

Statistique inférentielle

- *Statistiques III* -

Dr Sacha Varone

Édition 2009

h e g

Haute école de gestion de Genève
Geneva School of Business Administration

Bibliographie

- [1] David F. Groebner, Patrick W. Shannon, Philip C. Fry, and Kent D. Smith. *Business Statistics :A Decision-Making Approach*. Prentice Hall, 2005.
- [2] Thomas H. Wonnacott and Ronald J. Wonnacott. *Statistique*. Economica, Paris, 4th edition, 1991.

Table des matières

1	Introduction	1
2	Distributions continues	3
2.1	Loi normale	5
2.2	Loi de Student (\mathcal{T}_n)	9
2.3	Loi du chi carré (χ^2)	11
3	Estimation	13
3.1	Estimation ponctuelle	13
3.2	Théorème central limite	17
3.3	Intervalle de confiance	22
3.4	IC pour estimer μ, σ^2 connu	24
3.5	IC pour estimer μ, σ^2 inconnu	26
3.6	IC pour estimer une proportion	28
3.7	IC pour estimer σ^2, μ inconnu	30
3.8	Résumé	31
4	Tests paramétriques	33
4.1	Principe des tests d'hypothèses	33
4.2	Test de μ, σ^2 connu, grand échantillon	39
4.3	Test de μ, σ^2 inconnu, grand échantillon	40
4.4	Test de μ, σ^2 inconnu, petit échantillon	40
4.5	Value at Risk	41

4.6	Test de la proportion	42
4.7	Test de la variance avec moyenne inconnue	43
4.8	Méthode de la p -valeur	45
4.9	Relation entre IC et tests statistiques	48
4.10	Résumé	50
5	Tests non paramétriques	51
5.1	Test des rangs signés de Wilcoxon	51
5.2	Test d'ajustement	54
5.3	Test d'indépendance de deux variables catégorielles	55
6	Régression linéaire simple	57
6.1	Rappel	57
6.2	Validité d'une corrélation	58
6.3	Qualité d'un modèle linéaire	59
6.4	Validité d'un modèle linéaire	61
6.5	Analyse d'une régression linéaire simple	62
A	Tables statistiques	63
A.1	Loi normale centrée réduite	64
A.2	Table de la loi du χ^2	65
A.3	Table de la loi de Student	66
A.4	Table du test des rangs signés de Wilcoxon	67
B	Instructions pour logiciels	68
B.1	Loi normale	68
B.2	t -distribution	69
B.3	χ_n^2 distribution	69
B.4	Intervalle de confiance	69
B.5	Test d'une moyenne	70

B.6	Test d'une proportion	71
B.7	Test d'une variance	71
B.8	Test des rangs signés de Wilcoxon	71
B.9	Test d'indépendance	72
B.10	Test de corrélation linéaire	72
B.11	Régression linéaire	72

Remerciements

Je tiens à remercier le prof. André Berchtold, qui m'a fourni son support de cours, sur lequel une partie de ce cours a été construit.

Chapitre 1

Introduction

Les observations statistiques peuvent être classées en fonction de leur niveau, du plus particulier au plus général :

Définition 1.1 *Une unité statistique est le plus petit élément sur lequel porte l'analyse statistique.*

Définition 1.2 *Une variable statistique est une caractéristique d'une unité statistique.*

Exemple:

Unité statistique : un(e) étudiant(e) de la HEG.

Variables : la couleur des yeux, la taille, le poids, le sexe,

Définition 1.3 *Une population est un ensemble de toutes les unités statistiques sur lequel porte une étude statistique.*

Définition 1.4 *Un échantillon est un sous-ensemble de la population.*

Les données récoltées sont généralement issues d'un échantillon d'*individus* d'une population d'intérêt.

Exemple:

Population : ensemble des étudiants de la HEG.

Échantillon : ensemble des étudiants de première année de la HEG.

Individu : un(e) étudiant(e) de première année de la HEG.

On veut étudier une ou plusieurs caractéristiques que possède chaque individu d'une population.

Remarque: Lorsque les données proviennent d'observations collectées au même moment (ou presque), on parle de *données en coupe transversale*.

Statistique descriptive

En statistique descriptive, l'objectif est de décrire et comprendre les données à disposition. Le niveau auquel elles appartiennent n'a pas alors beaucoup d'importance. Lorsque l'on travaille sur l'ensemble des données d'une population, c'est parfois la seule phase statistique. L'analyse exploratoire suggère des hypothèses de travail et des modèles qui peuvent être formalisés et vérifiés en statistique inférentielle.

Statistique inférentielle

En statistique inférentielle au contraire, on travaille toujours à partir d'un échantillon de données issu d'une population que nous ne pouvons pas interroger ou examiner dans son ensemble, mais pour laquelle on aimerait disposer de résultats fiables. Elle conduit à des conclusions statistiques à partir de données en utilisant des notions de la théorie des probabilités. Cette partie s'occupe des méthodes de test et d'estimation des paramètres.

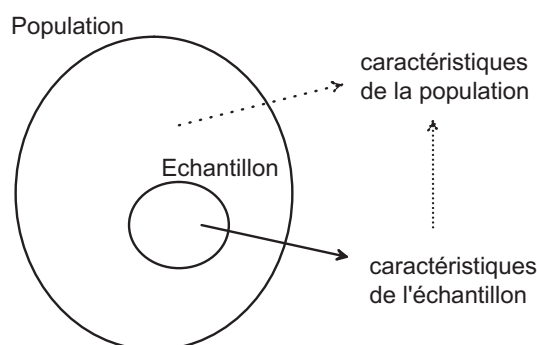


FIGURE 1.1 – Principe de l'inférence statistique

Définition 1.5 *Un paramètre est une mesure calculée à partir d'une population entière.*

Définition 1.6 *Une statistique est une mesure calculée à partir d'un échantillon.*

Aussi longtemps que la population ne change pas, la valeur d'un paramètre associé à cette population ne varie pas. Par contre, la valeur d'une statistique dépend de l'échantillon sélectionné. Il existe donc généralement une différence entre la valeur d'un paramètre, et son estimation par une statistique.

Définition 1.7 *L'erreur d'échantillonnage est la valeur de la différence entre une statistique et le paramètre évalué.*

Chapitre 2

Distributions continues

Les données peuvent provenir d'événements ponctuels comme le nombre de personnes entrant dans une banque en 1 heure, ou le nombre d'appels téléphonique à faire afin de trouver 10 acheteurs d'un nouveau produit. Mais souvent, les données sont continues, comme la quantité de panure sur les mets à base de poisson, la longueur de coupes d'aciers, le temps de recharge d'une batterie de téléphone,

Afin de modéliser ces phénomènes de type continu, nous utilisons des distributions continues, comme

- la loi uniforme
- la loi normale
- la loi exponentielle
-

Des *tables statistiques* sont disponibles pour chacune de ces lois, ainsi que des fonctions associées dans les logiciels de bureautique (MS Excel, Calc d'Openoffice, Gnumeric, ...) ou spécialisés (R).

Variable aléatoire continue

Une variable aléatoire continue X prend ses valeurs dans un intervalle qui est un sous-ensemble de l'ensemble des réels \mathbb{R} :

$$X \in [u, v]$$

avec $u, v \in \mathbb{R}, u < v$

Le nombre de valeurs possibles de X étant infini, chacune de ces valeurs a une probabilité nulle. En revanche, il est possible de calculer la probabilité associée à n'importe quel sous-intervalle $[a, b]$ de $[v, w]$:

$$\begin{aligned} P(X = x) &= 0 & \forall x \in [v, w] \\ P(X \in [a, b]) &\geq 0 & \forall a < b, [a, b] \subseteq [v, w] \end{aligned}$$

Fonction de densité

Une distribution continue est définie soit par sa fonction de densité, soit par sa fonction de répartition.

Une distribution continue peut être représentée par un histogramme. Si l'on dispose d'un grand nombre d'observations, il est possible de définir des classes de très petites amplitudes. A la limite, les classes deviennent d'amplitude nulle et l'histogramme devient une courbe. Cette courbe est la fonction de densité de la variable aléatoire continue X , notée $f(x)$.

La fonction de densité n'est pas une distribution de probabilité. En effet, pour tout x appartenant à l'intervalle $[v, w]$, $P(X = x) = 0$, mais $f(x) \geq 0$.

En revanche, l'aire sous la courbe d'une fonction de densité doit valoir 1, qui est la somme de toutes les probabilités. Ainsi, la relation suivante est vérifiée pour toute fonction de densité

$$P(X \in]-\infty, \infty[) = 1$$

Fonction de répartition

La probabilité d'être dans un intervalle de longueur dx très petit est donnée par $f(x)dx$:

$$P(X \in [x, x + dx]) = f(x)dx$$

La probabilité de se trouver dans un intervalle $[a, b]$ est définie comme l'aire sous la fonction de densité. Mathématiquement, cela revient à calculer l'intégrale de la fonction $f(x)$ entre a et b :

$$P(X \in [a, b]) = \int_a^b f(x)dx$$

Définition 2.1 La fonction de répartition, notée $F(a)$, exprime pour une variable aléatoire X la probabilité d'être inférieure ou égale à x . Cela correspond à la surface sous la fonction de densité à gauche de x :

$$F(a) = P(X \leq a) = \int_{-\infty}^a f(x)dx$$

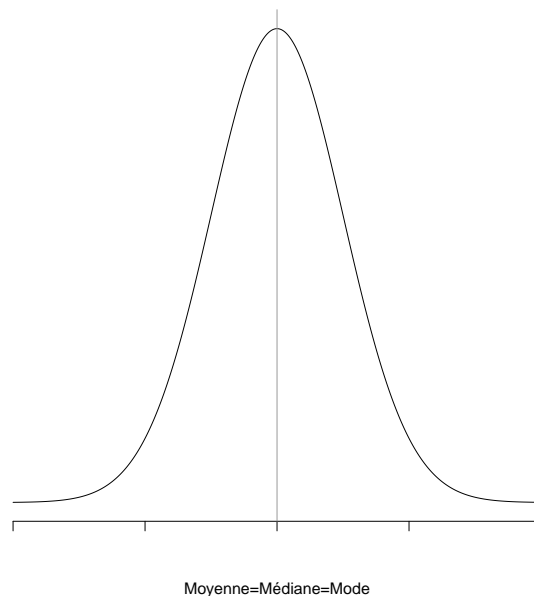
Propriété 2.1 La fonction de répartition est telle que

- $F(-\infty) = 0$,
- $F(\infty) = 1$
- $P(X < a) = P(X \leq a) = F(a)$
- $P(X \in [a, b]) = F(b) - F(a)$
- Par complémentarité, $P(X > a) = 1 - P(X \leq a) = 1 - F(a)$

2.1 Loi normale

La loi la plus utilisée pour décrire des phénomènes par une variable aléatoire continue est la loi normale. Par exemple, la description du poids, de la taille, du remplissage de récipients, ... Cette loi est abondamment utilisée lors de l'inférence statistique.

La représentation graphique de la loi normale est une courbe en forme de cloche, symétrique.



Définition 2.2 La densité de probabilité normale s'exprime par

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

où μ est la moyenne et σ est l'écart type

La loi normale, notée $\mathcal{N}(\mu, \sigma^2)$, comporte deux paramètres μ et σ^2 , qui déterminent la position et la forme de la distribution. Il existe donc une famille de lois normales, et non pas une seule, qui se différencient par leur moyenne et leur écart type.

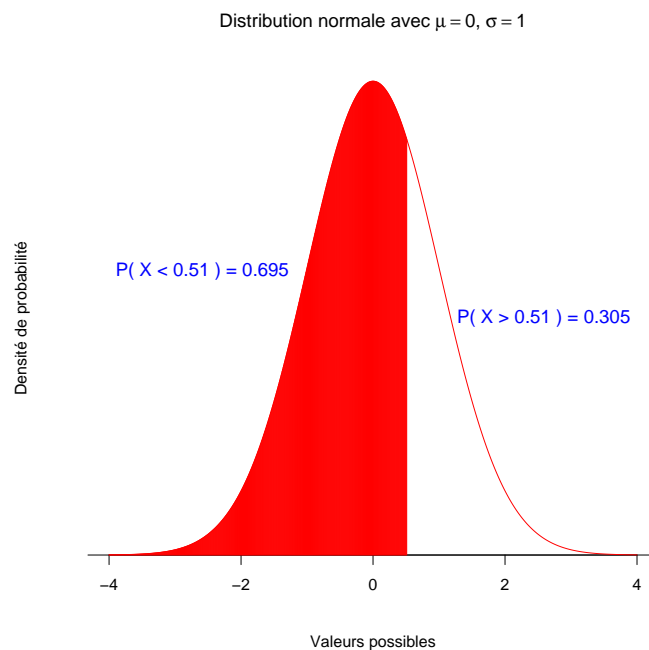
Propriété 2.2 – Le point le plus élevé de la courbe normale correspond à la moyenne, qui est aussi la médiane et le mode de la distribution.

- La distribution normale étant symétrique, son coefficient d'asymétrie (skewness) est nul.
- L'écart type détermine la largeur de la courbe. Plus sa valeur est élevée, plus la courbe sera large et aplatie.
- La variable aléatoire associée peut prendre n'importe quelle valeur réelle dans $]-\infty, \infty[$.

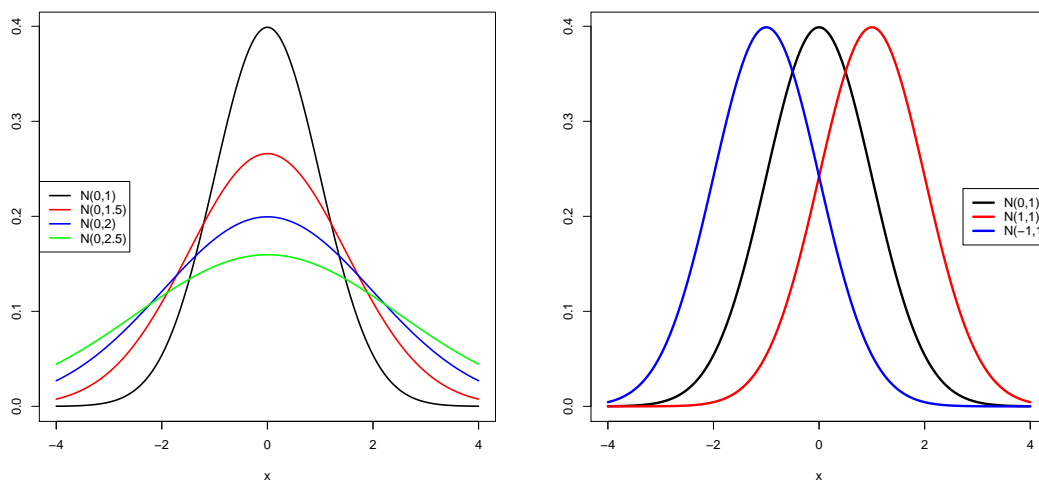
2.1. LOI NORMALE

La probabilité est représentée par l'aire sous la courbe de densité $f(x)$. L'aire totale vaut 1 car il s'agit d'une probabilité. Comme la loi normale est symétrique, la probabilité $P(X \leq \mu) = P(X \geq \mu) = 0.5$, et par conséquent

$$P(X \leq x) = 1 - P(X \geq x)$$



En variant les paramètres μ et σ^2 de la loi normale, nous obtenons différentes lois normales. Ainsi, il n'existe pas une seule loi normale, mais une multitudes de loi normales qui se différencient par leurs paramètres.



Il existe une distribution normale de référence, de moyenne 0 et de variance 1, dite aussi distribution Z .

Définition 2.3 Une loi normale de moyenne nulle et d'écart type 1 est dite loi normale centrée réduite, aussi dite loi normale standard. La fonction de densité est alors

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

La lettre Z est habituellement utilisée pour désigner la variable aléatoire continue dont la loi est normale centrée réduite. Il est toujours possible de transformer une variable aléatoire X distribuée selon une loi normale $\mathcal{N}(\mu, \sigma^2)$, en une variable Z distribuée selon une loi normale centrée réduite $\mathcal{N}(0, 1)$. La transformation consiste à soustraire de X la moyenne μ , puis de diviser par l'écart type σ

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Les probabilités suivantes sont alors équivalentes, avec $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned} X \leq x &\Rightarrow Z \leq \frac{x - \mu}{\sigma} \\ P(X \leq x) &= P(Z \leq \frac{x - \mu}{\sigma}) \end{aligned}$$

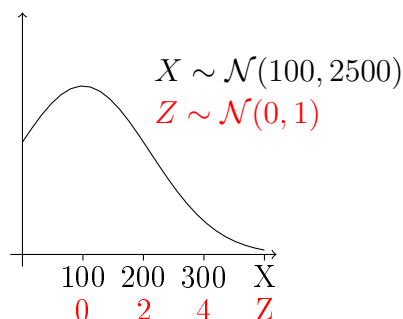
Exemple:

Si X est une variable aléatoire suivant une loi normale de centre $\mu = 100$ et d'écart type $\sigma = 50$, i.e.

$$X \sim \mathcal{N}(100, 2500)$$

alors la z valeur associée est

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



Ainsi, si $X = 200$, alors $Z = \frac{200-100}{50} = 2$. Cela signifie que $X = 200$ est éloigné de 2 écarts type au delà de la moyenne.

Nous avons alors $100 + 2 * 50 = 200 = X$

Table de la loi normale La table de la loi normale donne les probabilités d'occurrence jusqu'à la z-valeur considérée. La ligne donne la valeur de z jusqu'au dixième, et la colonne donne la valeur de z au centième.

Exemple:

- La z-valeur associée à la ligne 0.5 et à la colonne 0.01 correspond à $Z=0.51$.
L'intersection de la ligne et de la colonne donne la probabilité cherchée :
 $P(Z \leq 0.51) = 0.6950$
- $P(Z \geq 0.51) = 1 - 0.6950 = 0.305$ par complémentarité.
- $P(Z \leq -0.51) = P(Z \geq 0.51) = 0.305$ par symétrie.

Règle empirique En pratique, de nombreux ensembles de données ont une distribution en forme de cloche. Dans ce cas, on peut utiliser une règle empirique, fondée sur une distribution de probabilité normale :

- Environ 68% des observations se situent dans l'intervalle $[\bar{x} - s; \bar{x} + s]$
- Environ 95% des observations se situent dans l'intervalle $[\bar{x} - 2s; \bar{x} + 2s]$
- Environ 99.7% des observations (presque toutes) se situent dans l'intervalle $[\bar{x} - 3s; \bar{x} + 3s]$

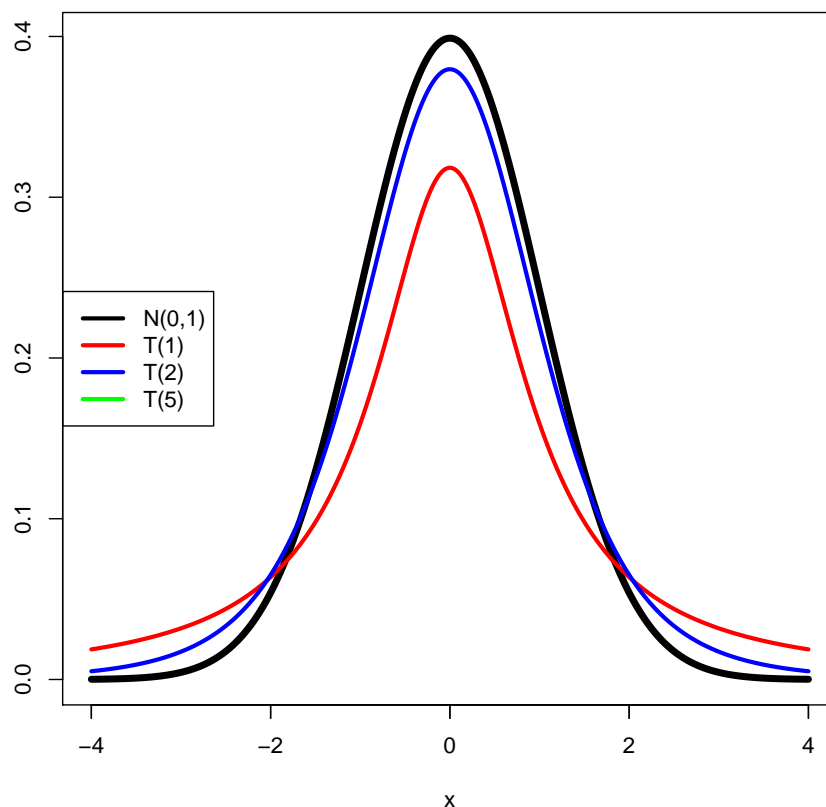
Les variables centrées réduites Z permettent de détecter des valeurs singulières de la manière suivante : si les données semblent être le résultat d'une variable aléatoire suivant une loi normale, presque toutes les observations devraient se trouver entre la moyenne et ± 3 écarts type. Il est alors recommandé de considérer toute observation dont la valeur de Z se situe en dehors de cet intervalle comme étant singulière.

2.2 Loi de Student (\mathcal{T}_n)

William Gosset, alors qu'il travaillait pour la brasserie Guinness à Dublin, avait l'interdiction de publier le résultat de ses travaux. Après avoir suivi un cours de Statistiques donné par Karl Pearson, il prit le pseudonyme de "Student" pour publier la distribution du même nom en 1908.

2.2.1 Loi de Student à n degrés de liberté

La distribution de Student, aussi appelée t -distribution, est une famille de distribution en forme de cloche et symétrique. Elle est similaire à la distribution normale, mais avec des valeurs plus grandes aux extrémités (x petit ou x grand). Elle se caractérise par le nombre de degrés de liberté.



Lorsque le nombre de degrés de liberté n tend vers l'infini, la loi de Student tend vers la loi normale $\mathcal{N}(0, 1)$.

Propriété 2.3 \mathcal{T}_n : loi de Student à n degrés de liberté

- $E(\mathcal{T}_n) = 0$, $n > 1$
L'espérance tend vers l'infini lorsque $n = 1$. On dit alors qu'elle n'existe pas.
- $\text{Var}(\mathcal{T}_n) = \frac{n}{n-2}$, $n > 2$
La loi de Student a une variance infinie pour $n \leq 2$
- *La loi de Student est symétrique autour de 0.*

On utilisera le théorème suivant qui permet de faire de l'inférence sur la moyenne d'une loi normale de moyenne μ inconnue et de variance σ^2 inconnue. :

Théorème 2.4 Soit un échantillon aléatoire de taille n , de moyenne \bar{x} et de variance s^2 , issu d'une loi normale $\mathcal{N}(\mu, \sigma^2)$. Alors

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim \mathcal{T}_{n-1}$$

2.2.2 Table de Student

La table de la loi de Student se lit pour l'essentiel comme la table de la loi du χ^2 . Chaque ligne correspond à un nombre de degrés de liberté n et chaque colonne à une erreur de première espèce α . L'intersection d'une ligne et d'une colonne donne le seuil $t_{\alpha,n}$ tel que

$$P(\mathcal{T}_n \geq t_{\alpha,n}) = \alpha$$

La relation entre p et α est $p = 1 - \alpha$.

Exemple:

$$P(\mathcal{T}_{10} \leq t_{\alpha,10}) = 0.95 \implies t_{0.05,10} = 1.8125$$

Étant donné que la loi de Student est symétrique autour de 0, seules les erreurs de première espèce α inférieures à 0.5 sont données. Les $\alpha > 0.5$ peuvent être retrouvées, comme pour la loi normale, par symétrie et complémentarité.

2.3 Loi du chi carré (χ^2)

2.3.1 Loi du chi-2 à n degrés de liberté

Définition 2.4 Soit n variables aléatoires normales centrées-réduites Z_i , indépendantes les unes des autres et identiquement distribuées : Z_i i.i.d. $\sim \mathcal{N}(0, 1)$, $i = 1, 2, \dots, n$. Alors la variable formée de la somme des carrés de ces variables

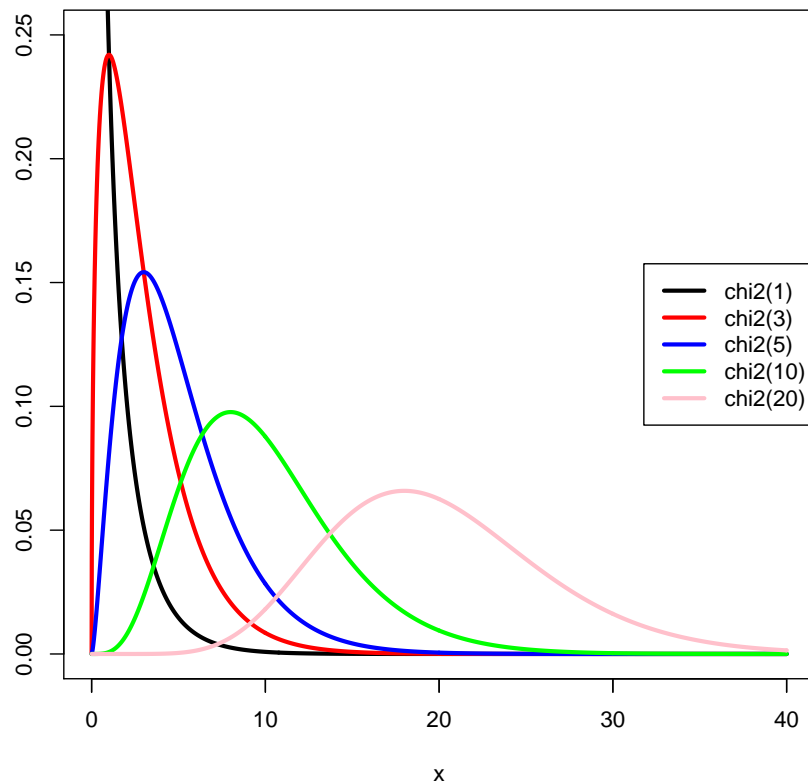
$$Q_n = \sum_{i=1}^n Z_i^2 \sim \chi^2$$

suit une loi du χ^2 à n degrés de liberté.

On note parfois $\chi^2(n)$ ou χ_n^2 pour indiquer une loi du χ^2 à n degrés de liberté

Propriété 2.5 La loi du χ^2 à n degrés de liberté possède les propriétés suivantes

- Son espérance vaut n
- Sa variance vaut $2n$



2.3. LOI DU CHI CARRÉ (χ^2)

Remarque: La loi du chi-2 étant définie comme une somme de carrés, elle ne prend que des valeurs positives.

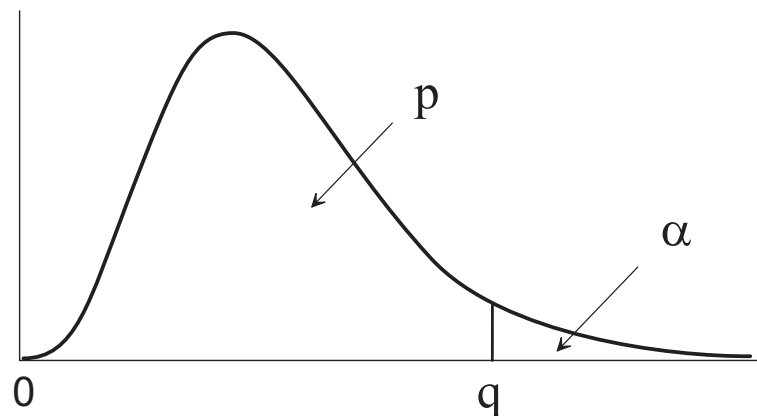
Nous utiliserons la propriété suivante lorsque nous calculerons (cf. 3.7) un intervalle de confiance pour la variance d'une population.

Propriété 2.6 La statistique χ^2 à $n - 1$ degrés de liberté vaut

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

où
 χ^2 = variable chi-2 standard
 s^2 = variance de l'échantillon
 σ^2 = variance de la population
 n = taille de l'échantillon

2.3.2 Table du χ^2



Pour une variable Q_n suivant une loi du χ^2 à n degrés de liberté, $p = P(Q_n \leq q_{\alpha,n})$ et $\alpha = P(Q_n > q_{\alpha,n})$. La relation entre p et α est donnée par

$$p = 1 - \alpha$$

La table du χ^2 en annexe A.2 donne le seuil q associé à une erreur de première espèce α (colonne) et un nombre de degrés de liberté n (ligne).

Exemple:

$$P(Q_7 \leq q_{\alpha,7}) = 0.95$$

$$\text{Alors } \alpha = 1 - 0.95 = 0.05$$

$$\text{Et donc } q_{0.05,7} = 14.0671$$

Chapitre 3

Estimation

L'estimation a pour objectif d'attribuer une valeur à un ou plusieurs paramètres de la population sur la base d'un échantillon issu de celle-ci. L'estimation peut être ponctuelle ou par intervalle.

3.1 Estimation ponctuelle

Définition 3.1 Une estimation ponctuelle, ou point d'estimation, est une valeur calculée à partir d'un échantillon pour estimer un paramètre d'une population.

Exemple:

Soit une variable aléatoire représentant la taille des ménages (nombre de personnes composant un ménage) et soit deux échantillons aléatoires X et Y :

échantillon X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
	1	3	2	4	4	1	2	6

échantillon Y	y_1	y_2	y_3	y_4	y_5
	2	2	1	1	3

Échantillon X : $\bar{x} = \frac{23}{8}$

Échantillon Y : $\bar{y} = \frac{9}{5}$

Les valeurs \bar{x} et \bar{y} sont des estimations ponctuelles de la taille moyenne des ménage (=paramètre de la population).

3.1.1 Propriétés des estimateurs

Chaque échantillon conduit à une estimation différente du paramètre. Il y a bien sûr des ensembles de valeurs qui sont plus adéquates que d'autres pour un estimateur.

Définition 3.2 Une distribution d'échantillonnage est la distribution des valeurs possibles d'une statistique pour un échantillon de taille fixée, sélectionné à partir d'une population.

Ainsi, un estimateur $\hat{\Theta}$ est considéré comme une variable aléatoire dont le comportement en moyenne est

$$E(\hat{\Theta})$$

et dont la variance est une mesure de dispersion des estimations

$$\text{Var}(\hat{\Theta})$$

Dans les situations où l'espérance de l'estimateur est égal au paramètre, on dit que l'estimateur est sans biais. Cependant, quel que soit le sondage (qui n'est pas un recensement), la différence entre le point $\hat{\theta}$ calculé et le paramètre θ de la population n'est pas connu. On attend d'un estimateur qu'il soit fiable, et aussi proche que possible de la véritable valeur du paramètre. Pour cela, deux propriétés sont souhaitables pour un estimateur $\hat{\Theta}$:

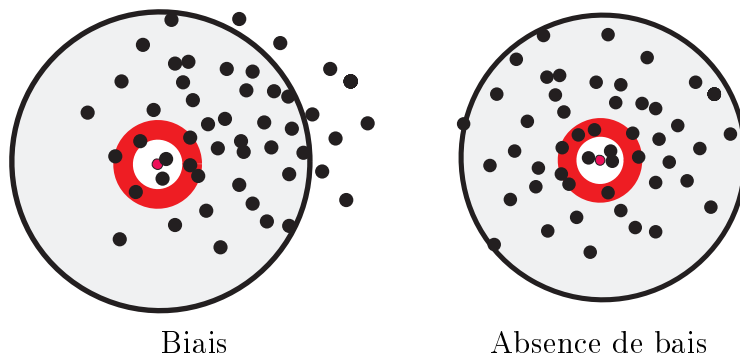
- l'absence de biais
- la convergence

Définition 3.3 Le biais est la quantité

$$E(\hat{\Theta} - \theta) = E(\hat{\Theta}) - \theta$$

Un estimateur est dit non-biaisé lorsque son espérance est égale à la vraie valeur du paramètre estimé :

$$E(\hat{\Theta}) = \theta$$



Définition 3.4 Un estimateur est dit convergent si, lorsque la taille n de l'échantillon devient grande

1. le biais disparaît :

$$\lim_{n \rightarrow \infty} \text{Biais}(\hat{\Theta}) = 0$$

2. la variance devient nulle

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\Theta}) = 0$$

Cela revient à dire que lorsque la taille de l'échantillon augmente, l'estimation devient de plus en plus précise.

Définition 3.5 *Un estimateur sans biais et convergent est dit absolument correct.*

3.1.2 Estimation de la moyenne de la population

Considérons un échantillon de taille n .

La moyenne de la population, μ , est estimée à l'aide de la moyenne de l'échantillon, notée \bar{x} :

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

Propriété 3.1 *\bar{x} est un estimateur absolument correct de la moyenne μ .*

Remarque: La moyenne tronquée, le mode et la médiane calculés à partir d'un échantillon sont aussi des estimateurs absolument corrects pour respectivement la moyenne tronquée, le mode et la médiane de la population.

3.1.3 Estimation de la variance de la population

Considérons un échantillon de taille n .

La variance de la population, σ^2 , peut être estimée à partir de la variance de l'échantillon.

La première idée serait de prendre comme estimateur de σ^2

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

Si nous calculons l'espérance de cet estimateur S^2 nous obtenons :

$$E(S^2) = \frac{(n-1)}{n} \sigma^2$$

S^2 est donc un estimateur biaisé de σ^2

En revanche, on peut montrer que S^2 est un estimateur convergent de σ^2

Ainsi, la variance de l'échantillon, S^2 , n'est pas un estimateur absolument correct de la variance de la population, car S^2 sous-estime la variance de la population.

Un meilleur estimateur $\hat{\sigma}^2$ de la variance de la population est obtenu en enlevant le biais de S^2 :

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

Propriété 3.2 s^2 est un estimateur absolument correct de la variance σ^2 de la population.

Nous allons donc considérer dorénavant s^2 comme estimateur de la variance σ^2 d'une population. L'exemple suivant permet de constater que s^2 est un meilleur estimateur que S^2 de la variance réelle σ^2 .

Exemple:

Supposons que la population est composée des trois nombres 1, 2 et 6.

La moyenne vaut

$$\mu = \frac{1 + 2 + 6}{3} = 3$$

et la variance vaut

$$\sigma^2 = \frac{(1-3)^2 + (2-3)^2 + (6-3)^2}{3} = \frac{14}{3}$$

Formons tous les échantillons possibles de taille 2 en prélevant 2 éléments au hasard dans la population. Comme la population n'est pas très vaste, on remplace le premier élément tiré dans la population avant de tirer le second. Nous obtenons les $\overline{A}_3^2 = 3^2 = 9$ échantillons possibles suivants :

$$(1; 1), (1; 2), (1; 6), (2; 1), (2; 2), (2; 6), (6; 1), (6; 2), (6; 6)$$

Les moyennes de ces couples de nombres sont :

$$1, \frac{3}{2}, \frac{7}{2}, \frac{3}{2}, 2, 4, \frac{7}{2}, 4, 6$$

Les variances S^2 de ces couples de nombres sont :

$$0, \frac{1}{4}, \frac{25}{4}, \frac{1}{4}, 0, 4, \frac{25}{4}, 4, 0$$

La moyenne des variances de ces couples de nombres est égale à $\frac{7}{3}$, ce qui n'est pas égale à la variance de la population ($\frac{14}{3}$). En revanche,

$$\frac{n}{n-1} \cdot \frac{7}{3} = \frac{2}{2-1} \cdot \frac{7}{3} = \frac{14}{3}$$

3.2. THÉORÈME CENTRAL LIMITE

Ainsi, afin d'obtenir une estimation non biaisée de la variance d'une population, on utilisera la formule

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

où n désigne le nombre d'observations dans l'échantillon

Remarque: Beaucoup de calculatrices et de logiciels statistiques donnent par défaut $\hat{\sigma}^2$ plutôt que S^2 . La valeur donnée par $\hat{\sigma}^2$ est correcte lorsque l'on veut estimer un paramètre de la population. En revanche, lorsque l'on fait de la statistique descriptive sur les données, c'est à dire lorsque les données sont considérée comme une population, il faut utiliser S^2 . La relation entre $\hat{\sigma}^2$ et S^2 est la suivante :

$$\hat{\sigma}^2 = \frac{n}{n-1} S^2$$

3.2 Théorème central limite

L'écart type $\sigma(\hat{\Theta})$ et la variance $\text{Var}(\hat{\Theta})$ sont des mesures de précision de l'estimateur $\hat{\Theta}$. Plus ils sont grands, moins bonne est l'estimation du paramètre étudié. Afin d'améliorer la précision, il faut alors soit trouver une meilleure formule pour estimer le paramètre, soit changer la méthode d'échantillonnage. En pratique, on cherche le plus souvent à réduire la variance en premier.

3.2.1 Distribution de la moyenne d'un échantillon

Le théorème suivant indique que si la population suit une loi normale, alors l'échantillon aussi.

Théorème 3.3 *Si une population est normalement distribuée, de moyenne μ et d'écart type σ , alors la distribution d'échantillonnage de la moyenne \bar{x} est aussi normalement distribuée de même moyenne que la population*

$$\mu_{\bar{x}} = \mu$$

et dont l'écart type vaut l'écart type de la population divisé par la racine carrée de la taille de l'échantillon

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Définition 3.6 *L'écart type de la distribution d'échantillonnage de la moyenne, aussi appelée erreur standard de la moyenne, est le terme*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3.2. THÉORÈME CENTRAL LIMITE

Remarque: Notons que l'erreur standard de la moyenne est toujours inférieure ou égale à l'écart type de la population.

La distribution d'échantillonnage de la moyenne est composée de toutes les moyennes possibles sur tous les échantillons de même taille. La distance relative d'une moyenne particulière à la moyenne d'échantillonnage (i.e. l'espérance) peut être déterminée par une variable centrée réduite z . Cette variable z mesure le nombre d'écart type entre la moyenne (i.e. l'espérance) et une valeur particulière (i.e. une moyenne d'un échantillon particulier).

Définition 3.7 La variable centrée réduite associée à la moyenne d'échantillonnage est la variable

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- \bar{x} = moyenne de l'échantillon
- μ = moyenne de la population
- σ = écart type de la population
- n = taille de l'échantillon

Exemple:

D'après l'étiquetage, la quantité de poisson contenue dans un plat surgelé de 1 kg suit une loi normale, de moyenne 500 g, avec une variation de ± 100 grammes (écart type). Un magazine de consommateurs achète 25 de ces plats surgelés et vérifie la quantité de poisson. Les résultats du test donne une moyenne de 490 g par plat. Peut-on affirmer qu'il y a tromperie du consommateur ?

1. La moyenne pour cet échantillon est de $\bar{x} = 490$
2. La distribution d'échantillonnage de la moyenne doit suivre une loi normale de moyenne $\mu_{\bar{x}} = \mu = 500$ et d'écart type $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{25}} = 20$
3. L'événement d'intérêt est le suivant : Comme la moyenne d'échantillon trouvée (490) est inférieure à la moyenne attendue (500), nous désirons connaître la probabilité d'un tel événement

$$P(\bar{x} \leq 490) = ?$$

4. Conversion de la moyenne d'échantillonnage en une valeur centrée réduite

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{490 - 500}{\frac{100}{\sqrt{25}}} = -0.5$$

5. Utiliser la distribution normale centrée réduite pour déterminer la probabilité désirée

$$P(z \leq -0.5) = 0.3085$$

3.2. THÉORÈME CENTRAL LIMITE

Il y a donc une probabilité d'environ 0.3 qu'une telle valeur (490) soit la moyenne d'un échantillon de 25 plats surgelé.

Remarque: Si la taille de l'échantillon est *plus du 5%* de la taille de la population, et que l'échantillon tiré est fait *sans remise*, alors un facteur de correction pour une population finie doit être considéré. Ce facteur de correction, qui s'applique à la valeur de l'écart type $\sigma_{\bar{x}}$, est de

$$\sqrt{\frac{N-n}{N-1}}$$

où

N = taille de la population

n = taille de l'échantillon

Beaucoup de méthodes statistiques reposent sur l'hypothèse selon laquelle les données observées sont distribuées selon une loi normale. En pratique, cela n'est pas toujours vérifié, mais le **théorème central limite** assure que même si la population ne satisfait pas à la normalité, la moyenne d'un échantillon de grande taille issu de celle-ci est distribuée de façon normale, ce qui est suffisant pour l'emploi de la plupart des outils statistiques.

Théorème 3.4 [*Théorème central limite*] Soit une suite (X_1, X_2, \dots, X_n) de n variables aléatoires identiquement et indépendamment distribuées (μ, σ^2) . Lorsque $n \rightarrow \infty$, la distribution de

$$\bar{X} = \frac{1}{n} \sum_i X_i$$

tend vers la loi $\mathcal{N}(\mu, \frac{\sigma^2}{n})$

Remarque: Plus la taille de l'échantillon augmente, meilleure est l'approximation par la loi normale.

En d'autres termes, pour des échantillons de taille suffisamment grande, en pratique de taille au moins 30, on remarque que la distribution d'échantillonnage a une forme en "courbe en cloche". Il s'agit de la loi de Gauss, ou loi normale. Un plan de sondage très précis correspond à une courbe très peu étalée. Si le plan de sondage est en plus très peu biaisé, alors le pic de la courbe est au voisinage de la vraie valeur du paramètre θ .

Ainsi, pour n suffisamment grand

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \approx \mathcal{N}(0, 1)$$

quelle que soit la distribution des X_i .

3.2. THÉORÈME CENTRAL LIMITE

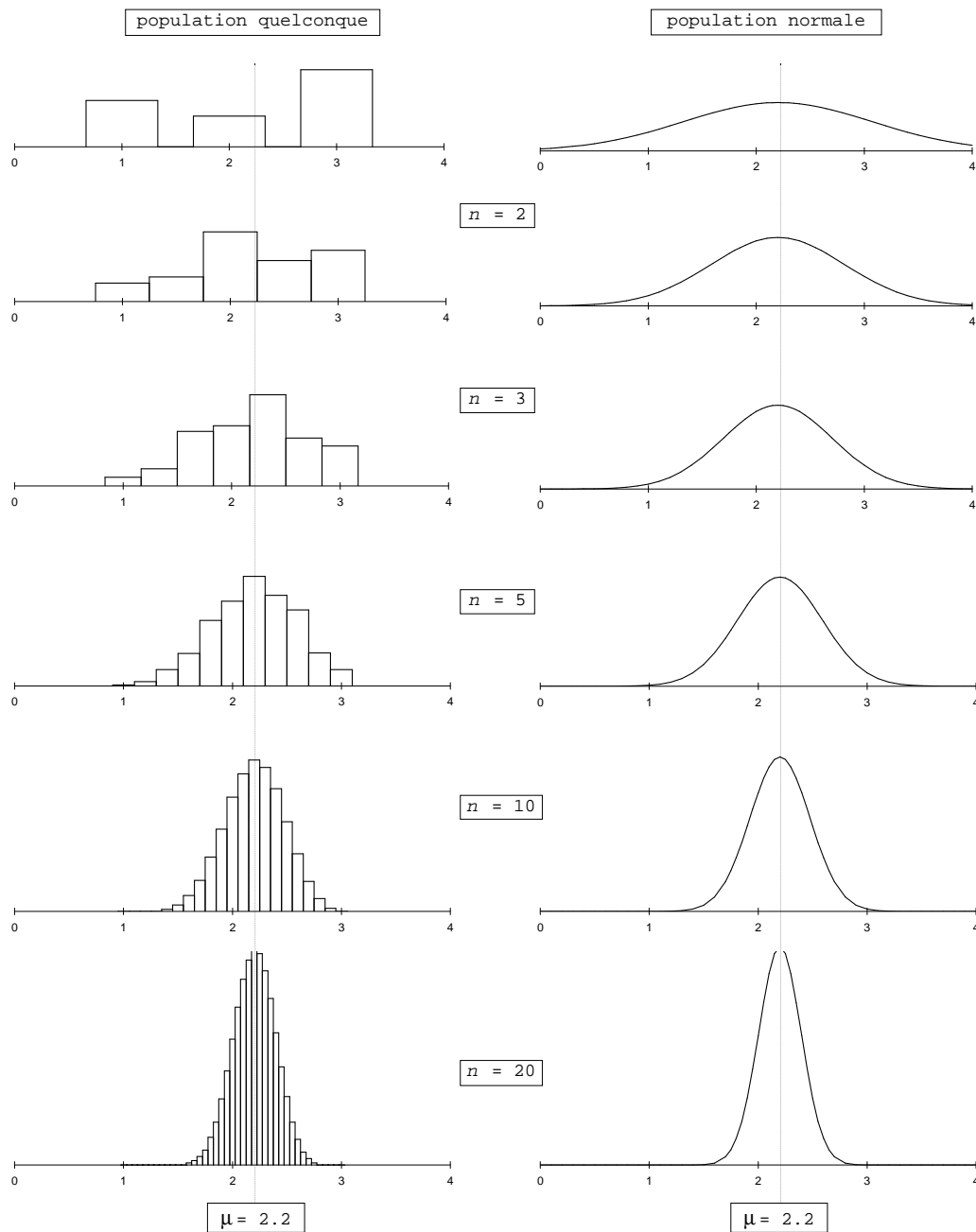


FIGURE 3.1 – Distribution de \bar{X} en fonction de la taille n d'échantillon.

3.2.2 Distribution d'une proportion d'un échantillon

Remarquons qu'une proportion peut être vue comme une moyenne particulière, dans laquelle la variable d'investigation vaut 1 si la caractéristique recherchée est présente, et 0 sinon. Une telle variable est appelée variable booléenne, ou dichotomique.

Notations :

X	nombre d'individus dans la population avec la caractéristique
N	taille de la population
x	nombre d'individus dans l'échantillon avec la caractéristique
n	taille de l'échantillon
$\pi = \frac{X}{N}$	proportion d'individus dans la population ayant la caractéristique
$\bar{p} = \frac{x}{n}$	proportion d'individus dans l'échantillon avec la caractéristique
$\bar{p} - \pi$	erreur d'échantillonnage

Si la taille de l'échantillon est suffisamment grande, en pratique, si

$$n\pi \geq 5 \quad \text{et} \quad n(1 - \pi) \geq 5$$

alors la distribution normale peut être utilisée.

Théorème 3.5 *Lorsque la taille de l'échantillon est suffisamment grande, en pratique si $n\pi \geq 5$ et $n(1 - \pi) \geq 5$, alors la distribution d'échantillonnage de la proportion \bar{p} est caractérisée par*

Moyenne	Écart-type
$\mu_{\bar{p}} = \pi$	$\sigma_{\bar{p}} = \sqrt{\frac{\pi(1-\pi)}{n}}$

π	proportion dans la population
n	taille de l'échantillon
\bar{p}	proportion dans l'échantillon

Exemple:

Le responsable d'une agence immobilière souhaite passer une annonce vantant la rapidité de traitement des affaires. Il pense que le 80% des propriétés à vendre trouvent preneur en au plus 4 mois. Il a sélectionné aléatoirement 100 affaires, et parmi celles-là, 73 se sont terminées en au plus 4 mois. Les étapes suivantes déterminent la probabilité de ce résultat :

1. Déterminer la proportion de la population.
La population est supposée avoir une proportion de 0.8, basée sur l'intuition du manager.
2. Calculer la proportion de l'échantillon.

$$\bar{p} = \frac{73}{100}$$

3. Déterminer la moyenne et l'écart type de la distribution d'échantillonnage.

$$\mu_{\bar{p}} = 0.8 \quad \sigma_{\bar{p}} = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.8(1-0.8)}{100}} = 0.04$$

4. Définir l'événement d'intérêt.

$$P(\bar{p} \leq 0.73) = ?$$

5. Vérifier les hypothèses.

Comme $n\pi = 80$ et $n(1-\pi) = 20$ sont suffisamment grands (supérieurs à 5), convertir \bar{p} en la variable centrée réduite z

$$z = \frac{\bar{p} - \pi}{\sigma_{\bar{p}}} = \frac{0.73 - 0.8}{\sqrt{\frac{0.8(1-0.8)}{100}}} = -1.75$$

6. Déterminer la probabilité.

$$P(\bar{p} \leq 0.73) = P(z \leq -1.75) = 0.0401$$

Ainsi, il y a seulement 4% de chance que sur un échantillon aléatoire de taille 100, 73 affaires ou moins trouvent preneur en au plus 4 mois.

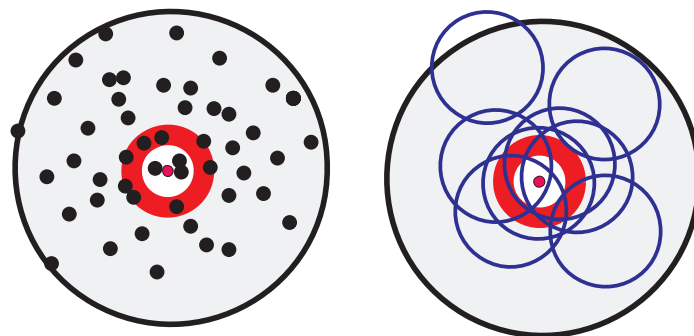
3.3 Intervalle de confiance

Une estimation ponctuelle attribue une valeur précise à un paramètre, mais engendre le risque que la valeur ainsi obtenue soit relativement éloignée de la réalité. L'idée de l'estimation par intervalle est de calculer un intervalle dans lequel se trouve la valeur du paramètre cherché, avec un niveau de confiance fixé.

Soit un paramètre θ . Au lieu de fournir une estimation $\hat{\theta}$, on construit un intervalle de valeurs de la forme

$$[\hat{\theta}_{inf} ; \hat{\theta}_{sup}]$$

dans lequel la vraie valeur du paramètre a une certaine probabilité fixée à l'avance, notée $1 - \alpha$, de se trouver.



Estimations ponctuelles Estimations par intervalle

Définition 3.8 $1 - \alpha$ est appelé le degré de confiance ou niveau de confiance ; il indique la probabilité que la vraie valeur du paramètre θ soit comprise dans l'intervalle de confiance.

$$\underbrace{1 - \alpha}_{\text{degré de confiance}} = P\left(\theta \in \underbrace{[\hat{\theta}_{\inf} ; \hat{\theta}_{\sup}]}_{\text{intervalle aléatoire}}\right)$$

Définition 3.9 Le risque de première espèce α est le risque que l'intervalle ne recouvre pas θ .

Le choix du degré de confiance est crucial, car il influence directement l'utilité des résultats :

- Si α est très petit, l'intervalle est très fiable, mais il devient tellement grand qu'il ne nous renseigne plus de façon utile sur la vraie valeur du paramètre.
- Si α est très grand, l'intervalle est très précis (= étroit), mais la probabilité qu'il recouvre effectivement la vraie valeur du paramètre est faible.

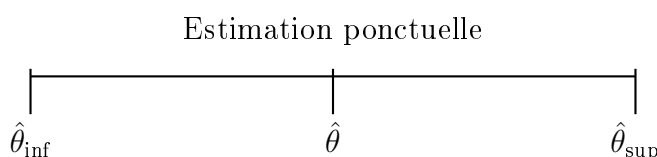
En pratique, on choisit généralement un risque α de 5% ou de 10 %.

Définition 3.10 Un intervalle de confiance de niveau $1 - \alpha$ pour un paramètre inconnu θ d'une population est un intervalle tel que la probabilité pour que θ appartienne à cet intervalle est $1 - \alpha$. Les bornes de cet intervalle se calculent à partir d'un échantillon.

3.3.1 Construction d'un intervalle de confiance

Tout intervalle de confiance se construit selon le schéma suivant :

$$\text{Estimation ponctuelle} \pm (\text{Valeur critique}) \cdot (\text{Écart type})$$



Après avoir décidé du degré de confiance à utiliser, la construction d'un intervalle de confiance pour un paramètre θ quelconque s'effectue en suivant les 3 étapes suivantes :

1. Choix d'une statistique dont la distribution est connue.

$$f(\theta) \sim \mathcal{D}$$

Remarque : Le paramètre pour lequel on construit l'intervalle de confiance doit pouvoir être explicité à partir de la distribution choisie.

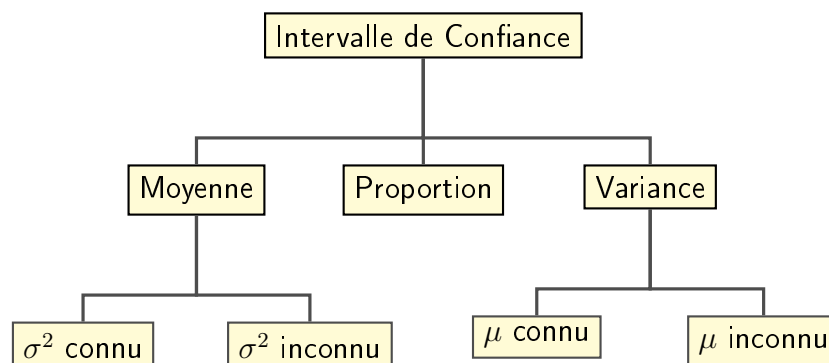
2. Construction de l'intervalle pour la statistique choisie :

$$[\mathcal{D}_{\text{inf}} ; \mathcal{D}_{\text{sup}}]$$

3. Construction de l'intervalle pour le paramètre :

$$[\hat{\theta}_{\text{inf}} ; \hat{\theta}_{\text{sup}}]$$

Nous allons tout d'abord étudier l'intervalle de confiance pour estimer une moyenne. Deux cas peuvent se présenter : soit la variance (ou l'écart type) de la population est connue, soit elle ne l'est pas. Ensuite, nous étudierons l'intervalle de confiance pour une proportion, et finalement pour une variance.



3.4 IC pour estimer μ , σ^2 connu

Lorsque la population dont on cherche à estimer la moyenne suit une loi normale de variance σ^2 connue, l'intervalle de confiance est calculé de la manière suivante :

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

où :

- \bar{x} = moyenne de l'échantillon
- $z_{\alpha/2}$ = valeur critique de la distribution normale standard pour un degré de confiance de $1 - \alpha$
- σ = écart type de la population
- n = taille de l'échantillon

Exemple:

Considérons la population suivante : l'ensemble des pots de peinture de 1lt remplis par une machine industrielle. Supposons que la quantité de peinture soit une variable aléatoire X suivant une loi normale d'écart type $\sigma = 0.04$ lt. Votre but en tant que responsable qualité est de contrôler qu'en moyenne, la machine remplisse 1lt de peinture par pot.

Vous prélevez 4 pots de peinture au hasard et mesurez la quantité de peinture dans chaque pot :

Pot	x_1	x_2	x_3	x_4
Quantité [lt]	1.0	0.98	1.1	1.1

Vous voulez connaître l'intervalle de confiance à 95% pour la quantité de peinture par pot.

1. La population d'intérêt est l'ensemble des pots de peinture d'1lt remplis par la machine
2. Le degré de confiance $1 - \alpha$ vaut 0.95. Donc $\alpha = 0.05$
3. La moyenne de l'échantillon est $\bar{x} = \frac{1.0+0.98+1.1+1.1}{4} = 1.045$
4. L'erreur standard de la moyenne (=écart type de l'estimateur) vaut

$$\sigma_{\bar{x}} = \frac{0.04}{\sqrt{4}} = 0.02$$

5. Les pots peuvent être trop peu remplis, ou trop remplis. L'erreur de première espèce α est alors divisée en 2 parties. La valeur critique est donc $z_{\alpha/2} = z_{0.025} = 1.96$
6. L'intervalle de confiance est

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.045 \pm 1.96 \cdot 0.02 = [1.0058; 1.0842]$$

Comme l'intervalle de confiance ne comprend pas la valeur de 1lt, vous concluez que la machine n'est pas bien réglée car elle remplit trop les pots en moyenne.

Les étapes suivantes permettent de calculer l'intervalle de confiance estimé pour une moyenne de population, lorsque l'écart type de la population est connue, et la moyenne suit une loi normale ou lorsque l'échantillon est de taille au moins 30.

1. Définir la population d'intérêt et sélectionner un échantillon aléatoire de taille n
2. Spécifier le degré de confiance $1 - \alpha$
3. Calculer la moyenne de l'échantillon

$$\bar{x} = \frac{\sum x_i}{n}$$

4. Déterminer l'erreur standard de la moyenne

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

5. Déterminer la valeur critique $z_{\alpha/2}$
6. Calculer l'intervalle de confiance

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

3.5 IC pour estimer μ, σ^2 inconnu

Considérons l'*hypothèse suivante* : la distribution de la population suit une loi normale. Dans la plupart des cas où la moyenne de la population est inconnue, la variance de la population est aussi inconnue. Il est alors nécessaire d'estimer la variance de la population à l'aide de l'échantillon. Il faut alors modifier la façon dont sont calculés la valeur critique et l'écart type.

$$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

où :

- \bar{x} = moyenne de l'échantillon
- $t_{\alpha/2, n-1}$ = valeur critique de la t -distribution à $n - 1$ degrés de liberté pour un degré de confiance de $1 - \alpha$
- s = écart type de l'échantillon
- n = taille de l'échantillon

Exemple:

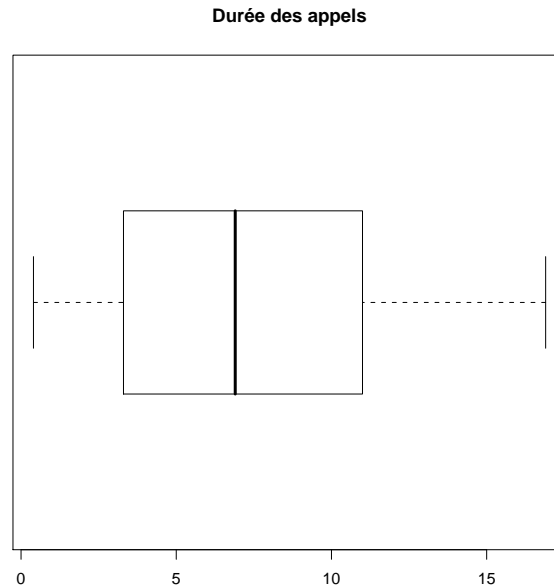
En tant que responsable d'un "backoffice" dans une entreprise, vous souhaitez calculer l'intervalle de confiance à 95% du temps moyen passé au téléphone par les employés du "backoffice" avec les clients. Vous avez recueilli les temps, en minutes, de 25 appels.

7.1	13.6	1.4	3.6	1.9
11.6	1.7	16.9	2.6	7.7
12.4	11.0	3.7	14.6	8.8
8.5	6.1	3.3	6.1	6.9
0.4	11.0	0.8	6.4	9.1

1. la population consiste en tous les appels des clients au "backoffice", et l'échantillon contient les 25 durées sélectionnées au hasard.
2. le niveau de confiance souhaité est de $1 - \alpha = 0.95$
3. la moyenne vaut $\bar{x} \approx 7.088$ et l'écart type vaut $s \approx 4.64$
4. l'erreur standard de la distribution d'échantillonnage vaut

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \approx 0.928$$

5. Comme vous ne savez pas à priori si la population est normalement distribuée, vous vérifiez à l'aide d'une boîte à moustache que la distribution de votre échantillon soit normalement distribuée :



la valeur critique vaut $t_{0.025,24} = 2.0639$

6. l'intervalle de confiance vaut alors

$$7.088 \pm 2.0639 \cdot 0.928 \text{ i.e. } [5.173; 9.003]$$

Les étapes suivantes permettent de calculer l'intervalle de confiance estimé pour une moyenne de population, lorsque la variance de la population est inconnue, mais que la population est distribuée suivant une loi normale, et avec des échantillons de petite taille (<30).

1. Définir la population d'intérêt et sélectionner un échantillon aléatoire de taille n
2. Spécifier le degré de confiance $1 - \alpha$
3. Calculer la moyenne et l'écart type de l'échantillon

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

4. Déterminer l'erreur standard de la moyenne

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

5. Déterminer la valeur critique $t_{\alpha/2, n-1}$
6. Calculer l'intervalle de confiance

$$\bar{x} \pm t_{\alpha/2, n-1} \sigma_{\bar{x}}$$

3.6. IC POUR ESTIMER UNE PROPORTION

Remarque: Sous R (version 2.9.1) vous pouvez utiliser la fonction `t.test()` pour trouver l'intervalle de confiance.

Remarque: Lorsque l'échantillon est de grande taille, c'est à dire de taille au moins 30, la statistique de test peut être approchée à l'aide d'une z -valeur. On peut donc dans ce cas utiliser la formule suivante pour calculer l'intervalle de confiance :

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

où :

- \bar{x} = moyenne de l'échantillon
- $z_{\alpha/2}$ = valeur critique de la distribution normale standard
pour un degré de confiance de $1 - \alpha$
- s = écart type de l'échantillon
- n = taille de l'échantillon

3.6 IC pour estimer une proportion

Lorsque la variable d'intérêt est une proportion, il est aussi possible d'utiliser le théorème central limite 3.4, car une proportion n'est rien d'autre qu'une moyenne particulière : la variable d'investigation vaut 1 si la caractéristique recherchée est présente, et 0 sinon.

Exemple:

$$Y_i = \begin{cases} 1 & \text{la personne a l'intention d'acheter le produit} \\ 0 & \text{sinon} \end{cases}$$
$$Y_i = \begin{cases} 1 & \text{la personne a l'intention de voter pour Mme L.U.} \\ 0 & \text{sinon} \end{cases}$$
$$Y_i = \begin{cases} 1 & \text{le client est satisfait du service} \\ 0 & \text{sinon} \end{cases}$$

Théorème 3.6 Dans un sondage aléatoire simple, la proportion dans l'échantillon \bar{p} est un estimateur sans biais de la proportion π dans la population.

Théorème 3.7 Lorsque la taille n de l'échantillon est suffisamment grande, i.e. $n\pi \geq 5$ et $n(1 - \pi) \geq 5$, la distribution d'échantillonnage peut être approchée par une distribution normale centrée en π , avec comme écart type

$$\sigma_{\bar{p}} = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

où

- π = proportion dans la population
- n = taille de l'échantillon

Comme le paramètre π n'est pas connu, il est alors simplement estimé par \bar{p} .

L'intervalle de confiance pour une proportion s'écrit donc, sous l'hypothèse que la taille de l'échantillon soit suffisamment grande :

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

où

- \bar{p} = proportion dans l'échantillon
- π = proportion dans la population
- n = taille de l'échantillon
- $z_{\alpha/2}$ = valeur critique de la distribution normale standard pour un degré de confiance de $1 - \alpha$

Exemple:

Une entreprise réunissant plusieurs marques désire estimer la proportion de ses clients connaissant plus de 5 de leur marques. Elle veut un degré de confiance de 0.9. Elle effectue alors un sondage aléatoire parmi 100 clients et obtient comme estimation ponctuelle 0.2.

1. La population d'intérêt est l'ensemble de ses clients.
2. L'échantillon aléatoire sélectionné est suffisamment grand si la proportion cherchée est entre 0.05 et 0.95, ce qui est le cas ici.
3. Le degré de confiance est $1 - \alpha = 0.9$
4. La valeur critique $z_{\alpha/2} = z_{0.05} = 1.645$
5. La proportion estimée est $\bar{p} = 0.2$
6. L'intervalle de confiance vaut donc

$$0.2 \pm 1.645 \sqrt{\frac{0.2(1 - 0.2)}{100}} = [0.134, 0.266]$$

Les étapes suivantes permettent de calculer l'intervalle de confiance estimé d'une proportion

1. Définir la population d'intérêt et la variable dont on veut estimer la proportion.
2. Sélectionner un échantillon aléatoire de taille n suffisamment grande, telle que

$$n\bar{p} \geq 5 \quad \text{et} \quad n(1 - \bar{p}) \geq 5$$

3. Spécifier le degré de confiance $1 - \alpha$
4. Déterminer la valeur critique $z_{\alpha/2}$ tirée d'une loi normale centrée réduite
5. Calculer la proportion \bar{p}
6. Calculer l'intervalle de confiance

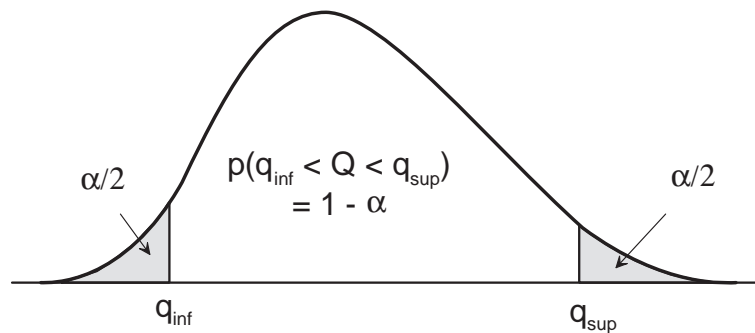
$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

3.7 IC pour estimer σ^2 , μ inconnu

L'estimation de la variance d'une population est utilisée par exemple pour mesurer la fiabilité d'un site de production ou d'un fournisseur. La fiabilité d'un instrument de mesure comme un altimètre est cruciale : il n'est pas suffisant de savoir qu'un altimètre donne en moyenne la bonne mesure (!) mais il faut aussi que les écarts par rapport à la moyenne soient suffisamment faibles.

La propriété 2.6 permet de calculer l'intervalle de confiance d'une variance :

$$\begin{aligned} P(q_{\inf} \leq Q_{n-1} \leq q_{\sup}) &= 1 - \alpha \\ P(\chi_{1-\alpha/2}^2 \leq Q_{n-1} \leq \chi_{\alpha/2}^2) &= 1 - \alpha \\ P\left(\chi_{1-\alpha/2}^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{\alpha/2}^2\right) &= 1 - \alpha \\ P\left(\frac{\chi_{1-\alpha/2}^2}{(n-1)s^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi_{\alpha/2}^2}{(n-1)s^2}\right) &= 1 - \alpha \\ P\left(\frac{(n-1)s^2}{\chi_{\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}\right) &= 1 - \alpha \end{aligned}$$



L'intervalle de confiance d'une variance s'écrit donc, sous l'hypothèse que l'échantillon aléatoire provient d'une population dont les éléments sont **iid de distribution normale** :

$$\left[\frac{(n-1)s^2}{q_{\alpha/2, n-1}}; \frac{(n-1)s^2}{q_{1-\alpha/2, n-1}} \right]$$

où

- n = taille de l'échantillon
- $q_{\alpha/2, n-1}$ = valeur critique de la distribution χ^2 à $n-1$ degrés de liberté pour un degré de confiance de $1-\alpha$
- s^2 = variance de l'échantillon

Remarque: L'utilisation de la statistique du χ^2 pour estimer la variance est très sensible à une violation de l'hypothèse d'une population normalement distribuée. Cette technique n'est donc pas une technique robuste.

Exemple:

Une entreprise fabriquant des photocopieuses assure aussi le service après-vente. Une nouvelle équipe vient d'être formée, et le responsable doit prévoir l'affectation des services. Il veut en estimer avec un degré de confiance de 0.9 le temps moyen de service et l'écart type, afin de pouvoir planifier les services de la nouvelle équipe. Pour cela, il inscrit la durée de 20 services pris aléatoirement parmi ceux effectué par la nouvelle équipe. Il calcule un écart type de 0.5h et peut supposer que les données proviennent d'une loi normale.

1. La population d'intérêt est l'ensemble des services effectués par la nouvelle équipe.
2. L'échantillon sélectionné est de taille 20, et provient d'une population normalement distribuée.
3. Le degré de confiance est de 0.9
4. Les valeurs critiques sont $\chi_{0.05}^2 = 30.14$ et $\chi_{0.95}^2 = 10.12$ avec 19 degrés de liberté.
5. La variance estimée est de $s^2 = 0.5^2 = 0.25$
6. L'intervalle de confiance associé à la variance est donc

$$\left[\frac{(19)0.25}{30.14}; \frac{(19)0.25}{10.12} \right] = [0.16; 0.47]$$

3.8 Résumé

Le tableau suivant résume les distributions utilisées pour représenter la moyenne, la proportion et la variance de la population :

estimé	hypothèse	distribution
μ	σ^2 connu, distr. normale ou $n \geq 30$	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$
	σ^2 inconnu, distr. normale	$T_{(n-1)} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \mathcal{T}_{n-1}$
π	$n\pi \geq 5$ et $n(1 - \pi) \geq 5$	$Z = \frac{\bar{P} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim \mathcal{N}(0, 1)$
σ^2	μ connu, distr. normale	$Q_{(n)} = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$
	μ inconnu, distr. normale	$Q_{(n-1)} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$

3.8. RÉSUMÉ

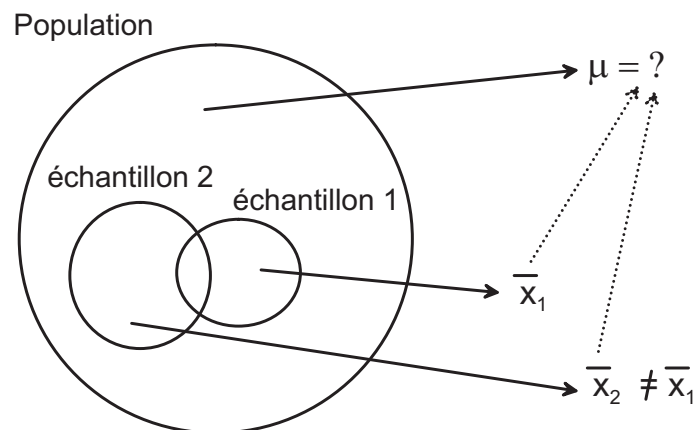
Rappel : lorsque la population est distribuée selon une loi quelconque de paramètres μ et σ^2 , il est possible d'utiliser le théorème central limite 3.4 pour estimer la moyenne, à condition que l'échantillon utilisé soit de grande taille. En pratique, un échantillon de taille au moins 30 suffit.

estimé	hypothèse	intervalle
μ	σ^2 connu, distr. normale ou $n \geq 30$	$\mu \in \bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
	σ^2 inconnu, distr. normale	$\mu \in \bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$
π	$n\bar{p} \geq 5$ et $n(1 - \bar{p}) \geq 5$	$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$
σ^2	μ connu, distr. normale	$\left[\frac{\sum (x_i - \mu)^2}{q_{\frac{\alpha}{2}, n}} ; \frac{\sum (x_i - \mu)^2}{q_{1-\frac{\alpha}{2}, n}} \right]$
	μ inconnu, distr. normale	$\left[\frac{\sum (x_i - \bar{x})^2}{q_{\frac{\alpha}{2}, n-1}} ; \frac{\sum (x_i - \bar{x})^2}{q_{1-\frac{\alpha}{2}, n-1}} \right]$

Chapitre 4

Tests paramétriques

4.1 Principe des tests d'hypothèses



Le but de l'estimation est de quantifier, alors que le but d'un test est de valider/invalider une hypothèse. Pour cela, on procède de la manière suivante :

1. On formule une hypothèse sur la population ou sur la distribution étudiée.
2. On vérifie si l'échantillon utilisé provient bien de la population ou de la distribution étudiée, avec un certain degré de confiance.

Le test statistique lui-même correspond à la règle de décision.

Définition 4.1 Les tests paramétriques présupposent que les données sont distribuées selon une loi particulière. Le test s'effectue alors en comparant certains paramètres de la distribution, comme la moyenne ou la variance.

Définition 4.2 Les tests non-paramétriques ne présupposent pas de distribution particulière des données. Les tests s'effectuent en tirant des conclusions à partir des valeurs

observées des échantillons, sans faire appel à des paramètres tels que la moyenne ou la variance.

Les tests non-paramétriques ont un champ d'application plus large que les tests paramétriques, car ils ne nécessitent pas une distribution particulière des données. En revanche, ils sont généralement moins puissants et le risque total d'erreur est plus grand.

4.1.1 Hypothèses nulle et alternative

Un test consiste à choisir entre deux hypothèses en fonction d'un échantillon. L'hypothèse nulle, H_0 , est celle que l'on veut tester. L'hypothèse alternative, H_1 , est son "contraire".

Définition 4.3 *L'hypothèse nulle H_0 est l'affirmation au sujet d'un paramètre de la population, qui sera testée. L'hypothèse nulle sera rejetée seulement si la statistique calculée à partir de l'échantillon, apporte une contradiction suffisamment forte.*

Définition 4.4 *L'hypothèse alternative H_1 est l'ensemble des valeurs d'un paramètre de la population non couvert par l'hypothèse nulle. L'hypothèse alternative est réputée vraie si l'hypothèse nulle est rejetée.*

H_0 : hypothèse nulle (à tester)

H_1 : hypothèse alternative

En fonction des hypothèses de départ, le résultat d'un test est toujours univoque : soit il y a suffisamment d'évidence pour rejeter l'hypothèse nulle, soit il n'y en a pas suffisamment (et donc *a fortiori* acceptation de l'hypothèse nulle).

Exemple:

On suppose que les petites entreprises suisses ont en moyenne 35 employés avec une variance égale à 220. Un échantillon aléatoire de taille 20 a donné un nombre moyen d'employés égal à 27 et une variance de 334.7.

2	2	5	6	7	8	12	14	23	26
28	31	40	42	46	47	48	49	52	52

En fonction du résultat donné par l'échantillon, est-il statistiquement admissible que le nombre moyen d'employés dans l'ensemble des petites entreprises du pays soit réellement égal à 35? Pour répondre à cette question, il faut former l'hypothèse nulle et l'hypothèse alternative suivantes :

H_0 : $\mu = 35$

H_1 : $\mu \neq 35$

Les règles suivantes aident à formuler l'hypothèse nulle et l'hypothèse alternative :

1. L'hypothèse nulle et l'hypothèse alternative doivent être formulées en termes du paramètre de la population d'intérêt.
2. L'hypothèse nulle représente un statu quo. Elle représente la condition qui est supposée exister, à moins qu'il n'y ait suffisamment d'évidence montrant que la situation a changé.
3. L'hypothèse nulle doit contenir un signe d'égalité $=$ ou \geq ou \leq .

Exemple:

Les étudiants en emploi bénéficient d'un horaire tenant compte de leur situation, à savoir un travail de 25h par semaine, correspondant à un temps partiel de 60%. Afin de tenir compte au mieux de leur situation réelle, le management désire savoir si la situation réelle a changé. Les étapes suivantes permettent de formuler l'hypothèse nulle et l'hypothèse alternative :

1. Déterminer le paramètre de la population d'intérêt
Le paramètre à considérer est le nombre d'heures de travail moyen par semaine des étudiants en emploi.
2. Définir la situation qui est supposée vraie à moins qu'il y ait suffisamment d'évidence que cela ne soit plus le cas.
Le statu quo est de considérer que les étudiants en emploi travaillent 25 heures ou moins par semaine en moyenne. S'il y a suffisamment d'évidence que la moyenne est plus grande que 25h/semaine, alors cela impliquera un effort considérable de réorganisation. Dans ce dernier cas, il s'agit de rejeter le status quo.
3. Formuler l'hypothèse nulle et l'hypothèse alternative

$$H_0 : \mu \leq 25$$

$$H_1 : \mu > 25$$

Remarque: Une *hypothèse simple* correspond à une valeur spécifique, une situation déterminée (une hypothèse sur une égalité). Une *hypothèse composite* correspond à un ensemble de valeurs, de situations (une hypothèse sur une inégalité).

4.1.2 Région critique et valeur critique

Le principe du test est de rejeter l'hypothèse H_0 si la valeur de la statistique Q_0 observée dans l'échantillon est trop différente de la valeur théorique postulée sous H_0 pour la population.

A cette fin, l'espace des valeurs possibles de la statistique Q_0 est divisé en deux zones :

- A : ensemble des valeurs probables de Q_0 lorsque l'hypothèse H_0 est vraie.

Il s'agit de la région d'acceptation de H_0 .

- R : ensemble des valeurs peu probables de Q_0 lorsque l'hypothèse H_0 est vraie.
Il s'agit de la région de rejet de H_0 (région critique).

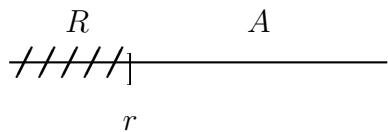
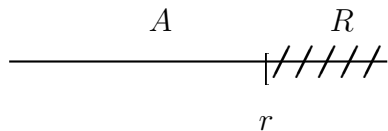
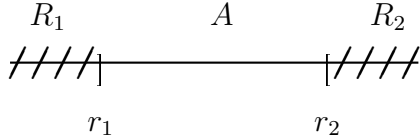
Règle de décision

$$q_0 \notin R \iff \text{Acceptation de } H_0$$

$$q_0 \in R \iff \text{Rejet de } H_0$$

Forme de la région critique R

La forme de la région critique R dépend du type d'hypothèse alternative retenue. On distingue les trois cas suivants :

- test unilatéral à gauche
 $H_1 : q = q_1 < q_0$

- test unilatéral à droite
 $H_1 : q = q_1 > q_0$

- test bilatéral
 $H_1 : q = q_1 \neq q_0$


Définition 4.5 La valeur critique r , aussi appelée seuil critique, est la valeur d'une statistique correspondante à un certain niveau de signification. Cette valeur seuil détermine la frontière entre les échantillons dont le test statistique conduit au rejet de l'hypothèse nulle, et les échantillons dont le test statistique ne conduit pas au rejet de l'hypothèse nulle.

La valeur critique r (ou les valeurs critiques r_1 et r_2) est choisie de façon à limiter le risque d'erreur.

4.1.3 Risques de première et de seconde espèce

Lorsqu'une hypothèse est testée, il y a un risque d'erreur dû à l'échantillonnage. Ce risque d'erreur, appelé aussi risque total d'erreur, se compose de deux parties : un risque α de première espèce et un risque β de deuxième espèce.

Définition 4.6 Le risque de première espèce α , dit aussi risque de type I, est le risque de rejeter l'hypothèse nulle H_0 alors qu'elle est en fait vraie. Il est aussi appelé niveau de signification.

$$\alpha = P(Q_0 \in R \mid H_0)$$

Définition 4.7 Le risque de deuxième espèce β , dit aussi risque de type II, est le risque d'accepter l'hypothèse nulle H_0 alors qu'elle est en fait fausse.

$$\beta = P(Q_0 \notin R \mid H_1)$$

Voici un tableau résumant ces risques. Seulement l'une de ces quatre possibilités peut se réaliser lors d'un test d'hypothèses.

	État de la nature	
	H_0 vraie	H_1 vraie
H_0 accepté	correct	β
H_1 accepté	α	correct

Remarque: α est spécifié avant d'effectuer le test et β est directement influencé par ce choix. Ces deux erreurs sont inversement liées.

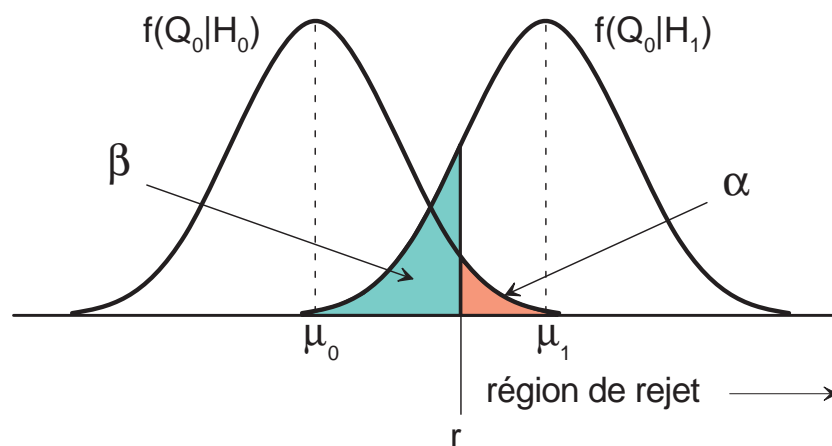


FIGURE 4.1 – Risques α et β

Définition 4.8 Le risque total d'erreur est défini par la relation

$$\alpha \underbrace{P(H_0)}_{\text{inconnu}} + \beta \underbrace{P(H_1)}_{\text{inconnu}}$$

Cette relation implique que le risque total est inconnu.

En pratique, on détermine le seuil critique r pour un α choisi arbitrairement petit (en général 5 % ou 10 %).

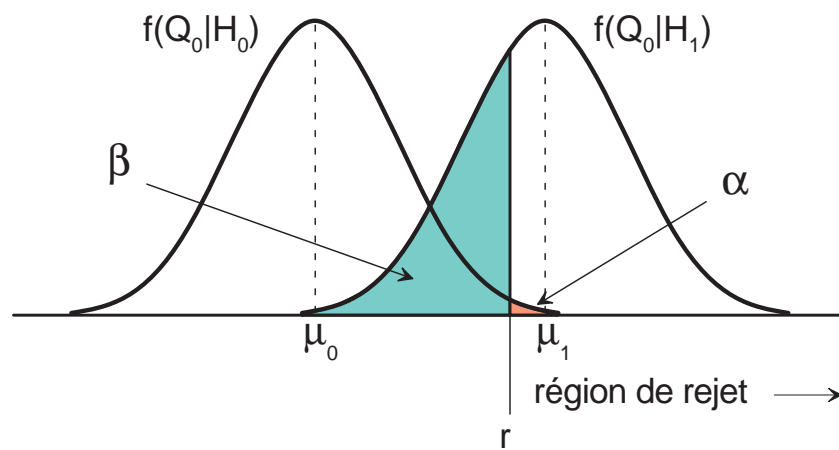


FIGURE 4.2 – α petit, donc β grand

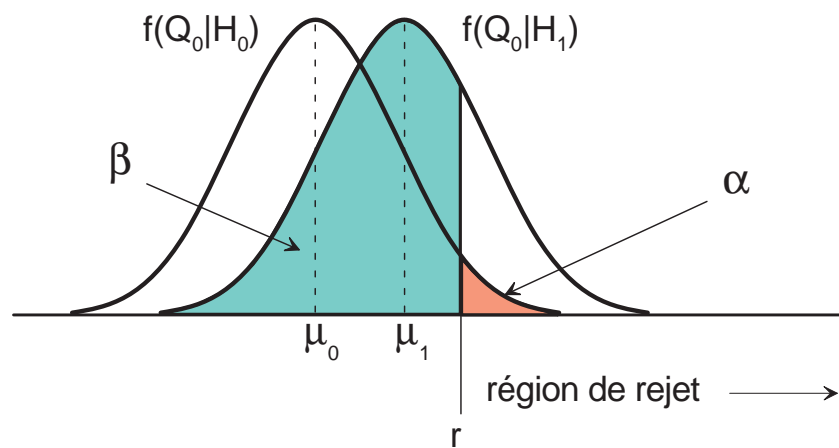


FIGURE 4.3 – H_1 peu différent de $H_0 \Rightarrow \beta$ grand

4.1.4 Procédure de test statistique

Pour tester une hypothèse, vous pouvez procéder ainsi :

1. Spécifier la valeur de la population d'intérêt.
2. Formuler l'hypothèse nulle H_0 et l'hypothèse alternative H_1
3. Choisir le niveau de signification α
4. Déterminer la région critique.
5. Calculer la statistique associée à l'échantillon.
6. Rejeter H_0 si la statistique appartient à la région critique.
Ne pas rejeter H_0 dans le cas contraire.
7. Énoncer une conclusion.

4.2 Test de μ , σ^2 connu, grand échantillon

Lorsque nous connaissons la variance et que l'échantillon est de grande taille $n \geq 30$, nous savons par le théorème central limite 3.4 que l'estimateur de la moyenne \bar{x} suit une loi normale de moyenne μ et d'écart type $\frac{\sigma}{\sqrt{n}}$. La statistique à utiliser est donc

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

où

- \bar{x} = moyenne de l'échantillon
- μ = moyenne supposée de la population étudiée
- σ = écart type de la population
- n = taille de l'échantillon

Exemple:

Une étude sur la mobilité des employés de l'entreprise "Jeux Mendors" affirme que la moyenne des temps de trajet des employés pour venir à leur travail excède 40 minutes. Vous devez tester cette affirmation avec une niveau de signification de 0.05. Pour cela, 100 employés vous ont indiqué leur temps de trajet actuel, dont la moyenne est 43.5 minutes. Basé sur une étude précédente, vous pouvez supposer que l'écart type de la population est de 8 minutes.

1. La valeur de la population d'intérêt est le temps de trajet.
2. Les hypothèses nulle et alternative sont :
 $H_0 \quad \mu \leq 40 \text{ minutes}$
 $H_1 \quad \mu > 40 \text{ minutes}$
3. Le niveau de signification est de 0.05
4. La région critique est $[z_{0.05} = 1.645; \infty[$
5. La statistique est $z = \frac{43.5-40}{\frac{8}{\sqrt{100}}} = 4.38$
6. Comme 4.38 appartient à la région critique, l'hypothèse H_0 est rejetée.
7. La conclusion est que la moyenne des temps de trajet excède 40 minutes.

4.3 Test de μ , σ^2 inconnu, grand échantillon

Dans la plupart des cas pratiques, la variance de la population n'est pas connue. Toutefois, il est possible d'estimer cette variance de la population à l'aide de la variance de l'échantillon. Si la taille de l'échantillon est suffisamment grande, en pratique d'au moins 30, la statistique de test peut être approchée par une z -valeur grâce au théorème central limite 3.4.

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

où

\bar{x} = moyenne de l'échantillon

μ = moyenne supposée de la population étudiée

s = écart type de l'échantillon

n = taille de l'échantillon

4.4 Test de μ , σ^2 inconnu, petit échantillon

Lorsque la variance de la population n'est pas connue et que l'on ne dispose que d'un échantillon de petite taille $n < 30$, nous pouvons utiliser la distribution de Student dans le test de la moyenne.

La statistique de test à utiliser suit une loi de Student à $n - 1$ degrés de libertés :

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

où

\bar{x} = moyenne de l'échantillon

μ = moyenne supposée de la population étudiée

s = écart type de l'échantillon

n = taille de l'échantillon

Toutefois, afin de pouvoir utiliser la t -distribution (= distribution de Student), nous devons faire la **supposition suivante** :

LA POPULATION EST NORMALEMENT DISTRIBUÉE

Exemple:

Nous voulons tester la moyenne d'un échantillon du nombre d'employés par petite entreprise suisse contre la moyenne de la population, $\mu = 35$, en ne connaissant pas la valeur de la variance de la population. Par des études précédentes, nous savons que le nombre d'employés est distribué selon une loi normale. Pour cela, 20 petites entreprises suisses ont été interrogées sur le nombre d'employés qu'elles occupent. La moyenne calculée vaut $\bar{x} = 27$ et la

variance calculée vaut $s^2 = 334.7$. Nous souhaitons un niveau de signification de 0.05.

1. La valeur de la population d'intérêt est le nombre d'employés de petites entreprises suisses.
2. Les hypothèses nulle et alternative sont :
 $H_0 \quad \mu = 35$ personnes
 $H_1 \quad \mu \neq 35$ personnes
3. Le niveau de signification est de 0.05
4. La région critique est composée de 2 zones distinctes car nous avons un test bilatéral à effectuer.

$$\begin{array}{c}
 R = R_1 \cup R_2 = \{t_0 \mid t_0 \leq r_1\} \cup \{t_0 \mid t_0 \geq r_2\} \\
 \begin{array}{ccc}
 R_1 & A & R_2 \\
 \hline
 \text{////|} & \text{-----} & \text{|\\\\} \\
 r_1 & & r_2
 \end{array}
 \end{array}$$

Les seuils critiques r_1 et r_2 sont déterminés à partir de la table de la loi de Student :

$$P(T_0 \leq r_1) = \frac{\alpha}{2} = 0.025 \text{ et donc } r_1 = t_{\alpha/2,19} = -t_{1-\alpha/2,19} = -t_{0.975,19} = -2.093$$

$$P(T_0 \geq r_2) = 1 - \frac{\alpha}{2} = 0.975 \text{ et donc } r_2 = t_{1-\alpha/2,19} = t_{0.975,19} = 2.093$$

5. La statistique est

$$t = \frac{27 - 35}{\sqrt{\frac{334.7}{20}}} = -1.955587$$

6. Comme $t = -1.955587$ n'appartient pas à la région critique, l'hypothèse H_0 n'est pas rejetée.
7. La conclusion est que la vraie moyenne de la population peut être égale à 35 employés par petite entreprise suisse.

4.5 Value at Risk

Nous traitons ici très sommairement d'une mesure de risque très utilisée en finance. Le but n'est pas d'expliquer de manière exhaustive tout ce que le concept de Value at Risk (VaR) recouvre, mais simplement de donner un exemple d'application de test unilatéral sur la moyenne. Nous allons voir comment est calculée une VaR paramétrique par l'approche delta-normale.

Définition 4.9 La Value at Risk est définie comme la perte maximale sur un horizon donné T , avec un niveau de confiance $1 - \alpha$.

De manière statistique, en considérant une variable aléatoire X suivant une loi normale $\mathcal{N}(\mu, \sigma^2)$, la VaR s'exprime comme

$$P(X \leq \text{VaR}(\alpha, T)) = \alpha \quad \Leftrightarrow \quad P(X \geq \text{VaR}(\alpha, T)) = 1 - \alpha$$

La notion de risque financier peut être estimé à l'aide de l'écart type annuelle des performances financières, appelé dans le jargon bancaire la *volatilité*. L'horizon est quant à lui souvent donné en jours (1, 10, ...), qui est à mettre en regard du nombre de jours ouvrables considéré dans la branche (250 ou 252 généralement). Considérons X une variable aléatoire indiquant les rendements, suivant une même loi normale $\mathcal{N}(\mu, \sigma^2)$ sur une seule période. Alors, sur la période T , les rendements sont également gaussiens de moyenne μT et de variance $\sigma^2 T$. La variable centrée réduite s'écrit :

$$z_\alpha = \frac{\text{VaR}(\alpha, T) - \mu T}{\sigma \sqrt{T}}$$

Et donc

$$\text{VaR}(\alpha, T) = \mu T + z_\alpha \sigma \sqrt{T}$$

Exemple:

Source : Rapport annuel 2007 de l'UBS

La position du secteur Banque d'investissement de l'UBS, sur un horizon de 1 jour, à un niveau de confiance de 99%, en utilisant des données sur 5 ans, est de

- 160 millions en 2007
- 169 millions en 2006

4.6 Test de la proportion

Il est souvent utile de pouvoir contrôler l'évolution d'une proportion. Par exemple, un responsable qualité doit s'assurer qu'une machine de production reste bien calibrée et satisfait ainsi aux normes établies pour cette machine. Comme une proportion peut être considérée comme une moyenne particulière, les mêmes concepts peuvent être appliqués dans le cas d'une proportion.

La statistique de test à utiliser, suivant une loi normale, est :

$$z = \frac{\bar{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

où

- \bar{p} = proportion de l'échantillon
- π = proportion supposée de la population étudiée
- n = taille de l'échantillon

Toutefois, afin de pouvoir utiliser la distribution normale, nous devons faire la **supposition suivante** :

LA TAILLE DE L'ÉCHANTILLON DOIT ÊTRE SUFFISAMMENT GRANDE
 en pratique, il suffit que

$$n\bar{p} \geq 5 \quad \text{et} \quad n(1 - \bar{p}) \geq 5$$

Exemple:

Un contrôle interne doit être effectué dans une banque pour vérifier que les contrats des hypothèques accordées comportent tous les documents nécessaires. Parfois, un contrat ne comporte pas tous les documents, auquel cas, le dossier n'est pas complet. La banque a octroyé 22500 hypothèques. et exige qu'il n'y ait pas plus d'1% de contrats incomplets. L'équipe chargée du contrôle n'a pas le temps d'examiner les 22500 contrats et décide de contrôler 600 contrats choisis aléatoirement, et fixe le niveau de signification à 0.02. Elle trouve 9 contrats incomplets.

1. La valeur de la population d'intérêt est la proportion de contrats incomplets

$$\pi = 0.01$$

Comme la taille de l'échantillon est suffisamment grande :

$$600 \times 0.01 = 6 \geq 5 \quad \text{et} \quad 600 \times (1 - 0.01) = 594 \geq 5$$

l'utilisation de la statistique z est possible.

2. $H_0 : \pi \leq 0.01$
 $H_1 : \pi > 0.01$
3. Le niveau de signification est $\alpha = 0.02$
4. La région critique est $[z_\alpha \approx 2.05; \infty[$
5. La statistique associée à l'échantillon est

$$\bar{p} = 9/600 = 0.015 \quad z = \frac{0.015 - 0.01}{\sqrt{\frac{0.01(1-0.01)}{600}}} = 1.23$$

6. Comme $z = 1.23 < 2.05$, H_0 n'est pas rejetée.
7. L'équipe de contrôle interne peut donc supposer que l'exigence de la banque est satisfaite.

Remarque:

- Lors d'un test unilatéral, la valeur critique est déterminée par z_α
- Lors d'un test bilatéral, les valeurs critiques sont déterminées par $\pm z_{\alpha/2}$

4.7 Test de la variance avec moyenne inconnue

Lorsque l'intérêt se porte sur la dispersion des données, l'indicateur utilisé le plus souvent est l'écart type. Or, il n'existe pas de test sur l'écart type, mais par contre, des tests

existent sur la variance. Comme généralement la moyenne de la population n'est pas connue, afin de calculer la variance, il faudra aussi l'estimer grâce à l'échantillon.

La statistique de test à utiliser suit une loi du χ^2 à $n - 1$ degrés de libertés :

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

où
 σ^2 = variance supposée de la population
 s^2 = variance de l'échantillon
 n = taille de l'échantillon

Hypothèse : échantillon provenant d'une population dont les éléments sont **indépendants et identiquement distribués, de distribution normale**.

Remarque: Le nombre de degrés de liberté n'est pas n mais $n - 1$ car la véritable moyenne (de la population) n'est pas connue. Il faut donc l'estimer par la moyenne de l'échantillon.

Exemple:

Le système de satellites Galileo est un projet européen prévu pour concurrencer le système GPS américain actuel. Supposons qu'un constructeur de GPS fournisse un appareil utilisant Galileo, et annonce une précision de ± 5 cm. Un test est effectué pour vérifier cette affirmation, en utilisant un échantillon de taille 20, et un niveau de signification de 0.05. Les données de l'échantillon suivent une loi normale, et l'on peut supposer qu'il en est de même pour la population. La variance de l'échantillon est calculée $s^2 = 0.0108$.

1. La valeur de la population d'intérêt est la précision de l'appareil, qui est une variance.
2. $H_0 : \sigma^2 \leq 0.0025$
 $H_1 : \sigma^2 > 0.0025$
3. Le niveau de signification est fixé à $\alpha = 0.05$
4. La région de rejet du test est l'ensemble des valeurs supérieures à

$$\chi_{0.05}^2 \approx 30.1435$$

La valeur limite calculée est celle pour une distribution χ^2 à $20 - 1 = 19$ degrés de liberté, et un niveau de signification de 0.05.

5. La statistique associée à l'échantillon est

$$s^2 = 0.0108 \quad \chi^2 = \frac{(20-1)0.0108}{0.0025} = 82.08$$

6. Comme $\chi^2 = 82.08$ appartient à la région critique, l'hypothèse H_0 est rejetée.
7. Il est donc très peu probable que la précision annoncée soit correcte.

4.8 Méthode de la p -valeur

Il existe une autre méthode permettant d'interpréter le résultat de n'importe quel test d'hypothèse selon la même procédure. Cette procédure, dite de la p -valeur, est notamment utilisée lorsque l'on travaille avec un ordinateur.

Définition 4.10 La p -valeur, appelée niveau (ou degré) de signification observé, est la probabilité d'observer l'échantillon réellement utilisé sachant que l'hypothèse nulle H_0 est vraie.

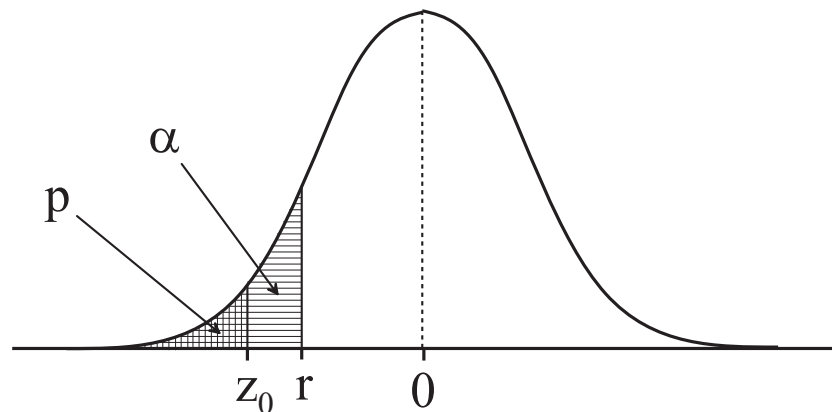
La p -valeur s'interprète aussi comme la probabilité d'obtenir à partir d'un autre échantillon tiré de la même population une valeur du paramètre testé au moins aussi extrême (plus éloignée de H_0) que la valeur réellement observée.

Voici comment interpréter cette p -valeur :

Si la p -valeur calculée est plus petite que la probabilité α associée à la région de rejet, alors l'hypothèse nulle H_0 est rejetée. Sinon, l'hypothèse nulle H_0 n'est pas rejetée. Cette méthode est populaire car la plupart des logiciels, y-compris Excel, Calc, Gnumeric, ..., permettent de calculer cet p -valeur. L'avantage de rapporter cette p -valeur est qu'elle donne plus d'information que simplement le rejet ou non d'une hypothèse. En effet, elle donne le degré de signification associé au résultat.

Exemple:

Dans le cas d'un test unilatéral à gauche, la situation décrite par le graphique suivant conduit au rejet de H_0 , car la p -valeur (zone hachurée verticalement) est plus petite que le risque α (zone hachurée horizontalement).



4.8.1 Procédure de test statistique utilisant la p -valeur

Pour tester une hypothèse en utilisant la p -valeur, vous pouvez procéder ainsi :

1. Spécifier la valeur de la population d'intérêt.
2. Formuler l'hypothèse nulle H_0 et l'hypothèse alternative H_1
3. Choisir le niveau de signification α
4. Déterminer la région critique :
Si la p -valeur est inférieure à α , alors rejeter H_0
Sinon, ne pas rejeter H_0
5. Calculer la p -valeur associée à l'échantillon.
6. Prendre une décision.
(Rejeter H_0 seulement si la p -valeur est inférieure à α)
7. Énoncer une conclusion.

Exemple:

Test unilatéral

Reprenons l'exemple 4.2.

Une étude sur la mobilité des employés de l'entreprise "Jeux Mendors" affirme que la moyenne des temps de trajet des employés pour venir à leur travail excède 40 minutes. Vous devez tester cette affirmation avec un niveau de signification de 0.05. Pour cela, 100 employés vous ont indiqué leur temps de trajet actuel, dont la moyenne est 43.5 minutes. Basé sur une étude précédente, vous pouvez supposer que l'écart type de la population est de 8 minutes.

1. La valeur de la population d'intérêt est le temps de trajet moyen.
2. Les hypothèses nulle et alternative sont :
 $H_0 \quad \mu \leq 40 \text{ minutes}$
 $H_1 \quad \mu > 40 \text{ minutes}$
3. Le niveau de signification est de 0.05
4. La région critique est celle dont la p -valeur est inférieure à α
5. Le calcul de la p -valeur s'effectue de la manière suivante :

$$z = \frac{43.5 - 40}{\frac{8}{\sqrt{100}}} = 4.38$$

Il faut ensuite calculer la probabilité suivante :

$$p\text{-valeur} = P(z > 4.38) = 1 - P(z \leq 4.38) \approx 0$$

6. Comme la p -valeur appartient à la région critique, l'hypothèse H_0 est rejetée.
7. La conclusion est que la moyenne des temps de trajet excède 40 minutes.

Lors d'un test bilatéral, on ne peut se trouver que dans l'une ou l'autre région de rejet du test, et non pas dans les deux à la fois. Considérons une statistique suivant une distribution symétrique. Afin de calculer la p -valeur, on commence par calculer la probabilité que la statistique prenne des valeurs plus extrêmes que celles obtenue à partir de l'échantillon, du côté correspondant à la situation observée. Puis, cette probabilité est doublée pour obtenir la p -valeur.

Exemple:

Test bilatéral

Reprenons l'exemple 4.4

Nous voulons tester la moyenne d'un échantillon du nombre d'employés par petite entreprise suisse contre la moyenne de la population, $\mu = 35$, en ne connaissant pas la valeur de la variance de la population. Par des études précédentes, nous savons que le nombre d'employés est distribué selon une loi normale. Pour cela, 20 petites entreprises suisses ont été interrogées sur le nombre d'employés qu'elles occupent. La moyenne calculée vaut $\bar{x} = 27$ et la variance calculée vaut $s^2 = 334.7$. Nous souhaitons un niveau de signification de 0.05.

1. La valeur de la population d'intérêt est le nombre moyen d'employés de petites entreprises suisses.
2. Les hypothèses nulle et alternative sont :
 $H_0 \quad \mu = 35$ employés
 $H_1 \quad \mu \neq 35$ employés
3. Le niveau de signification est de 0.05
4. La région critique est celle dont la p -valeur est inférieure à α
5. Le calcul de la p -valeur s'effectue de la manière suivante :

$$t = \frac{27 - 35}{\sqrt{\frac{334.7}{20}}} = -1.955587$$

Comme la moyenne de l'échantillon est inférieure à la moyenne de la population postulée sous H_0 , on considère tout d'abord la région de rejet de gauche. Il faut donc calculer la probabilité suivante :

$$P(t_{\frac{\alpha}{2}, 19} < -1.955587) \approx 0.0327$$

S'agissant d'un test bilatéral, cette probabilité doit être doublée

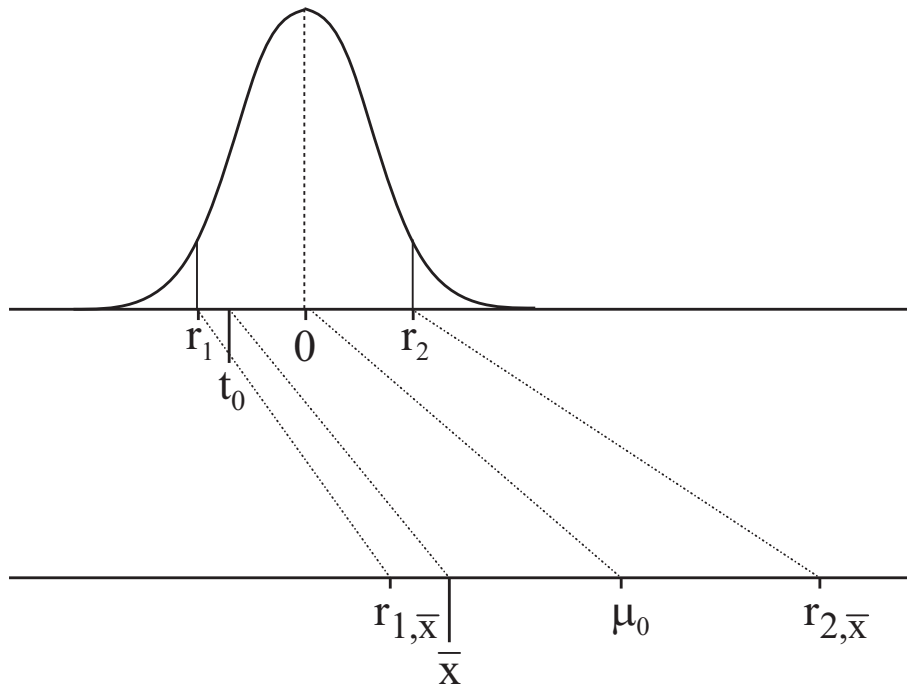
$$p\text{-value} = 2 \times P(t_{\frac{\alpha}{2}, 19} < -1.955587) \approx 0.0654$$

6. Comme la p -valeur n'appartient pas à la région critique, l'hypothèse H_0 n'est pas rejetée.
7. La conclusion est que la vraie moyenne de la population peut être égale à 35 employés par petite entreprise suisse.

4.9 Relation entre IC et tests statistiques

4.9.1 Test dans l'unité du paramètre

Dans l'exemple précédent 4.8.1, nous avons transformé la valeur observée $\bar{x} = 27$ en une valeur sans unité $t_0 = -1.955587$ pouvant être comparée avec les seuils de la loi de Student, en conservant un niveau de signification $\alpha = 0.05$. De façon équivalente, il est également possible d'effectuer le test en comparant la valeur observée $\bar{x} = 27$ avec des transformations $r_{1,\bar{x}}$ et $r_{2,\bar{x}}$ des seuils de la loi de Student.



Étant donné que nous avons utilisé la statistique

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$$

nous pouvons effectuer la transformation suivante

$$\bar{x} = \mu + t \hat{\sigma}_{\bar{x}}$$

et les seuils de rejet dans l'unité de mesure de l'échantillon s'écrivent

$$\begin{aligned} r_{1,\bar{x}} &= \mu + t_{1-\frac{\alpha}{2},19} \hat{\sigma}_{\bar{x}} \\ &\approx 35 - 2.093024 \cdot 4.090843 \approx 26.43777 \end{aligned}$$

et

$$\begin{aligned} r_{2,\bar{x}} &= \mu + t_{\frac{\alpha}{2},19} \hat{\sigma}_{\bar{x}} \\ &\approx 35 + 2.093024 \cdot 4.090843 \approx 43.56223 \end{aligned}$$

Comme $\bar{x} = 27$ se trouve entre les deux seuils, l'hypothèse H_0 est acceptée.

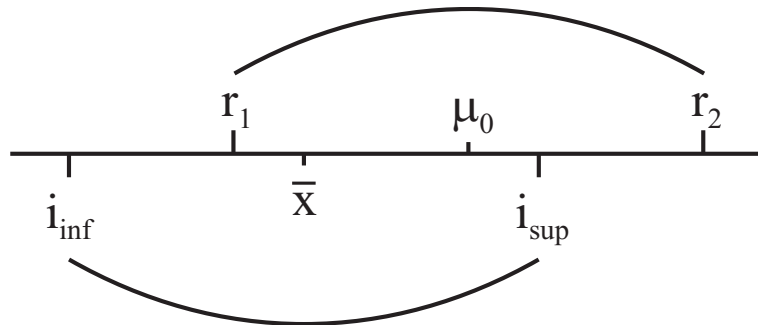
4.9.2 Intervalle de confiance et test statistique

Reprenons encore l'exemple précédent avec variance inconnue et calculons un intervalle de confiance pour μ en gardant un risque $\alpha = 0.05$.

$$\begin{aligned}\mu &\in \bar{x} \pm \hat{\sigma}_{\bar{x}} t_{1-\frac{\alpha}{2}, n-1} \\ &\approx 27 \pm 4.09 \cdot 2.093 \\ &= 27 \pm 8.56\end{aligned}$$

Et donc $P(\mu \in [18.44, 35.56]) = 95\%$

La largeur de l'intervalle de confiance construit autour de \bar{x} est identique à la distance séparant les deux seuils de rejet du test.



Nous pouvons déduire de la figure ci-dessus que lorsque le test bilatéral est accepté,

$$r_1 < \bar{x} < r_2$$

ce qui implique

$$i_{\inf} < \mu_0 < i_{\sup}$$

Il y a donc équivalence entre le test d'hypothèse bilatéral et l'intervalle de confiance :

Si μ_0 se trouve dans l'intervalle de confiance construit autour de \bar{x} , alors l'hypothèse $H_0 : \mu = \mu_0$ est acceptée et vice versa.

4.10 Résumé

Le tableau suivant résume les distributions utilisées pour représenter la moyenne, la proportion et la variance d'une population.

estimé	hypothèse	distribution
μ	σ^2 connu, distr. normale ou $n \geq 30$ σ^2 inconnu, distr. normale	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ $T_{(n-1)} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \mathcal{T}_{n-1}$
π	$n\pi \geq 5$ et $n(1 - \pi) \geq 5$	$Z = \frac{\bar{P} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim \mathcal{N}(0, 1)$
σ^2	μ connu, distr. normale μ inconnu, distr. normale	$Q_{(n)} = \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$ $Q_{(n-1)} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$

Chapitre 5

Tests non paramétriques

Lorsque nous voulons tester des hypothèses, nous devons tenir compte des conditions sous lesquelles les tests peuvent être faits. Par exemple, si nous désirons tester si la moyenne d'une population s'écarte significativement d'une valeur donnée, et que sa distribution ne suit pas une loi normale, nous ne pouvons pas utiliser le test de la moyenne décrit en 4.4. Beaucoup de populations, comme par exemple le revenu moyen par famille, ont une distribution fortement asymétrique. Dans ce cas, l'emploi d'une procédure de test non paramétrique est nécessaire. Ces dernières sont moins restrictives que les procédures de test paramétrique.

Des tests du χ^2 peuvent aussi se baser sur les différences entre des effectifs théoriques et des effectifs observés. Dans ce cas, ces tests ne comportent que très peu de restrictions quant à la distribution sous-jacente, c'est pourquoi ils sont souvent classés parmi les tests non paramétriques.

5.1 Test des rangs signés de Wilcoxon

Ce test non paramétrique n'est pas soumis à une restriction sur une forme particulière de distribution. Il est utilisé pour tester la valeur hypothétique de la médiane d'une population. Il se base sur le principe suivant :

Puisque la médiane sépare une population en deux parties de même taille, la médiane hypothétique de la population sera rejetée si les données se répartissent trop fortement au-dessus ou au-dessous de cette valeur.

5.1.1 Échantillon de petite taille

La statistique de test W suit la loi tabulée pour le test des rangs signés de Wilcoxon en annexe A.4.

1. Calculer les différences d_i entre chaque valeur et la médiane postulée $\tilde{\mu}$

2. Calculer la valeur absolue des différences précédentes : $|d_i|$
3. Déterminer le rang pour chacune des valeurs $|d_i|$, en ne tenant pas compte des valeurs nulles.
Si des observations ont la même valeur $|d_i|$, alors affecter le rang moyen de ces observations.
4. Calculer la statistique W qui est la somme des rangs dont les d_i sont positifs.

L'hypothèse nulle est rejetée si la valeur de la statistique est supérieure à celle de la valeur critique associée.

Exemple:

Une étude a été faite sur les salaires en début de carrière des diplômés de l'école "Jess Aitou", afin de savoir si le salaire médian est supérieur à 35000 Euros, à un seuil $\alpha = 0.05$. Les données récoltées sont les salaires en euros :

36400
38500
27000
35000
29000
40000
52000
34000
38900
41000

Nous pouvons utiliser le test des rangs signés de Wilcoxon afin de répondre à l'objet de l'étude.

1. La valeur de la population d'intérêt est le salaire médian en début de carrière.
2. Les hypothèses nulle et alternative sont :
 $H_0 \quad \tilde{\mu} \leq 35'000 \text{ euros}$
 $H_1 \quad \tilde{\mu} > 35'000 \text{ euros}$
3. Le niveau de signification est de 0.05

4.

Salaire = x_i	$d_i = x_i - \tilde{\mu}$	$ d_i $	Rang	R_+	R_-
36400	1400	1400	2	2	
38500	3500	3500	3	3	
27000	-8000	8000	8		8
35000	0	0			
29000	-6000	6000	6.5		6.5
40000	5000	5000	5	5	
52000	17000	17000	9	9	
34000	-1000	1000	1		1
38900	3900	3900	4	4	
41000	6000	6000	6.5	6.5	
				$W = 29.5$	15.5

Soit n le nombre de d_i différents de 0. Il y a donc $n = 9$ rangs attribués.

La valeur critique du test est $W_\alpha = 37$

5. La statistique est $W = 29.5$

6. Comme

$$W = 29.5 < 37$$

l'hypothèse H_0 ne peut pas être rejetée.

7. Nous ne pouvons donc pas conclure que le salaire médian en début de carrière dépasse 35000 euros.

5.1.2 Échantillon de grande taille

Lorsque la taille de l'échantillon est suffisamment grande, en pratique, lorsque $n > 20$, alors la statistique de test W peut être approchée par une distribution normale. La statistique de test est transformée en une z -valeur, qui suit une loi normale centrée réduite.

$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

où

n = taille de l'échantillon

W = somme des rangs positifs R_+

5.2 Test d'ajustement

Certains tests exigent que la distribution de la population suive une loi spécifique, comme la loi normale. Comment peut-on vérifier que cette condition est satisfaite ? Le test d'ajustement du χ^2 permet justement de le faire. Le test d'ajustement du χ^2 peut être utilisé pour déterminer si un échantillon provient d'une hypothétique distribution.

La procédure de test débute par l'acquisition d'un échantillon de taille suffisamment grande. Les données sont ensuite classées en k différentes catégories, afin d'obtenir les fréquences absolues observées o_i , associées à chaque catégorie. En utilisant la distribution hypothétique, les fréquences absolues théoriques e_i sont calculées. Il suffit alors de comparer les fréquences observées à celles théoriques. Si une trop grande différence existe, alors l'hypothèse d'un échantillon tiré de la distribution hypothétique est rejetée.

La statistique à utiliser suit une loi du χ^2 à $k - 1$ degrés de liberté et est calculée ainsi :

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

où

o_i = fréquence observée pour la catégorie i

e_i = fréquence théorique pour la catégorie i

k = nombre de catégories

Toutefois, afin de pouvoir utiliser ce test, nous devons faire la **supposition suivante** :

LA TAILLE DE L'ÉCHANTILLON EST SUFFISAMMENT GRANDE

En pratique, un échantillon de taille au moins 30 est suffisant, pour autant que les fréquences observées par cellule sont suffisamment élevées.

Exemple:

Un nouveau directeur d'un centre d'appels pour FAI constate que le personnel est réduit de 20% les mercredi, jeudi et dimanche. Son prédécesseur avait procédé ainsi car le nombre d'appels était 20% moins élevé ces jours-là. Afin de savoir si tel est toujours le cas, il fait relever le nombre d'appels sur 1 mois pour chaque jour de la semaine et obtient les données agrégées suivantes :

Jours	Lu	Ma	Me	Je	Ve	Sa	Di
Nombre d'appels	1000	1200	900	1000	1200	1100	800

Il souhaite savoir s'il y a effectivement une baisse de 20% les mercredi, jeudi et dimanche, avec un niveau de signification de 0.05.

1. Les hypothèses nulle et alternative sont :

H_0 La distribution des appels est identique les lu, ma, ve et sa,
et 20% moins élevée les me, je et di.

H_1 La distribution des appels n'est pas celle décrite en H_0

2. Le niveau de signification est de 0.05

3. La valeur critique est celle d'une distribution du χ^2 à 6 degrés de libertés.

$$\chi_{0.05}^2 = 12.5916$$

4. Le nombre total d'appels sur la période observée est 7200

Jours	L	Ma	Me	Je	Ve	Sa	Di
o_i	1000	1200	900	1000	1200	1100	800
e_i	1125	1125	900	900	1125	1125	900

La statistique est alors

$$\chi^2 = 46.6$$

5. Comme $46.6 > 12.5916$, l'hypothèse H_0 est rejetée.
 6. La conclusion est que la distribution du nombre d'appels n'est plus telle qu'indiquée par le précédent directeur.

5.3 Test d'indépendance de deux variables catégorielles

Afin de tester l'indépendance de deux variables de type catégoriel, nous allons utiliser une table de contingence. Nous comparons alors les fréquences observées avec les fréquences théoriques en cas d'indépendance. Si une trop grande différence existe, alors l'hypothèse d'indépendance des variables est rejetée.

Les fréquences théoriques sont calculées ainsi :

$$e_{ij} = \frac{(\text{Total ligne } i) \cdot (\text{Total colonne } j)}{\text{Taille de l'échantillon}}$$

La statistique à utiliser est celle du χ^2 à $dl = (c - 1)(r - 1)$ degrés de liberté

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

où

- o_i = fréquence observée de la cellule (i, j)
- e_i = fréquence théorique de la cellule (i, j)
- r = nombre de lignes
- c = nombre de colonnes

Toutefois, afin de pouvoir utiliser ce test, nous devons faire la **supposition suivante** :

LA TAILLE DE L'ÉCHANTILLON EST SUFFISAMMENT GRANDE

En pratique, l'effectif observé par cellule doit être d'au moins 5 unités.

Exemple:

Une recherche est effectuée afin de savoir si le nombre de sorties le week-end est indépendant des résultats aux examens, avec un niveau de signification de 0.05. La table de contingence suivante résume les données récoltées :

5.3. TEST D'INDÉPENDANCE DE DEUX VARIABLES CATÉGORIELLES

	Résultat				<i>Total</i>
	Insuffisant	Acquis	Bien	Excellent	
Sortie WE					
Jamais	84	50	50	16	200
Occasionnel	82	64	34	20	200
Fréquent	34	36	16	14	100
<i>Total</i>	200	150	100	50	500

1. Les hypothèses nulle et alternative sont :
 H_0 Les sorties du week-end sont indépendantes des résultats aux examens.
 H_1 les sorties du week-end ne sont PAS indépendantes des résultats aux examens
2. Le niveau de signification est de 0.05
3. La valeur critique est celle d'une distribution du χ^2 à $(3 - 1)(4 - 1) = 6$ degrés de libertés.

$$\chi_{0.05}^2 = 12.5916$$

4. Les fréquences théoriques de la table de contingence sont

	Résultat				<i>Total</i>
	Insuffisant	Acquis	Bien	Excellent	
Sortie WE					
Jamais	80	60	40	20	200
Occasionnel	80	60	40	20	200
Fréquent	40	30	20	10	100
<i>Total</i>	200	150	100	50	500

La statistique est alors

$$\chi^2 = 10.88$$

5. Comme $10.88 < 12.5916$, l'hypothèse H_0 ne peut pas être rejetée.
6. La conclusion est qu'il n'y a pas suffisamment d'évidence pour conclure que les sorties du week-end et le résultat aux examens ne sont pas indépendants.
... et ceci n'est aucunement une incitation à faire la fête le week-end !

Chapitre 6

Régression linéaire simple

6.1 Rappel

Définition 6.1 *L'analyse statistique de la relation entre deux ou plusieurs variables s'appelle l'analyse de régression. Si seulement deux variables sont étudiées, il s'agit d'une analyse de régression simple. Lorsque la relation entre deux variables est linéaire, elle porte le nom de régression linéaire simple.*

Définition 6.2 *Le modèle de régression linéaire simple est défini par l'équation suivante :*

$$y = \beta_0 + \beta_1 x + \epsilon$$

avec

- y = variable dépendante (ou variable expliquée)
- x = variable indépendante (ou variable explicative)
- β_0 = constante de la droite de régression pour la population
- β_1 = pente de la droite de régression pour la population
- ϵ = terme d'erreur ou résidu.

Le modèle de régression linéaire simple se base sur les postulats suivants :

1. Les résidus ϵ sont indépendants et identiquement distribués, suivant une loi normale.
2. Le modèle de régression linéaire est légitime.

Ce modèle est estimé par la méthode des moindres carrés : la somme des carrés des résidus est minimisée, ce qui permet d'obtenir l'équation suivante :

$$\hat{y} = b_0 + b_1 x$$

avec

$$\begin{aligned}\hat{y} &= \text{valeur estimée de } y \\ x &= \text{valeur de la variable indépendante} \\ b_1 &= \frac{s_{xy}}{s_x^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ b_0 &= \bar{y} - b_1\bar{x}\end{aligned}$$

Définition 6.3 *Le coefficient de corrélation linéaire d'un échantillon est la valeur*

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

6.2 Validité d'une corrélation

Une corrélation linéaire r étant le plus souvent calculée à partir d'un échantillon, sa valeur est sujette à des erreurs d'échantillonnage. Ainsi, r_{xy} n'est qu'une estimation de la véritable valeur du coefficient de corrélation linéaire ρ . Il faut donc utiliser un test sur l'existence ou non d'une corrélation linéaire :

$$\begin{aligned}H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0\end{aligned}$$

La statistique de test à considérer est la suivante :

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad dl = n - 2$$

avec

$$\begin{aligned}t &= \text{nombre d'écart-type de } r \text{ depuis } 0 \\ r &= \text{coefficient de corrélation linéaire} \\ n &= \text{taille de l'échantillon}\end{aligned}$$

Notons que ce test de Student suppose que

1. Les données sont quantitatives.
2. Les deux variables x et y suivent une distribution bivariée normale (i.e. leur distribution conjointe suit une loi normale).

Exemple:

Une entreprise souhaite analyser la relation entre la taille d'une annonce publicitaire, et le nombre d'appels reçus générés par l'annonce. Elle veut savoir s'il existe une corrélation linéaire positive entre ces deux variables, au seuil 0.05. Pour cela, elle demande à ses clients d'indiquer quelle annonce leur ont fait connaître l'entreprise. Le dépouillement de l'enquête donne :

Taille [cm carrés]	90	160	250	160	200	160	200	200	160	90
Prop. d'appels	0.13	0.16	0.21	0.18	0.18	0.19	0.15	0.17	0.13	0.11

1. Le paramètre d'intérêt est la corrélation linéaire ρ entre la taille d'une annonce publicitaire et la proportion d'appels générés par l'annonce. L'entreprise
2. L'hypothèse nulle et alternative sont :

$$H_0 : \rho \leq 0$$

$$H_1 : \rho > 0$$

3. Le niveau de signification choisi est $\alpha = 0.05$
4. La p -valeur associée est 0.003921
5. Comme la p -valeur est inférieure au niveau de signification, l'hypothèse nulle est rejetée
6. L'échantillon supporte la possibilité d'une relation linéaire positive entre la taille d'une annonce publicitaire et la proportion d'appels générés par l'annonce.

6.3 Qualité d'un modèle linéaire

Définition 6.4 La somme des carrés totale ($SST = \text{Total Sum of Squares}$) est la valeur

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

avec

n = taille de l'échantillon

y_i = i ème valeur de la variable dépendante

\bar{y} = moyenne de la variable dépendante

La somme des carrés des erreurs ($SSE = \text{Sum of Squares Errors}$) est la valeur

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

avec

n = taille de l'échantillon

y_i = i ème valeur de la variable dépendante

\hat{y}_i = i ème valeur prédite

La somme des carrés de régression ($SSR = \text{Sum of Squares Regression}$) est la valeur

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

avec

n = taille de l'échantillon

\hat{y}_i = i ème valeur prédite

\bar{y} = moyenne de la variable dépendante

Propriété 6.1 La variance totale se décompose en une partie expliquée et une partie non-expliquée (ou résiduelle)

$$SST = SSE + SSR$$

Cette propriété est à la base de l'analyse de variance, utilisée pour tester si plusieurs populations sont significativement différentes les unes des autres.

Propriété 6.2 Soit une droite de régression linéaire, basée sur la minimisation de la somme des carrés.

– La somme des résidus est nulle.

$$\sum_{i=1}^n (y_i - \hat{y}_i) = 0$$

– La somme des carrés des résidus ($SSE = \text{Sum of Squares Errors}$) est minimale.

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

– La droite de régression simple linéaire passe par le point (\bar{x}, \bar{y})

– Les coefficients estimés b_0 et b_1 de β_0 et β_1 sont des estimateurs sans biais. (Rappel : un estimateur d'un paramètre est dit sans biais si son espérance est égale au paramètre.)

Définition 6.5 Le coefficient de détermination R^2 est la proportion de variation totale dans la variable dépendante qui est expliquée par sa relation avec la variable dépendante.

$$R^2 = \frac{SSR}{SST}$$

Le coefficient de détermination est une mesure de la qualité d'un modèle de régression linéaire. R^2 varie entre 0 et 1. Plus il se rapproche de 1, meilleur est le modèle. En pratique, des valeurs supérieures ou égales à 0.7 indiquent que le modèle est satisfaisant.

Exemple:

L'entreprise souhaitant analyser la relation entre la taille d'une annonce publicitaire, et le nombre d'appels reçus générés par l'annonce, est parvenue à la conclusion qu'une relation linéaire existe probablement.

Taille [cm carrés]	90	160	250	160	200	160	200	200	160	90
Prop. d'appels	0.13	0.16	0.21	0.18	0.18	0.19	0.15	0.17	0.13	0.11

Le coefficient de détermination vaut $R^2 = 0.7795506$. Ainsi, environ 78% de la variance du nombre d'appels est expliquée par la taille de l'annonce.

Remarque:

1. Dans le cas d'une seule variable indépendante, le coefficient de détermination est égal au carré de la valeur du coefficient de corrélation linéaire de Pearson : $R^2 = r^2$
2. Lorsque le modèle comporte au moins 2 variables indépendantes (= explicatives), il est préférable d'utiliser le "coefficient de détermination ajusté" plutôt que R^2 . En effet, R^2 a tendance à augmenter automatiquement avec l'augmentation du nombre de variables indépendantes.

6.4 Validité d'un modèle linéaire

Pour tout modèle, il est nécessaire d'en vérifier sa validité. Autrement dit, il est nécessaire de connaître si le modèle est statistiquement significatif. Pour le modèle linéaire simple, en supposant valides les postulats sur les résidus (i.i.d. et suivant une loi normale), il existe deux méthodes de test équivalentes :

1. Test de signification de la corrélation entre x et y
2. Test de signification du coefficient de la pente de régression

Le test de signification de la corrélation a déjà été présenté. Le test de signification du coefficient de la pente de régression utilise les hypothèses nulle et alternatives

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

L'estimation de l'écart type de la pente de régression est donnée par

$$s_{b_1} = \frac{s_\epsilon}{\sqrt{(x - \bar{x})^2}}$$

avec

$$s_\epsilon = \sqrt{\frac{SSE}{n-2}}$$

La variable de test à utiliser est

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \quad dl = n - 2$$

avec

β_1 = pente supposée de la droite de régression

b_1 = pente calculée de la droite de régression

s_{b_1} = estimation de l'écart type de la pente de régression

6.5 Analyse d'une régression linéaire simple

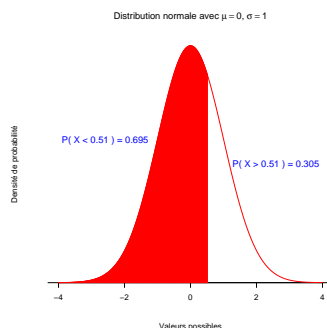
Les étapes suivantes montrent comment effectuer une analyse de régression linéaire simple :

1. Définir la variable dépendante y et la variable indépendante x .
2. Dessiner un diagramme de dispersion sur x et y afin de vérifier visuellement si une relation linéaire est probable.
3. Calculer le coefficient de corrélation r_{xy} .
4. Calculer les coefficients de la régression, ainsi que le coefficient de détermination R^2 .
5. Effectuer un test statistique pour valider ou invalider le modèle linéaire simple.
6. Établir une conclusion.

Annexe A

Tables statistiques

A.1 Loi normale centrée réduite



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

A.2 Table de la loi du χ^2

	Valeurs de α									
dl	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.7055	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.6052	5.9915	7.3778	9.2103	10.5966
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.2514	7.8147	9.3484	11.3449	12.8382
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.7794	9.4877	11.1433	13.2767	14.8603
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.2364	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.6446	12.5916	14.4494	16.8119	18.5476
7	0.9893	1.2390	1.6899	2.1673	2.8331	12.0170	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.3616	15.5073	17.5345	20.0902	21.9550
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.6837	16.9190	19.0228	21.6660	23.5894
10	2.1559	2.5582	3.2470	3.9403	4.8652	15.9872	18.3070	20.4832	23.2093	25.1882
11	2.6032	3.0535	3.8157	4.5748	5.5778	17.2750	19.6751	21.9200	24.7250	26.7568
12	3.0738	3.5706	4.4038	5.2260	6.3038	18.5493	21.0261	23.3367	26.2170	28.2995
13	3.5650	4.1069	5.0088	5.8919	7.0415	19.8119	22.3620	24.7356	27.6882	29.8195
14	4.0747	4.6604	5.6287	6.5706	7.7895	21.0641	23.6848	26.1189	29.1412	31.3193
15	4.6009	5.2293	6.2621	7.2609	8.5468	22.3071	24.9958	27.4884	30.5779	32.8013
16	5.1422	5.8122	6.9077	7.9616	9.3122	23.5418	26.2962	28.8454	31.9999	34.2672
17	5.6972	6.4078	7.5642	8.6718	10.0852	24.7690	27.5871	30.1910	33.4087	35.7185
18	6.2648	7.0149	8.2307	9.3905	10.8649	25.9894	28.8693	31.5264	34.8053	37.1565
19	6.8440	7.6327	8.9065	10.1170	11.6509	27.2036	30.1435	32.8523	36.1909	38.5823
20	7.4338	8.2604	9.5908	10.8508	12.4426	28.4120	31.4104	34.1696	37.5662	39.9968
21	8.0337	8.8972	10.2829	11.5913	13.2396	29.6151	32.6706	35.4789	38.9322	41.4011
22	8.6427	9.5425	10.9823	12.3380	14.0415	30.8133	33.9244	36.7807	40.2894	42.7957
23	9.2604	10.1957	11.6886	13.0905	14.8480	32.0069	35.1725	38.0756	41.6384	44.1813
24	9.8862	10.8564	12.4012	13.8484	15.6587	33.1962	36.4150	39.3641	42.9798	45.5585
25	10.5197	11.5240	13.1197	14.6114	16.4734	34.3816	37.6525	40.6465	44.3141	46.9279
26	11.1602	12.1981	13.8439	15.3792	17.2919	35.5632	38.8851	41.9232	45.6417	48.2899
27	11.8076	12.8785	14.5734	16.1514	18.1139	36.7412	40.1133	43.1945	46.9629	49.6449
28	12.4613	13.5647	15.3079	16.9279	18.9392	37.9159	41.3371	44.4608	48.2782	50.9934
29	13.1211	14.2565	16.0471	17.7084	19.7677	39.0875	42.5570	45.7223	49.5879	52.3356
30	13.7867	14.9535	16.7908	18.4927	20.5992	40.2560	43.7730	46.9792	50.8922	53.6720
40	20.7065	22.1643	24.4330	26.5093	29.0505	51.8051	55.7585	59.3417	63.6907	66.7660
50	27.9907	29.7067	32.3574	34.7643	37.6886	63.1671	67.5048	71.4202	76.1539	79.4900
60	35.5345	37.4849	40.4817	43.1880	46.4589	74.3970	79.0819	83.2977	88.3794	91.9517
70	43.2752	45.4417	48.7576	51.7393	55.3289	85.5270	90.5312	95.0232	100.4252	104.2149
80	51.1719	53.5401	57.1532	60.3915	64.2778	96.5782	101.8795	106.6286	112.3288	116.3211
90	59.1963	61.7541	65.6466	69.1260	73.2911	107.5650	113.1453	118.1359	124.1163	128.2989
100	67.3276	70.0649	74.2219	77.9295	82.3581	118.4980	124.3421	129.5612	135.8067	140.1695

A.3 Table de la loi de Student

t	Valeurs de α									
dl	0.45	0.4	0.3	0.25	0.2	0.1	0.05	0.025	0.01	0.005
1	0.1584	0.3249	0.7265	1.0000	1.3764	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.1421	0.2887	0.6172	0.8165	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.1366	0.2767	0.5844	0.7649	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.1338	0.2707	0.5686	0.7407	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.1322	0.2672	0.5594	0.7267	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.1311	0.2648	0.5534	0.7176	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.1303	0.2632	0.5491	0.7111	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.1297	0.2619	0.5459	0.7064	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.1293	0.2610	0.5435	0.7027	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.1289	0.2602	0.5415	0.6998	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.1286	0.2596	0.5399	0.6974	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.1283	0.2590	0.5386	0.6955	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.1281	0.2586	0.5375	0.6938	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.1280	0.2582	0.5366	0.6924	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.1278	0.2579	0.5357	0.6912	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467
16	0.1277	0.2576	0.5350	0.6901	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.1276	0.2573	0.5344	0.6892	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.1274	0.2571	0.5338	0.6884	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.1274	0.2569	0.5333	0.6876	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.1273	0.2567	0.5329	0.6870	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.1272	0.2566	0.5325	0.6864	0.8591	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.1271	0.2564	0.5321	0.6858	0.8583	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.1271	0.2563	0.5317	0.6853	0.8575	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.1270	0.2562	0.5314	0.6848	0.8569	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.1269	0.2561	0.5312	0.6844	0.8562	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.1269	0.2560	0.5309	0.6840	0.8557	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.1268	0.2559	0.5306	0.6837	0.8551	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.1268	0.2558	0.5304	0.6834	0.8546	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.1268	0.2557	0.5302	0.6830	0.8542	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.1267	0.2556	0.5300	0.6828	0.8538	1.3104	1.6973	2.0423	2.4573	2.7500
40	0.1265	0.2550	0.5286	0.6807	0.8507	1.3031	1.6839	2.0211	2.4233	2.7045
50	0.1263	0.2547	0.5278	0.6794	0.8489	1.2987	1.6759	2.0086	2.4033	2.6778
60	0.1262	0.2545	0.5272	0.6786	0.8477	1.2958	1.6706	2.0003	2.3901	2.6603
70	0.1261	0.2543	0.5268	0.6780	0.8468	1.2938	1.6669	1.9944	2.3808	2.6479
80	0.1261	0.2542	0.5265	0.6776	0.8461	1.2922	1.6641	1.9901	2.3739	2.6387
90	0.1260	0.2541	0.5263	0.6772	0.8456	1.2910	1.6620	1.9867	2.3685	2.6316
100	0.1260	0.2540	0.5261	0.6770	0.8452	1.2901	1.6602	1.9840	2.3642	2.6259
200	0.1258	0.2537	0.5252	0.6757	0.8434	1.2858	1.6525	1.9719	2.3451	2.6006
500	0.1257	0.2535	0.5247	0.6750	0.8423	1.2832	1.6479	1.9647	2.3338	2.5857
∞	cf. Distribution Normale									

A.4 Table du test des rangs signés de Wilcoxon

Les valeurs critiques sont données par la table suivante :

	unilatéral	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
	bilatéral	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.02$
n	Inférieur, Supérieur			
5	0,15			
6	2,19 0,21			
7	3,25 2,26 0,28			
8	5,31 3,33 1,35			
9	8,37 5,40 3,42			
10	10,45 8,47 5,50			
11	13,53 10,56 7,59			
12	17,61 13,65 10,68			
13	21,70 17,74 12,79			
14	25,80 21,84 16,89			
15	30,90 25,95 19,101			
16	35,101 29,107 23,113			
17	41,112 34,119 27,126			
18	47,124 40,131 32,139			
19	53,137 46,144 37,153			
20	60,150 52,158 43,167			

Annexe B

Instructions pour logiciels

Voici quelques instructions pour quelques logiciels disponibles sur les ordinateurs de la HEG Genève.

B.1 Loi normale

Logiciel	Version	
Calc d'OpenOffice	2.4	$z_\alpha = \text{LOI.NORMALE.STANDARD.INVERSE}()$ $p = \text{LOI.NORMALE}()$
MS Excel	2007	$z_\alpha = \text{LOI.NORMALE.STANDARD.INVERSE}()$ $p = \text{LOI.NORMALE}()$
R	2.9.1	$z_\alpha = \text{qnorm}()$ $p = \text{pnorm}()$

Exemple:

Calculer les valeurs critiques $z_{\alpha/2}$ pour un degré de confiance $1 - \alpha = 0.9$ dans un test bilatéral, i.e. tels que $P(z_{\alpha/2}) = 0.9$

R	$\text{qnorm}(0.1/2) ; \text{qnorm}(1-0.1/2)$
Ms Excel	$=\text{LOI.NORMALE.STANDARD.INVERSE}(0.1/2)$
Calc	$= \text{LOI.NORMALE.STANDARD.INVERSE}(0.1/2)$

B.2 t -distribution

Logiciel	Version	
Calc d'OpenOffice	2.4	$t_\alpha = \text{LOI.STUDENT.INVERSE}()$ $p = \text{LOI.STUDENT}()$
MS Excel	2007	$t_\alpha = \text{LOI.STUDENT.INVERSE}()$ $p = \text{LOI.STUDENT}()$
R	2.9.1	$t_\alpha = \text{qt}()$ $p = \text{pt}()$

Exemple:

Calculer les valeurs critiques pour un degré de confiance $1 - \alpha = 0.95$ dans un test unilatéral avec 10 degrés de liberté.

R	<code>qt(0.95, 10)</code>
Ms Excel	<code>=LOI.STUDENT.INVERSE(2*0.05;10)</code>
Calc	<code>=LOI.STUDENT.INVERSE(2*0.05;10)</code>

B.3 χ_n^2 distribution

Logiciel	Version	
Calc d'OpenOffice	2.4	$\chi_{1-\alpha}^2 = \text{KHI DEUX.INVERSE}()$ $p = 1 - \text{LOI.KHI DEUX}()$
MS Excel	2007	$\chi_{1-\alpha}^2 = \text{KHI DEUX.INVERSE}()$ $p = 1 - \text{LOI.KHI DEUX}()$
R	2.9.1	$\chi_\alpha^2 = \text{QCHISQ}()$ $p = \text{pchisq}()$

Exemple:

Calculer la valeur critique pour un degré de confiance $1 - \alpha = 0.9$ dans un test unilatéral avec 19 degrés de liberté.

R	<code>qchisq(0.9,19)</code>
Ms Excel	<code>=KHI DEUX.INVERSE(1-0.9;19)</code>
Calc	<code>=KHI DEUX.INVERSE(0.1;19)</code>

B.4 Intervalle de confiance

La fonction `INTERVALLE.CONFIANCE` d'Excel et Calc d'OpenOffice ne calculent pas un intervalle de confiance, mais la marge d'erreur dans le cas suivant : la population suit une loi normale, et son écart type est connu.

Logiciel	Version	
Calc d'OpenOffice	2.4	$\text{Marge}_z = \text{INTERVALLE.CONFIANCE}()$
MS Excel	2007	$\text{Marge}_z = \text{INTERVALLE.CONFIANCE}()$
R	2.9.1	$IC_t = \text{t.test()}\$conf.int$

Exemple:

Calculer la marge de l'IC pour une moyenne, dont l'écart type connu de la population est 0.04, le degré de confiance vaut 0.95, la taille de l'échantillon est 4 et la population suit une loi normale.

Ms Excel	=INTERVALLE.CONFIANCE(0.05;0.04;4)
Calc	=INTERVALLE.CONFIANCE(0.05;0.04;4)

Reprenons l'exemple en 3.5

R : x <- c(7.1,13.6,1.4,3.6,1.9,11.6,1.7,16.9,2.6,7.7, 12.4,11,3.7,14.6,8.8,8.5,6.1,3.3,6.1,6.9,0.4,11,0.8,6.4,9.1) t.test(x)\$conf.int

B.5 Test d'une moyenne

Logiciel	Version	
Ms Excel	2007	=TEST.STUDENT()
R	2.9.1	t.test library(TeachingDemos); z.test()

Exemple:

Test unilatéral de la moyenne : $H_0 : \mu \leq 5.2$ lorsque la variance est inconnue, pour un échantillon de taille 15, et un seuil de signification de 0.05.

Ms Excel =TEST.STUDENT(X,5.2;1;3) R : x <- c(4,4.1,6,5.5,5.8,4.2,3.1,6,4.9,3.9,4.8,5.6,5.7,5.3,5.8) t.test(x, mu = 5.2, alternative = "greater", correct=FALSE)

Exemple:

Test bilatéral de la moyenne : $H_0 : \mu = 99$ lorsque l'écart type est connu ($\sigma = 5$), avec un seuil de signification de 0.05.

R : x <- rnorm(25, 100, 5) library(TeachingDemos); z.test(x, mu = 99, stdev= 5)

B.6 Test d'une proportion

Logiciel	Version	
R	2.9.1	prop.test

Exemple:

Reprenons l'exemple 4.6, dans lequel 9 contrats sur 600 étaient incomplets. La banque exige qu'il n'y ait pas plus d'1% de contrats incomplets, avec un niveau de signification fixé à 0.02.

R : `prop.test(9, 600, p=0.01, alternative="greater", correct=FALSE)`

B.7 Test d'une variance

Exemple:

Tester si la variance est en dessous de 0.005.

Valeurs : -0.04, 0.11, -0.10, 0.05, 0.20, -0.05, -0.04, 0.13, 0.01, 0.05, -0.05, 0.04, -0.03, 0.11, 0.05, -0.03, -0.21, 0.19, 0.04, -0.14

$H_0 : \sigma^2 \leq 0.005$

$H_1 : \sigma^2 > 0.005$

```
y <- c( -0.04,  0.11, -0.10,  0.05,  0.20, -0.05, -0.04,  0.13,  0.01,  0.05,
        -0.05,  0.04, -0.03,  0.11,  0.05, -0.03, -0.21,  0.19,  0.04, -0.14)
sigmacarre = 0.005
p-valeur   = pchisq( var(y)*(length(y)-1)/sigmacarre, length(y)-1, lower.tail=FALSE)
```

B.8 Test des rangs signés de Wilcoxon

Logiciel	Version	
R	2.9.1	= wilcox.test()

Exemple:

Tester si la valeur de la médiane est de 40.

```
x <- c(31, 48, 23, 56, 28, 29, 44)
wilcox.test(x, alternative = "two.sided", mu=40,
            exact = TRUE, conf.level=0.9, conf.int = TRUE)
```

B.9 Test d'indépendance

Logiciel	Version	
R	2.9.1	<code>= chisq.test()</code>

Exemple:

Déterminer si deux variables sont indépendantes.

```
amendes <- matrix(c(240,160,80,40,32,18,11,9,5,4),ncol=2,byrow=TRUE)
rownames(amendes)<-c("Homme","Femme")
colnames(amendes)<-c("Vitesse","Parcage","Feux grillé","Service AP", "Autre")
amendes <- as.table(amendes)
chisq.test(amendes)
```

B.10 Test de corrélation linéaire

Logiciel	Version	
R	2.9.1	<code>= cor.test()</code>

Exemple:

Déterminer si deux variables sont positivement corrélées (linéairement).

```
taille <- c(90, 160, 250, 160, 200, 160, 200, 200, 160, 90)
proportion <- c(0.13, 0.16, 0.21, 0.18, 0.18, 0.19, 0.15, 0.17, 0.13, 0.11)
cor.test(taille, proportion, method = "pearson", alternative = "greater")
```

B.11 Régression linéaire

Logiciel	Version	
R	2.9.1	<code>= lm()</code>

Exemple:

```
x <- c(3,5,2,8,2,6,7,1,4,2,9,6)
y <- c(487,445,272,641,187,440,346,238,312,269,655,563)
lm(y~x) # coefficients
summary(lm.D90 <- lm(y ~ x)) # avec tests
```