

Le coefficient de corrélation

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

satisfait toujours $|r| \leq 1$ et $|r| = 1$ si et seulement si $y_i = a + bx_i$ pour un certain a pour $i = 1, \dots, n$.

Donc les valeurs $r = 1$ et $r = -1$ dénotent une corrélation linéaire parfaite entre les x_i et les y_i .

En comparant les expressions de b_1 et de r , on constate que r et b_1 sont de même signe et que $r = 0 \iff b_1 = 0$. Nous avons la relation suivante :

$$r = b_1 \frac{S_x}{S_y}, \text{ et donc } b_1 = r \frac{S_y}{S_x}$$

Donc $r > 0$ si et seulement si la droite des moindres carrés est de pente positive, et $r = 0$ si et seulement si la droite des moindres carrés est horizontale. Pour interpréter les valeurs intermédiaires de r , nous avons l'égalité suivante :

$$r^2 = SCE/SCT$$

Donc r^2 est la *proportion de la dispersion des y qui est expliquée par la dispersion des x* .

Dans l'exemple 10.1, on trouve après calculs : $r^2 = 0,544$, ce qui indique que la relation entre le poids et le taux de cholestérol est relativement forte.

Revenons à l'exemple 10.1 et demandons à un logiciel statistique (par exemple, MINITAB) d'effectuer l'analyse de régression avec comme variable dépendante le taux de cholestérol et comme variable indépendante, le poids.

La plupart des logiciels fourniront alors un tableau très similaire à celui exhibé ci-dessous :

Regression Analysis: Taux de cholestérol versus Poids

The regression equation is

$$\text{Taux de cholestérol} = 217 + 0,777 \text{ Poids}$$

Predictor	Coef	SE Coef	T	P
Constant	217,47	14,46	15,04	0,000
Poids	0,7767	0,1779	4,37	0,000

$$S = 12.7423 \quad R\text{-Sq} = 54,4\% \quad R\text{-Sq}(\text{adj}) = 51,5\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3095,3	3095,3	19,06	0,000
Residual Error	16	2597,9	162,4		
Total	17	5693,1			

Comment interpréter tous ces nombres?

Annexe A : Aire pour la loi normale

Note : Un nombre dans la table correspond à l'aire sous la courbe à gauche. *Exemple* : Si $X \sim N(0,1)$, $P(X \leq 1,25) = 0,8944$.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,60	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,70	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,80	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,90	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,00	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,10	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,20	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,30	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,40	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,50	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,60	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,70	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,80	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,90	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4,00	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Annexe B : Table de la loi khi-deux

Chaque case donne la valeur x pour laquelle $P(X \leq x) = p$ lorsque X est de loi χ^2 à v degrés de liberté χ_v^2 .

Exemple : Si $X \sim \chi_5^2$, $P(X \leq 1,610) = 0,1$.

v	P							
	0,010	0,025	0,05	0,1	0,9	0,95	0,975	0,99
1	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635
2	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210
3	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345
4	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277
5	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086
6	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812
7	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475
8	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090
9	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666
10	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209
11	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725
12	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217
13	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688
14	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141
15	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578
16	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000
17	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409
18	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805
19	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191
20	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566
21	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932
22	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289
23	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638
24	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980
25	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314
26	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642
27	12,879	14,573	16,151	18,114	36,741	40,113	43,195	46,963
28	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278
29	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588
30	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892

Annexe C : Table de la loi de Student

Chaque case donne la valeur x pour laquelle $P(X \geq x) = p$ lorsque X est de loi de Student à v degrés de liberté : t_v
 Exemple : Si $X \sim t_{14}$, $P(X \geq 1,07628) = 0,15$.

v	p : Surface à droite							
	0,25	0,2	0,15	0,10	0,05	0,025	0,01	0,005
1	1,00000	1,37638	1,96261	3,07768	6,31375	12,70620	31,82052	63,65674
2	0,81650	1,06066	1,38621	1,88562	2,91999	4,30265	6,96456	9,92484
3	0,76489	0,97847	1,24978	1,63774	2,35336	3,18245	4,54070	5,84091
4	0,74070	0,94096	1,18957	1,53321	2,13185	2,77645	3,74695	4,60409
5	0,72669	0,91954	1,15577	1,47588	2,01505	2,57058	3,36493	4,03214
6	0,71756	0,90570	1,13416	1,43976	1,94318	2,44691	3,14267	3,70743
7	0,71114	0,89603	1,11916	1,41492	1,89458	2,36462	2,99795	3,49948
8	0,70639	0,88889	1,10815	1,39682	1,85955	2,30600	2,89646	3,35539
9	0,70272	0,88340	1,09972	1,38303	1,83311	2,26216	2,82144	3,24984
10	0,69981	0,87906	1,09306	1,37218	1,81246	2,22814	2,76377	3,16927
11	0,69745	0,87553	1,08767	1,36343	1,79588	2,20099	2,71808	3,10581
12	0,69548	0,87261	1,08321	1,35622	1,78229	2,17881	2,68100	3,05454
13	0,69383	0,87015	1,07947	1,35017	1,77093	2,16037	2,65031	3,01228
14	0,69242	0,86805	1,07628	1,34503	1,76131	2,14479	2,62449	2,97684
15	0,69120	0,86624	1,07353	1,34061	1,75305	2,13145	2,60248	2,94671
16	0,69013	0,86467	1,07114	1,33676	1,74588	2,11991	2,58349	2,92078
17	0,68920	0,86328	1,06903	1,33338	1,73961	2,10982	2,56693	2,89823
18	0,68836	0,86205	1,06717	1,33039	1,73406	2,10092	2,55238	2,87844
19	0,68762	0,86095	1,06551	1,32773	1,72913	2,09302	2,53948	2,86093
20	0,68695	0,85996	1,06402	1,32534	1,72472	2,08596	2,52798	2,84534
21	0,68635	0,85907	1,06267	1,32319	1,72074	2,07961	2,51765	2,83136
22	0,68581	0,85827	1,06145	1,32124	1,71714	2,07387	2,50832	2,81876
23	0,68531	0,85753	1,06034	1,31946	1,71387	2,06866	2,49987	2,80734
24	0,68485	0,85686	1,05932	1,31784	1,71088	2,06390	2,49216	2,79694
25	0,68443	0,85624	1,05838	1,31635	1,70814	2,05954	2,48511	2,78744
26	0,68404	0,85567	1,05752	1,31497	1,70562	2,05553	2,47863	2,77871
27	0,68368	0,85514	1,05673	1,31370	1,70329	2,05183	2,47266	2,77068
28	0,68335	0,85465	1,05599	1,31253	1,70113	2,04841	2,46714	2,76326
29	0,68304	0,85419	1,05530	1,31143	1,69913	2,04523	2,46202	2,75639
30	0,68276	0,85377	1,05466	1,31042	1,69726	2,04227	2,45726	2,75000
35	0,68156	0,85201	1,05202	1,30621	1,68957	2,03011	2,43772	2,72381
40	0,68067	0,85070	1,05005	1,30308	1,68385	2,02108	2,42326	2,70446
50	0,67943	0,84887	1,04729	1,29871	1,67591	2,00856	2,40327	2,67779
60	0,67860	0,84765	1,04547	1,29582	1,67065	2,00030	2,39012	2,66028
70	0,67801	0,84679	1,04417	1,29376	1,66691	1,99444	2,38081	2,64790
80	0,67757	0,84614	1,04320	1,29222	1,66412	1,99006	2,37387	2,63869
90	0,67723	0,84563	1,04244	1,29103	1,66196	1,98667	2,36850	2,63157
100	0,67695	0,84523	1,04184	1,29007	1,66023	1,98397	2,36422	2,62589

Annexe D : Rappels sur les ensembles

Concepts de base

Définition 1 : Un ensemble A est une collection d'objets. Les objets sont appelés éléments.

Notation : On écrit $p \in A$ (et on lit : " p appartient à A ") si p est un élément de A .

Exemple : $\mathbf{N} = \{0,1,2,3,4,\dots\}$ est l'ensemble des entiers naturels.

On a $1 \in \mathbf{N}$, $4 \in \mathbf{N}$ mais $-2 \notin \mathbf{N}$.

Définition 2 : Soient A et B deux ensembles. Si chaque élément de A est également un élément de B , alors A est un sous-ensemble de B . On écrit $A \subset B$ (et on lit « A est inclus dans B »).

Exemple : Let $B = \mathbf{N} = \{0,1,2,3,\dots\}$ and $A = \{0,2,4,6,8,\dots\}$. Il est clair que $A \subset B$.

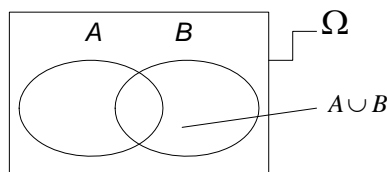
Remarques :

- 1) Si $A \subset B$ et $B \subset A$, alors $A = B$.
- 2) La négation de $p \in A$, $A \subset B$, $A = B$ est $p \notin A$, $A \not\subset B$, $A \neq B$ respectivement.

Définition 3 : Dans le contexte de la théorie des probabilités, tous les ensembles considérés sont des sous-ensembles d'un ensemble universel que l'on désigne par Ω . L'ensemble vide qui est l'ensemble ne contenant aucun élément est désigné par \emptyset .

Opérations dans les ensembles :

- (a) **Union :** Soient A et B deux ensembles. L'union de A et B est l'ensemble des éléments qui appartiennent à A ou à B . On désigne l'union de A et B par $A \cup B$. On dira que $x \in A \cup B$ si x appartient à au moins l'un des deux ensembles A et B .

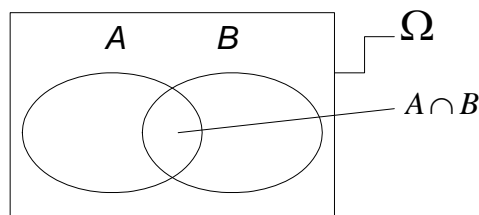


Remarque : La figure ci-dessus est appelée diagramme de Venn.

Exemples :

- 1) Soient $A = \{1,2,3,4\}$ et $B = \{2,4,5,6\}$. Alors $A \cup B = \{1,2,3,4,5,6\}$.
- 2) Soit A l'ensemble des individus aux cheveux blonds; soit B l'ensemble des individus aux cheveux bruns. Alors $A \cup B$ est l'ensemble des individus qui ont les cheveux blonds ou les cheveux bruns.

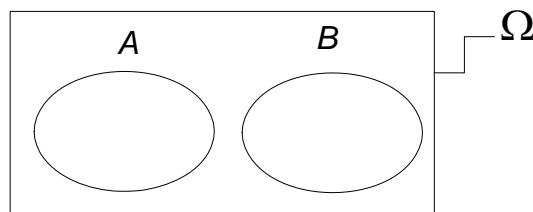
(b) **Intersection :** Soient A and B deux ensembles. L'intersection de A et B est l'ensemble des éléments qui appartiennent à A et à B . On désigne l'intersection de A et B par $A \cap B$.



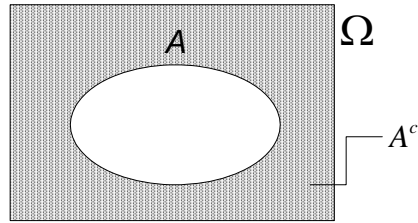
Exemples :

- 1) Soit $A = \{1,2,3,4\}$ et $B = \{2,4,5,6\}$. Alors, $A \cap B = \{2,4\}$.
- 2) Soit $A = \{1,3,5,7,\dots\}$ et $B = \{0,2,4,6,8,\dots\}$. Alors $A \cap B = \emptyset$.

Remarque : Si $A \cap B = \emptyset$, on dit que A et B sont *mutuellement exclusifs* ou *disjoints* ou *incompatibles*. Dans ce cas, le diagramme de Venn est le suivant :



(c) **Complément:** Le complément d'un ensemble A est l'ensemble des éléments qui n'appartiennent pas à A . Le complément de A est désigné par A^c .



Exemple: Si A est l'ensemble des fumeurs au Canada, A^c est l'ensemble des non-fumeurs.

Remarque: A et A^c sont toujours disjoints; i.e. $A \cap A^c = \emptyset$.

Règles de base

Union

$$\begin{aligned} A \cup A &= A \\ A \cup B &= B \cup A \\ A \cup \emptyset &= A \\ A \cup \Omega &= \Omega \\ A \cup A^c &= \Omega \end{aligned}$$

$$(A^c)^c = A$$

$$(A \cup B)^c = A^c \cap B^c$$

Intersection

$$\begin{aligned} A \cap A &= A \\ A \cap B &= B \cap A \\ A \cap \emptyset &= \emptyset \\ A \cap \Omega &= A \\ A \cap A^c &= \emptyset \end{aligned}$$

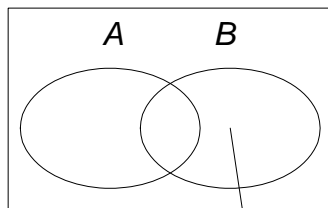
$$\Omega^c = \emptyset, \emptyset^c = \Omega$$

$$(A \cap B)^c = A^c \cup B^c$$

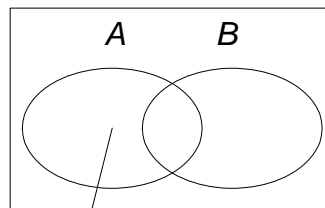
Lois de DeMorgan

Autres ensemble intéressants:

- $A \cap B^c$: $x \in (A \cap B^c)$ si et seulement si $x \in A$ et $x \notin B$.
- $A^c \cap B$: $x \in (A^c \cap B)$ si et seulement si $x \notin A$ et $x \in B$.



$$A^c \cap B$$



$$A \cap B^c$$

