5

IICT – BAS
eISSN: 2367-8666
Lecture Notes in Computer Science and Technologies

Statistique inférentielle

Vera Angelova

eISBN: 978-619-7320-00-8

The series **Lectures Notes in Computer Science and Technologies of the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences** presents in an electronic format textbooks for undergraduate, graduate and PhD students studied various programs related to Informatics, Computational Mathematics, Mathematical Modeling, Communication Technologies, etc., as well as for all readers interested in these scientific disciplines. The Lecture Notes are based on courses taught by scientists of the Institute of Information and Communication Technologies - BAS in various Bulgarian universities and the Center for Doctoral Training in BAS. The published materials are with open access - they are freely available without any charge.

Editorial board

Gennady Agre (Editor-in-Chef), IICT-BAS
e-mail: agre@iinf.bas.bg

Vera Angelova, IICT-BAS
e-mail: vangelova@iit.bas.bg

Pencho Marinov, IICT-BAS
e-mail: pencho@bas.bg

eISSN: 2367-8666

The series is subject to copyright. All rights reserved in translation, printing, using illustrations, citations, distribution, reproduction on microfilm or in other ways, and storage in a database of all or part of the material in the present edition. The copy of the publication or part of the content is permitted only with the consent of the authors and / or editors

Avec la collaboration de madame Viviane Baligand et monsieur François Mimiague - Professeur à l'Université de Bordeaux IV, qui ont posé les basses de l'enseignement en Statistique au programme français de la Faculté de gestion et d'économie à l'Université de Sofia.

Table des matières

1 Echantillonnage - rappel **1**

1.1 Introduction 1

1.2 Les problèmes de distribution d'échantillonnage 4

1.2.1 Distribution d'échantillonnage de la moyenne $\bar{X}$ 4

1.2.2 Distribution de la variance d'échantillon $S'^2_{\bar{X}}$ 10

1.2.3 Distribution d'échantillonnage d'une proportion F 11

1.3 Synthèse sur les distributions d'échantillonnage 14

2 Estimation **16**

2.1 Estimation ponctuelle 18

2.1.1 Qualités d'un estimateur 18

2.1.2 Les estimateurs les plus utilisés 18

2.2 Estimation par intervalle de confiance 24

2.2.1 Intervalle de confiance de la moyenne d'une population : μ 26

2.2.2 Intervalle de confiance de la proportion d'une population : p 30

2.2.3 Précision - Taille d'échantillon - Risque d'erreur 31

2.2.4 Intervalle de confiance de la variance de la population : σ^2 32

2.3 Comparaisons 34

2.3.1 Estimation ponctuelle de la différence de 2 moyennes 34

2.3.2 Intervalle de confiance de la différence de 2 moyennes 35

2.3.3 Différence de 2 proportions 42

2.3.4 Rapport de 2 variances (comparaison de 2 variances) 43

2.3.5 Synthèse sur l'estimation 45

3 Les tests d'hypothèse **48**

3.1 Généralités 48
3.1.1 Principe d'un test d'hypothèses 48
3.1.2 Définition des concepts utiles à l'élaboration des tests d'hypothèse 49
3.2 Tests permettant de déterminer si un échantillon appartient à une population donnée 52
3.2.1 Test sur une moyenne : comparaison d'une moyenne expérimentale à une moyenne théorique dans le cas d'un caractère quantitatif 52
3.2.2 Tests sur une proportion 53
3.3 Risques de première et de deuxième espèce 55
3.3.1 Définitions 55
3.3.2 Schématisation des deux risques d'erreur sur la distribution d'échantillonnage 57
3.3.3 Exemples d'application 61
3.4 Comparaisons. Tests permettant de déterminer si deux échantillons appartiennent à la même population 66
3.4.1 Comparaison de deux moyennes d'échantillon : "test T" 66
3.4.2 Comparaison de deux variances d'échantillon : "test F" 68
3.4.3 Comparaison de deux proportions d'échantillon 69
3.5 Tests non-paramétriques 72
3.5.1 Test d'ajustement de deux distributions : "test du khi-deux" 73
3.5.2 Test d'indépendance du khi-deux 76
3.5.3 Test d'homogénéité de plusieurs populations 79

Bibliographie **82**

Annexe **83**
Schémas 84
Synthèse sur les distributions d'échantillonnage 85
Synthèse sur les distributions d'échantillonnage 85
Estimation ponctuelle. Synthèse 86
Intervalle de confiance. Synthèse 87
Tables statistiques 89
Table de la loi Normale 90
Fractiles de la loi Normale 91
Fractiles de la loi du χ^2_ν 92

Table de la loi de Student 94
Table de la loi de Fisher-Snedecor $p = 0.05$ 96
Table de la loi de Fisher-Snedecor $p = 0.025$ 97
Table de la loi de Fisher-Snedecor $p = 0.01$ 98
Feuilles 99
Feuille 1 : Échantillonnage 100
Feuille 2 : Estimation 102
Feuille 3 : Les tests d'hypothèse 111
Feuille 4 : Préparation pour les contrôles 116

Chapitre 1

Echantillonnage - rappel

1.1 Introduction

L'échantillonnage représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée.

Avantages de l'échantillonnage

L'analyse d'un échantillon, par rapport à celle de la population, cout moindre, gain de temps et c'est la seule méthode qui donne des résultats dans le cas d'un test destructif.

FIGURE 1.1 : Statistique descriptive

Inconvénients de l'échantillonnage

L'échantillonnage a pour but de fournir suffisamment d'informations pour pouvoir faire des déductions sur les caractéristiques de la population. Les résultats obtenus d'un échantillon à l'autre sont en général différents et différents également de la valeur de la caractéristique correspondante dans la population. Ces différences sont dues aux fluctuations d'échantillonnage. Pour pouvoir tirer des conclusions valables, il faut déterminer les lois de probabilités qui régissent ces fluctuations.

Pour que les résultats observés lors d'une étude soient généralisables à la population statistique, **l'échantillon doit être représentatif** de cette dernière, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité. Seul l'échantillonnage aléatoire assure la représentativité de l'échantillon.

Un échantillon est qualifié d'**aléatoire** lorsque chaque individu de la population a une probabilité connue et non nulle d'appartenir à l'échantillon.

Le cas particulier le plus connu est celui qui attribue à chaque individu la même probabilité d'appartenir à l'échantillon.

Il y a 2 grandes catégories de méthodes d'échantillonnage :
— **l'échantillonnage non aléatoire :** l'analyste utilise son expérience et son jugement pour constituer l'échantillon avec tous les risques de non représentativité de celui-ci. On identifie dans la population-mère, quelques critères de répartition significatifs puis on essaye de respecter cette répartition dans l'échantillon d'individus interrogés.
La méthode d'échantillonnage non-probabiliste est utilisée lorsqu'il n'est pas possible de constituer une liste exhaustive de toutes les unités du sondage.
— **l'échantillonnage aléatoire ou probabiliste :** il permet de calculer précisément l'erreur due à l'échantillonnage et par conséquent de juger de la valeur de l'information partielle obtenue (donc de la représentativité de l'échantillon).

Par la suite, nous ne parlerons que de l'échantillon aléatoire simple : c'est un échantillon choisi de telle sorte que chaque unité de la population ait la même probabilité d'être sélectionnée dans l'échantillon et que chaque échantillon de même taille tiré de la population ait la même probabilité d'être choisi. On laisse dans ce cas le hasard choisir l'échantillon en utilisant par exemple une table de nombres au hasard.

Un échantillon aléatoire simple peut être tiré avec ou sans remise.

Dans l'*échantillon aléatoire simple avec remise*, chaque unité est remise dans la population après avoir été observée et avant qu'une autre unité soit choisie. Il y a donc indépendance entre les résultats d'un tirage à l'autre et chaque unité conserve la même probabilité d'être sélectionnée.

Dans l'*échantillon aléatoire simple sans remise* (échantillonnage exhaustif), l'unité tirée n'est pas remise ce qui modifie, pour une unité particulière, la probabilité d'être choisie d'un tirage à l'autre (si l'échantillon est choisi dans une population finie de N unités, chaque unité a une probabilité $\frac{1}{N}$ d'être choisie au 1er tirage, chaque unité restante une probabilité $\frac{1}{N-1}$ d'être choisie au 2e tirage, etc...). Dans ce cas, il n'y a pas d'indépendance d'un tirage à l'autre.

Si l'on a affaire à une population infinie ou si n, taille de l'échantillon, est relativement petite par rapport à N, taille de la population mère, on peut supposer qu'il y a indépendance d'une épreuve à l'autre, même si les tirages sont effectués sans remise. Dans le cas contraire, lorsque la population est finie et lorsque $n > 0,05N$, il faut tenir compte d'un facteur de correction ou d'exhaustivité (voir l'estimation de l'écart type).

On distingue 2 catégories de problèmes :
— **les problèmes de distribution d'échantillonnage :** lorsque on connaît la valeur de certains paramètres de la population mère et on cherche à induire des renseignements sur les

valeurs que peuvent prendre ces paramètres dans l'échantillon.
— **les problèmes d'estimation :** on connaît la valeur de certains paramètres dans l'échantillon et on cherche à induire des renseignements sur les valeurs que peuvent prendre ces paramètres dans la population mère.

Dans la suite du cours on utilisera les notations les suivantes :
— pour la population mère : taille : N, moyenne arithmétique de la variable étudiée : μ, variance : σ^2, écart type : σ.
— pour l'échantillon : taille : n, moyenne arithmétique mesurée sur l'échantillon : $\bar{x}$, variance : s^2, écart-type : s.

	Population	Échantillon
Définition	Ensemble des unités considérées par le statisticien	Sous-ensemble de la population choisie pour étude
Caractéristiques	Paramètres	Statistiques
Taille	N	n
Caractère **quantitatif**	moyenne de la population $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	moyenne de l'échantillon $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
	écart-type de la population $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$	écart-type de l'échantillon $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ $s' = \sqrt{\frac{n}{n-1}}s$
Caractère **qualitatif**	proportion dans la population p	proportion dans l'échantillon f

1.2 Les problèmes de distribution d'échantillonnage

1.2.1 Distribution d'échantillonnage de la moyenne $\bar{X}$

Dans une population mère de taille N, on peut tirer plusieurs échantillons de taille n : $\left(C_N^n = \frac{N!}{n!(N-n)!}\right)$.

Pour chaque échantillon, on peut calculer une moyenne :

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

et une variance

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

La valeur de l'espérance mathématique $\bar{x}$ et de la variance s^2 varient d'un échantillon à l'autre. C'est cette variation qui donne naissance à la distribution des variables aléatoires :

- **échantillonnage de la moyenne ou moyenne d'échantillon** $\bar{X}$, caractérisée par :

$E(\bar{X})$: l'espérance mathématique des moyennes calculées sur tous les échantillons de taille n.

$s_{\bar{X}}$: l'écart type de la distribution d'échantillonnage, qui représente la dispersion de l'ensemble des moyennes d'échantillons de taille n autour de $E(\bar{X})$

- **variance d'échantillon** $S'^2_{\bar{X}}$ définie par

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

L'espérance de $S'^2_{\bar{X}}$ est la variance de la population et $S'^2_{\bar{X}}$ $/E(S'^2_{\bar{X}}) = \sigma^2/$ est un estimateur **sans biais** de σ^2.

I. Cas : moyenne μ et écart-type σ de la population connus :

A) Si la population est infinie ou si l'échantillonnage est non exhaustif (tirage avec remise) :

— l'espérance mathématique de $\bar{X}$ est égale à la moyenne de la population :

$$E(\bar{X}) = \mu$$

— la variance de $\bar{X}$ est égale à la variance de la population divisée par la taille n de l'échantillon :

$$s^2_{\bar{X}} = \frac{\sigma^2}{n} \rightarrow s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Soit E_1, E_2, ..., E_p : p échantillons de taille n issues d'une même population mère de moyenne μ et de variance σ^2.
Soit $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_p$: leurs moyennes respectives.
Soit $\bar{X}$: la variable aléatoire qui prend pour valeur ces moyennes :

$$\bar{X} = \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$$

Alors lorsque $n \geq 30$, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ en vertu du théorème central limite.

Exemple 1.2.1 /Feuille 1/ Une machine effectue l'ensachage d'un produit.
On sait que les sacs ont un poids moyen de 250g avec un écart-type de 25g.
Quelles sont les caractéristiques de la moyenne des poids d'un échantillon de 100 sacs ?

Solution.
$(P) : \mu = 250, \ \ \sigma = 2,5; \ \ (E) : n = 100 > 30$
$\bar{X}$ suit la loi normale de paramètres $\mu = 250$ et $\frac{\sigma}{\sqrt{n}} = \frac{25}{10} = 2,5$.

Remarque 1 1. La moyenne de la distribution d'échantillonnage des moyennes est égale à la moyenne de la population.
2. On constate que plus n croît, plus $Var(\bar{X})$ décroît.

La distribution des moyennes d'échantillon est moins dispersée que la distribution initiale. En effet, à mesure que la taille de l'échantillon augmente, nous avons accès à une plus grande quantité d'informations pour estimer la moyenne de la population. Par conséquent, la différence probable entre la vraie valeur de la moyenne de la population et la moyenne échantillonnage diminue. L'étendue des valeurs possibles de la moyenne échantillonnale diminue et le degré de dispersion de la distribution aussi.

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ est aussi appelé l'**erreur-type** de la moyenne.

B) Si l'échantillonnage est exhaustif (tirage sans remise) dans une population finie (avec $n > 0.05N$) : on doit tenir compte d'un facteur d'exhaustivité pour déterminer $s_{\bar{X}}$.

Celui-ci devient : $s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

Échantillonnage exhaustif (tirage sans remise) dans une population finie avec $n > 0.05N$

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

Exemple 1.2.2 /Feuille 1/ Dans une usine textile, on utilise une machine automatique pour couper des morceaux de tissu. Lorsque la machine est correctement ajustée, la longueur des morceaux de tissu est en moyenne de 90 cm avec un écart type de 0.60 cm.

Pour contrôler la longueur des morceaux de tissu, on tire dans la production d'une journée un échantillon aléatoire de 200 morceaux.

a) Si l'on suppose que la longueur X des morceaux de tissu suit une loi normale, calculer la probabilité que la moyenne de l'échantillon soit au plus égale à 89.90 cm, ceci dans 2 cas :
— production de la journée : 10 000 morceaux
— production de la journée : 2 000 morceaux.

b) Déterminer la même probabilité sans faire l'hypothèse que X soit distribuée normalement.

c) Si la moyenne observée sur cet échantillon est de 90.30 cm, celui-ci est-il représentatif de la population mère en prenant un risque de 5 % de se tromper ? (avec $N = 10\ 000$).

Solution :

a) Production journalière $= N = 10\ 000$; Taille de l'échantillon $= n = 200$; $\frac{n}{N} = 0.02$

Même si l'échantillonnage est exhaustif, ce n'est pas la peine de tenir compte du coefficient d'exhaustivité.

Dans ce cas $E(\bar{X}) = 90$ cm et $s_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.6}{\sqrt{200}} = 0.042$.

Comme $X \sim \mathcal{N}(90, 0.6) \rightarrow \bar{X} \sim \mathcal{N}(90, 0.042)$

$$P(\bar{X} \leq 89.9) = P\left(T \leq \frac{89.9 - 90}{0.042}\right) = P(T \leq -2.38) = 1 - \pi(2.38) = 0.0087 \rightarrow 0.87\%$$

Production journalière $= N = 2\ 000 \rightarrow \frac{n}{N} = 0.1 \rightarrow$ on doit tenir compte du coefficient d'exhaustivité

$$s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{0.6}{\sqrt{200}}\sqrt{\frac{2000-200}{2000-1}} = 0.04$$
$$\bar{X} \sim \mathcal{N}(90, 0.04)$$

$$P(\bar{X} \leq 89.9) = P\left(T \leq \frac{89.9 - 90}{0.04}\right) = P(T - 2.5) = 1 - \pi(2.5) = 0.0062 \rightarrow 0.62\%$$

b) Même si l'on ne fait plus l'hypothèse que X soit une variable normale, comme $n = 200 > 30$, le théorème central limite permet de dire quel $\bar{X} \sim \mathcal{N}(90, 0.042)$ pour $N = 10000$.

On trouvera donc la même probabilité $P(\bar{X} \leq 89.9) = 0.0087 \rightarrow 0,87\%$.

c) L'échantillon est représentatif de la population mère avec un intervalle de confiance de 95 % lorsque :

$$P(\mu - t\, s_{\bar{X}} \leq \bar{x} \leq \mu + t\, s_{\bar{X}}) = 0.95$$

Lorsque la probabilité d'un intervalle symétrique est de 0.95, on a

$$t = 1.96 \quad \left(\pi(t) - \pi(-t) = 2\pi(t) - 1 = 0.95 \rightarrow \pi(t) = \frac{1.95}{2} = 0,975 \rightarrow t = 1,96\right).$$

$$P(90 - 1.96 \times 0.042 \leq \bar{x} \leq 90 + 1.96 \times 0.042) = 0.95$$

L'intervalle est donc $[89.917; 90.082]$. Comme $\bar{x} = 90.3$ cm, ne se situe pas dans cet intervalle de confiance, l'échantillon n'est pas jugé représentatif de la population mère (avec un risque de 5 % de se tromper).

C) Distribution de $\bar{X}_1 - \bar{X}_2$

Il peut arriver en statistique que l'on désire comparer 2 populations relativement à une certaine caractéristique X.
Population 1 : caractéristique X_1, moyenne : μ_1, variance σ_1^2, écart-type σ_1
Population 2 : caractéristique X_2, moyenne : μ_2, variance σ_2^2, écart-type σ_2

Pour comparer ces 2 populations, on tire indépendamment un échantillon aléatoire de taille n_1 dans la 1re et un échantillon aléatoire de taille n_2 dans la 2e et on considère la distribution de la différence $(\bar{X}_1 - \bar{X}_2)$.

D'après les propriétés de l'espérance mathématique et de la variance, on a :

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad \rightarrow \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}}$$

Si $\quad X_1 \sim \mathcal{N}(\mu_1, \sigma_1),\ X_2 \sim \mathcal{N}(\mu_2, \sigma_2) \rightarrow \left(\bar{X}_1 - \bar{X}_2\right) \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

— Si n_1 et n_2 sont grands (supérieurs à 30), quelles que soient les distributions de X_1 et X_2, $(\bar{X}_1 - \bar{X}_2)$ suivra une loi normale de mêmes paramètres, en vertu du théorème central limite.
— On utilisera le facteur d'exhaustivité dans les mêmes conditions (tirages sans remise, populations finies avec $n_i > 0.05N_i$).

Exemple 1.2.3 /Feuille 1/ Deux sociétés fabriquent des piles électriques d'un certain format.

Les piles de la société 1 ont une durée d'utilisation moyenne de 230 heures avec un écart type de 30 heures. Les piles de la société 2 ont une durée d'utilisation moyenne de 210 heures avec un écart type de 20 heures. Quelle est la probabilité que la durée d'utilisation moyenne d'un échantillon aléatoire simple de 100 piles de la société 1 soit d'au moins 30 heures de plus que la durée d'utilisation moyenne d'un échantillon aléatoire simple de 125 piles de la société 2 ?

Solution :
Soit :

X_1 la durée d'utilisation des piles de la société 1,

X_2 la durée d'utilisation des piles de la société 2.

On ne connaît pas les distributions de X_1 et X_2, mais comme les tailles $n_1 = 100$ et $n_2 = 125$ sont grandes (> 30), on peut dire que :

$$\begin{aligned}
(\bar{X}_1 - \bar{X}_2) &\sim \mathcal{N}\left(230 - 210, \sqrt{\frac{30^2}{100} + \frac{20^2}{125}}\right) \\
(\bar{X}_1 - \bar{X}_2) &\sim \mathcal{N}(20; 3.493) \\
P(\bar{X}_1 - \bar{X}_2 \geq 30) &= P\left(T > \frac{30 - 20}{3.493}\right) = P(T > 2.86) \\
&= 1 - \pi(2.86) = 0.0021 = 0.21\%
\end{aligned}$$

II. Cas : variance σ^2 de la population inconnue

A. Un grand échantillon ($n \geq 30$) permet de déduire une valeur fiable pour σ^2 en calculant la variance de l'échantillon s^2 et en posant

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

Les remarques précédentes restent valables :

Un grand échantillon $n \geq 30$ de variance s

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{s}{\sqrt{n-1}}\right)$$

B. Cas des petits échantillons : $n < 30$
On considère exclusivement le cas où X suit une loi normale dans la population.

Lorsque l'échantillonnage s'effectue à partir d'une population normale de variance inconnue et que la taille de l'échantillon est petite ($n < 30$), l'estimation de la variance effectuée par la

variance de l'échantillon n'est plus fiable. Comme s^2 varie trop d'échantillon en échantillon, on ne peut plus écrire que $\sigma^2 \approx \frac{n}{n-1}s^2$. L'écart-type de la distribution de $\bar{X}$ $\frac{\sigma}{\sqrt{n}}$, approximé par $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n-1}}$ n'est plus une constante et sa valeur varie dans chaque échantillon.

La variable écart-type d'échantillon, notée S est une variable aléatoire, définie par $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Considérons la variable aléatoire $T = \frac{\bar{X}-\mu}{s/\sqrt{n-1}} = \frac{\sqrt{n-1}(\bar{X}-\mu)}{s}$, dont le dénominateur n'est pas une constante. Alors, la variable T ne suit une loi normale.

En divisant numérateur et dénominateur par σ, on écrit T sous la forme

$$T = \frac{\sqrt{n-1}(\bar{X}-\mu)}{s} = \frac{\sqrt{n-1}\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}}.$$

dont le numérateur est composé par une variable aléatoire qui suit une loi $\mathcal{N}(0,1)$, multipliée par un facteur$\sqrt{n-1}$, et le dénominateur est une somme de carrés de variables suivant aussi la loi $\mathcal{N}(0,1)$. Le carré du dénominateur suit donc une loi du χ^2. Pour pouvoir utiliser correctement les tables du χ^2 il faut déterminer le nombre de degrés de liberté. Le nombre de degrés de liberté est toujours associée à une somme de carrés et représente le nombre de carrés indépendants dans cette somme. On peut calculer le nombre de degrés de liberté d'après deux règles :
- on effectue la différence entre le nombre total de carrés et le nombre de relations qui lient les différents éléments de la somme ;
- on effectue la différence entre le nombre total de carrés et le nombre de paramètres que l'on doit estimer pour effectuer le calcul.

Pour déterminer les degrés de liberté de la somme $\sum_{i=1}^{n}(\frac{X_i-\bar{X}}{\sigma})^2$, d'après la première règle le nombre de carrés dans la somme est n. Il y a une relation entre les variables $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$. Le nombre de degrés de liberté est donc $n-1$.

D'après la deuxième règle le nombre de carrés dans la somme est n. Lorsqu'on dit que $\sum_{i=1}^{n}(\frac{X_i-\bar{X}}{\sigma})^2$ est une somme de carrés de variables normales centrées réduites, on remplace μ par $\bar{X}$. On a estimé un paramètre. donc le nombre de degrés de liberté est $n-1$.

Si $n < 30$, et σ inconnu, la variable $T = \frac{\bar{X}-\mu}{s/\sqrt{n-1}}$ suit une loi de Student à $n-1$ degrés de liberté, notée T_{n-1}.

Exemple 1.2.4 /Feuille 1/ Le responsable d'une entreprise a accumulé depuis des années les résultats à un test d'aptitude à effectuer un certain travail. Il semble plausible de supposer que les résultats au test d'aptitude sont distribués suivant une loi normale de moyenne $\mu = 150$ et de variance $\sigma^2 = 100$. On fait passer le test à 25 individus de l'entreprise. Quelle est la probabilité que la moyenne de l'échantillon soit entre 146 et 154 ?

Solution :

On considère la variable aléatoire $\bar{X}$ moyenne d'échantillon pour les échantillons de taille $n = 25$. On cherche à déterminer $P(146 < \bar{X} < 154)$.

Pour cela, il nous faut connaître la loi suivie par $\bar{X}$. Examinons la situation. Nous sommes en présence d'un petit échantillon ($n < 30$) et heureusement dans le cas où la variable X (résultat au test d'aptitude) suit une loi normale. De plus, σ est connu. Donc $\bar{X}$ suit $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) = \mathcal{N}(150, 10/5)$. On en déduit que $Z = \frac{\bar{X}-150}{2}$ suit $\mathcal{N}(0,1)$.

La table donne

$$\begin{aligned} P(146 < \bar{X} < 154) &= P\left(\frac{146-150}{2} < Z < \frac{154-150}{2}\right) = P(-2 < Z < 2) \\ &= 2P(0 < Z < 2) = 2 \times (P(Z < 2) - P(Z < 0)) = 2 \times (0,9772 - 0,5) \\ &= 2 \times 0.4772 = 0.9544. \end{aligned}$$

1.2.2 Distribution de la variance d'échantillon $S'^2_{\bar{X}}$

Supposons que X suit une loi normale.

On considère la variable $Y = \frac{nS^2_{\bar{X}}}{\sigma^2} = \frac{n}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$.

Y est une somme d'écarts réduits relatifs à une variable normale, donc Y suit une loi du χ^2 à $n-1$ degrés de liberté (on perd un degré de liberté car on a estimé le paramètre μ par $\bar{X}$).

$$Y = \frac{nS^2_{\bar{X}}}{\sigma^2} \sim \chi^2_{n-1}.$$

Comme $S'^2 = \frac{n}{n-1}S^2$ et d'ici $S^2 = \frac{n-1}{n}S'^2$ on pet écrire

$$Y = \frac{(n-1)S'^2_{\bar{X}}}{\sigma^2} \sim \chi^2_{n-1}.$$

Approximation de la distribution de S'^2 dans le cas des grands échantillons : $n \geq 30$

Lorsque n est grand ($n \geq 30$), on peut approcher la loi χ^2_ν par la loi $\mathcal{N}(\nu, \sqrt{2\nu})$. Donc Y suit approximativement une loi normale, $E(Y) \approx n-1 = \nu$ et $Var(Y) \approx 2(n-1) = 2\nu$.

$$Y = \frac{(n-1)S'^2_{\bar{X}}}{\sigma^2} = \frac{\nu S'^2_{\bar{X}}}{\sigma^2} \sim \chi^2_\nu \longrightarrow_{n\geq 30} \mathcal{N}(\nu, \sqrt{2\nu}).$$

De $Y = \frac{\nu S'^2_{\bar{X}}}{\sigma^2} \sim \mathcal{N}(\nu, \sqrt{2\nu})$ et $S'^2 = \frac{Y\sigma^2}{\nu}$ on a

$$E(Y) = \nu; \quad E(S'^2_{\bar{X}}) = E\left(\frac{Y\sigma^2}{\nu}\right) = \frac{\sigma^2}{\nu}E(Y) = \frac{\sigma^2}{\nu}\nu = \sigma^2$$

$$V(Y) = 2\nu; \quad V(S'^2_{\bar{X}}) = V\left(\frac{Y\sigma^2}{\nu}\right) = \frac{\sigma^4}{\nu^2}V(Y) = \frac{\sigma^4}{\nu^2}2\nu = \frac{2\sigma^4}{\nu} = \frac{2\sigma^4}{n-1}$$

et d'ici on obtient la distribution de $S'^2_{\bar{X}}$, lorsque $n \geq 30$

Si $n \geq 30$, $S'^2_{\bar{X}} \sim \mathcal{N}\left(\sigma^2, \sigma^2\sqrt{\frac{2}{n-1}}\right)$ en première approximation.

La loi de $S'^2_{\bar{X}}$ est alors approximativement normale, son espérance vaut σ^2 et sa variance approximativement

$$Var(S'^2_{\bar{X}}) = Var\left(\frac{\sigma^2}{n-1}Y\right) = \frac{\sigma^4}{(n-1)^2}Var(Y) \approx \frac{2\sigma^4}{n-1}.$$

1.2.3 Distribution d'échantillonnage d'une proportion F

Dans certaines circonstances en gestion, on peut traiter les données sous forme de proportions (taux d'absentéisme, de rebuts, de réussite...).

Notations :

Population mère : p : proportion moyenne ; $q = 1 - p$ = proportion complémentaire

Echantillon : f : fréquence observée de l'échantillon de taille n.

Soit F la fréquence d'apparition du caractère dans un échantillon de taille n. Donc $F = X/n$ où X est le nombre de fois où le caractère apparaît dans le n-échantillon.

Par définition X suit $\mathcal{B}(n, p)$. Donc $E(X) = np$ et $Var(X) = npq$.

A) Si la population est infinie ou si l'échantillonnage est non exhaustif (tirage avec remise), on montre que :

$$E(F) = p; \quad s_F^2 = \frac{pq}{n}; \quad s_F = \sqrt{\frac{pq}{n}}$$

Si n est grand ($n \geq 30$) et $np \geq 15, nq \geq 15$, alors $\mathcal{B}(n, \frac{p}{n}) \to \mathcal{N}(p, \sqrt{\frac{pq}{n}})$ et d'ici $F \sim \mathcal{N}(p, \sqrt{\frac{pq}{n}})$

B) Si l'échantillonnage est exhaustif (tirage sans remise) dans une population finie (avec $n > 0.05N$**) :** on doit tenir compte du facteur d'exhaustivité.

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

Échantillonnage exhaustif (tirage sans remise) dans une population finie (avec $n > 0.05N$)

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

Exemple 1.2.5 /Feuille 7/ [2] Le directeur financier d'une société sait par expérience que 12 % des factures émises ne sont pas réglées dans les 10 jours ouvrables suivant l'échéance. Il fait prélever un échantillon aléatoire de 500 factures.
Quelle est la probabilité qu'au moins 70 factures ne sont pas réglées dans le délais, sachant que l'ensemble des factures pouvant être étudiées est de plusieurs dizaines de milliers.

Solution :

Soit F = "proportion d'échantillon dans un échantillon de taille 500". $P\left(F \geq \frac{70}{500}\right)$ =? - Distribution d'échantillonnage d'une proportion F ; échantillonnage exhaustif (tirage sans remise) dans une population finie, mais $n < 0,05N$, donc il ne faut pas tenir compte du facteur d'exhaustivité.
Ici $p = 0.12$, $q = 1 - p = 1 - 0.12 = 0.88$.
Comme $n = 500 > 30$, $np = 500 * 0,12 = 60 > 15$, $nq = 500 * 0,88 = 440 > 15 \Longrightarrow$ approximation de la loi binomiale par la loi normale :

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right) = \mathcal{N}\left(0,12; \sqrt{\frac{0,12 * 0,88}{500}}\right) = \mathcal{N}(0,12;\ 0,015)$$

$$\begin{aligned} P\left(F \geq \frac{70}{500}\right) &= P\left(F > \frac{69,5}{500}\right) = P\left(Z > \frac{0,139 - 0,12}{0,015}\right) \\ &= 1 - P\left(Z < \frac{0,019}{0,015}\right) = 1 - P(Z < 1,27) = 1 - \pi(1,27) = 1 - 0,8997 \approx 0,1 \end{aligned}$$

$\approx 10\%$ de chances pour que plus de 70 factures dans un 500 échantillon soient non réglées dans le délais.

Exemple 1.2.6 Selon une étude sur le comportement du consommateur, 25% d'entre eux sont influencés par la marque, lors de l'achat d'un bien. Si on interroge 100 consommateurs pris au hasard, quelle est la probabilité pour qu'au moins 35 d'entre eux se déclarent influencés par la marque ?

Solution :

Soit F = "proportion d'échantillon dans un échantillon de taille 100".
$P(F > 0,35)$ =? $\Longrightarrow$ il faut déterminer la loi de F. $n = 100 > 30$; $np = 100 \times 0,25 = 25 > 15$ et $nq = 100 \times 0,75 = 75 > 15$

$$\Longrightarrow F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right) = \mathcal{N}(0,25, 0,0433).$$

On utilise la variable $Z = \frac{F-0,25}{0,0433}$ qui suit la loi $\mathcal{N}(0,1)$.

$$\begin{aligned} P(F > 0,35) = P(Z > 2,31) &= 0,5 - P(0 < Z < 2,31) \\ &= 0,5 - 0,4896 = 0,0104. \end{aligned}$$

Conclusion :. Il y a environ une chance sur 100 pour que plus de 35 consommateurs dans un 100 - échantillon se disent influencés par la marque lorsque l'ensemble de la population contient 25% de tels consommateurs.

C) Distribution de $F_1 - F_2$

Lorsque n_1 et n_2 sont grands, alors :

$$(F_1 - F_2) \sim \mathcal{N}\left(p_1 - p_2; \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

1.3 Synthèse sur les distributions d'échantillonnage

<table>
<tr><th>Variable aléatoire</th><th>Définition</th><th>Paramètres descriptifs</th><th>Loi</th></tr>
<tr><td rowspan="3">F Proportion d'échantillon</td><td rowspan="3">$F = X/n$,
$X \sim \mathcal{B}(n,p)$
$E(X) = np$
$Var(X) = npq$</td><td rowspan="3">$E(F) = p$
$Var(F) = \frac{pq}{n}$</td><td>$n \geq 30$, $np > 15$, $nq > 15$
$\mathcal{B}(n, \frac{p}{n}) \to \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$</td></tr>
<tr><td>tirage avec remise
(sans remise et $n < 0,05N$)
$F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$</td></tr>
<tr><td>tirage sans remise et $n > 0,05N$
$F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$</td></tr>
<tr><td>$F_1 - F_2$
$F_1 \sim \mathcal{N}\left(p_1, \sqrt{\frac{p_1q_1}{n_1}}\right)$
$F_2 \sim \mathcal{N}\left(p_2, \sqrt{\frac{p_2q_2}{n_2}}\right)$</td><td>$F_1 - F_2$</td><td>$E(F_1 - F_2) = p_1 - p_2$
$Var(F_1 - F_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$</td><td>$n_1 \geq 30$; $n_2 \geq 30$
$F_1 - F_2 \sim \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}\right)$</td></tr>
</table>

TABLE 1.1 : Synthèse sur les distributions d'échantillonnage

<table>
<tr><th>Variable aléatoire</th><th>Définition</th><th>Paramètres descriptifs</th><th colspan="4">Loi</th></tr>
<tr><td rowspan="4">$\bar{X}$ Moyenne d'échantillon</td><td rowspan="4">$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$
$= \frac{1}{n}\sum_{i=1}^{n} X_i$</td><td rowspan="4">$E(\bar{X}) = \mu$
$Var(\bar{X}) = \frac{\sigma^2}{n}$</td><td colspan="2">$n \geq 30$</td><td colspan="2">$n < 30$, $X \sim \mathcal{N}(\mu, \sigma)$</td></tr>
<tr><td>σ connu</td><td>σ inconnu
estimation fiable
$\hat{\sigma}^2 = \frac{n}{n-1}s^2$</td><td>$\sigma$ connu</td><td rowspan="3">σ inconnu
estimation fiable
$\hat{\sigma}^2 = \frac{n}{n-1}s^2$

$T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n-1}}}$
$= \frac{\bar{X}-\mu}{\frac{s'}{\sqrt{n}}}$
$T \sim T_{n-1}$</td></tr>
<tr><td colspan="2">tirage avec remise ;
tirage sans remise et $n < 0,05N$
$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$</td></tr>
<tr><td colspan="2">tirage sans remise et $n > 0,05N$
$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right)$</td></tr>
<tr><td rowspan="3">X_1 : n_1, μ_1, σ_1
X_2 : n_2, μ_2, σ_2</td><td rowspan="3">$\bar{X}_1 - \bar{X}_2$</td><td rowspan="3">$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$;
$Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$</td><td colspan="2">$n_1, n_2 \geq 30$; $n_i < 0,05N$</td><td colspan="2">$n_1, n_2 < 30$ et
$X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$,
$X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$</td></tr>
<tr><td colspan="4">$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$</td></tr>
<tr><td colspan="4">$n_i > 0,05N \rightarrow$ facteur d'exhaustivité</td></tr>
<tr><td rowspan="2">$S'^2_{\bar{X}}$ Variance d'échantillon
- estimation de σ^2</td><td>$S^2_{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$</td><td>$E(S^2_{\bar{X}}) = \frac{n-1}{n}\sigma^2$,</td><td colspan="2">$n \geq 30$</td><td>$n < 30$</td></tr>
<tr><td>$S'^2_{\bar{X}} = \frac{n}{n-1}S^2$
$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$</td><td>$E(S'^2_{\bar{X}}) = \sigma^2$</td><td colspan="2">$S'^2_{\bar{X}} \sim \mathcal{N}\left(\sigma^2, \sigma^2\sqrt{\frac{2}{(n-1)}}\right)$</td><td>$\frac{(n-1)S'^2_{\bar{X}}}{\sigma^2} \sim \chi^2_{n-1}$</td></tr>
</table>

TABLE 1.2 : Synthèse sur les distributions d'échantillonnage

Chapitre 2

Estimation

L'estimation fait part de la Statistique inférentielle

L'estimation répond au problème réciproque à celui de l'échantillonnage : obtenir de l'information sur la population à partir d'échantillons. Ce problème comporte des incertitudes. Il ne pourra être résolu que moyennant un certain "risque d'erreur".

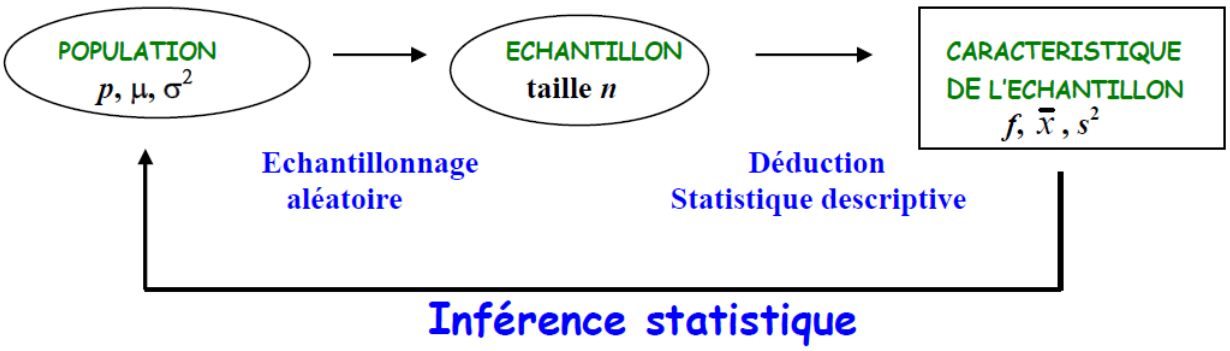


FIGURE 2.1 : [2] Statistiques inférentielle

Dans les problèmes d'estimation, on cherche à se faire une idée de la valeur d'un paramètre inconnu de la population mère à partir de données observées dans un échantillon - induction du particulier au général.

L'objectif est d'obtenir une bonne estimation de μ, p et σ à partir de $\bar{x}$, f et s, compte tenu de l'existence d'une dispersion dans la distribution d'échantillonnage.

Les méthodes d'estimation se divisent en 2 grandes catégories :
— **l'estimation ponctuelle** : on estime la valeur du paramètre inconnu de la population mère par un seul nombre à partir de l'information fournie par l'échantillon.
— **l'estimation par intervalle de confiance** : on estime un paramètre d'une population donnée par deux nombres qui forment un intervalle à l'intérieur duquel le paramètre de la population a de grandes chances de se trouver.
Les estimations par intervalles indiquent la précision d'une estimation et sont donc préférables aux estimations ponctuelles.

Exemple 2.0.1 Considérons la v.a. discrète X définie par la face obtenue en lançant le dé. En relançant le dé 100 fois puis 1000 fois, nous avons obtenu les répartitions suivantes : Les

Faces	1	2	3	4	5	6
Probabilités	1/6	1/6	1/6	1/6	1/6	1/6
Proportions (100 valeurs)	0.16	0.12	0.16	0.14	0.21	0.21
Proportions (1000 valeurs)	0.175	0.162	0.154	0.164	0.162	0.183

TABLE 2.1 : Résultats des lancers d'un dé équilibré à 6 faces

moyennes sont donc :
- Moyenne théorique :

$$\begin{aligned}\mu &= p_1x_1 + p_2x_2 + p_3x_3 + p_4x_4 + p_5x_5 + p_6x_6 \\ &= (1/6) \times 1 + (1/6) \times 2 + (1/6) \times 3 + (1/6) \times 4 + (1/6) \times 5 + (1/6) \times 6\end{aligned}$$

- Moyenne observée

$$\bar{x} = f_1x_1 + f_2x_2 + f_3x_3 + f_4x_4 + f_5x_5 + f_6x_6$$

sur les 100 valeurs :

$$\bar{x}_{100} = 0.16 \times 1 + 0.12 \times 2 + 0.16 \times 3 + 0.14 \times 4 + 0.21 \times 5 + 0.21 \times 6$$

sur les 1000 valeurs :

$$\bar{x}_{1000} = 0.175 \times 1 + 0.162 \times 2 + 0.154 \times 3 + 0.164 \times 4 + 0.162 \times 5 + 0.183 \times 6$$

On trouve :

$$\mu = 3.5 \quad \bar{x}_{100} = 3.75 \quad \bar{x}_{1000} = 3.525$$

La proximité entre la moyenne théorique (3.5) et les moyennes observées (3.75 et 3.525) est due à la convergence des proportions observées f_i vers les probabilités p_i. Plus les effectifs sont importants, plus ces proportions sont proches des probabilités, et plus la moyenne observée est proche de la moyenne théorique (au sens de la convergence en probabilité).

Le calcul détaillé pour la variance donne :
- Variance théorique :

$$\sigma^2 = p_1x_1^2 + p_2x_2^2p_3x_3^2 + p_4x_4^2 + p_5x_5^2 + p_6x_6^2 - \mu^2$$

- Variance observée :

$$s^2 = f_1x_1^2 + f_2x_2^2 + f_3x_3^2 + f_4x_4^2 + f_5x_5^2 + f_6x_6^2 - \bar{x}^2$$

On trouve, en notant s^2_{100} et s^2_{1000} les variances des échantillons de taille 100 et 1000 :

$$\sigma^2 = 2.917 \quad s^2_{100} = 3.0008 \quad s^2_{1000} = 3.045.$$

Les convergences des proportions f_i vers les probabilités p_i et de la moyenne empirique $\bar{x}$ vers la moyenne théorique μ assurent celle de la variance empirique vers la variance théorique. Mais cette convergence en probabilité est soumise au hasard, et c'est pour cela que la variance empirique s^2_{100} précédente est plus proche de la variance théorique σ^2 que s^2_{1000}.

2.1 Estimation ponctuelle

2.1.1 Qualités d'un estimateur

- **estimateur sans biais :** Comme un estimateur est une variable aléatoire (il y a autant d'estimateurs que d'échantillons de taille n), on dit que T est un estimateur sans biais d'un paramètre θ de la population si $E(T) = \theta$.

- **estimateur convergent :** T est un estimateur convergent pour θ si à mesure que la taille de l'échantillon augmente, T tend à prendre une valeur de plus en plus rapprochée de θ.

- **estimateur efficace :** T est l'estimateur le plus efficace de θ s'il est non biaisé et si sa variance est au moins aussi petite que celle de tout autre estimateur T' non biaisé : $E(T) = \theta$ et $V(T) \leq V(T')$.

- **estimateur exhaustif :** T est un estimateur exhaustif de θ si T résume toute l'information, contenue dans l'échantillon, qui est pertinente à θ. Plus un estimateur possédera de ces qualités, meilleur il sera.

2.1.2 Les estimateurs les plus utilisés

- **Estimation ponctuelle de la moyenne de la population :** $\widehat{\mu} = \bar{x}$

Soit $(X_1; X_2; \ldots; X_n)$ indépendantes et identiquement distribuées (i.i.d.) n observations de $X \sim \mathcal{N}(\mu; \sigma)$ ou grand échantillon ($n \geq 30$). $\forall i = \overline{1; n}\ E(X_i) = \mu\,;\, V(X_i) = \sigma^2$.

Pour estimer la moyenne μ de la population, on utilise le plus souvent la distribution d'échantillonnage de la moyenne dont l'estimateur est :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

1. Estimateur sans biais :

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu.$$

2. Estimateur convergent en probabilité :
Cas : population infinie ou tirage non exhaustif :

$$\begin{aligned} V(\bar{X}) &= s^2_{\bar{X}} = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0 \\ &\Rightarrow V(\bar{X}) = s^2_{\bar{X}} \to 0 \text{ quand } n \to +\infty. \end{aligned}$$

Cas : population finie et tirage exhaustif (sans remise) :
Si la population échantillonnée a un nombre fini d'individus de taille N, on conçoit que la loi de la population change après chaque tirage et que les tirages ne soient pas indépendants. On doit apporter le facteur de correction : $\frac{N-n}{N-1} \approx 1 - \frac{n}{N}$ à la variance de l'estimateur, si le taux de sondage $t = \frac{n}{N} > 5\%$.

$$V(\bar{X}) = s^2_{\bar{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)^2, \quad \text{comme} \quad \frac{N-n}{N-1} \xrightarrow[n\to\infty]{} 1 \Rightarrow V(\bar{X}) = s^2_{\bar{X}} \to 0 \text{ quand } n \to +\infty.$$

Toutefois ce facteur de correction peut être ignoré (≈ 1) si le taux de sondage est inférieur à 5%.

La distribution d'échantillonnage de la moyenne :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

est un excellent estimateur de μ.
La moyenne $\bar{x}$ observée sur l'échantillon est une estimation ponctuelle de la moyenne μ de la population : $\widehat{\mu} = \bar{x}$.

- **Estimation ponctuelle de la proportion de la population : $\widehat{p} = f$**

Soient $A_1; \ldots; A_i; \ldots; A_n$ n événements indépendants de probabilité p. Pour estimer la proportion p de la population, on utilise la proportion F de réalisation des événements A_i dans l'échantillon :

$$F = \frac{1}{n}\sum_{i=1}^{n} n_{A_i}$$

L'estimateur ainsi définit est :
1. Estimateur sans biais :

Démonstration :

On est ramené au cas estimation de la moyenne d'une loi de Bernoulli. En effet,

$$\begin{cases} (X_1;\ldots;X_i;\ldots;X_n) \text{ i.i.d.} \\ X_i \sim \mathcal{B}(p) \text{ Bernoulli} \\ \forall i \;\; E(X_i)=p \text{ et } V(X_i)=pq \end{cases} \Rightarrow \begin{cases} Y=\sum_i^n X_i \sim \mathcal{B}(n,p) \text{ Binomiale} \\ E(Y)=np \\ V(Y)=npq \text{ avec } q=1-p \end{cases}$$

$$\begin{cases} F=\frac{Y}{n}=\frac{1}{n}\sum_i^n X_i \\ E(F)=\frac{1}{n}E(Y)=p \\ V(F)=\frac{1}{n^2}V(Y)=\frac{pq}{n} \end{cases} \Rightarrow (n \text{ grand} : N \geq 30) \begin{cases} F \sim \mathcal{N}(p,\sqrt{\frac{pq}{n}}) \\ E(F)=p \text{ et } V(F)=\frac{pq}{n} \\ \text{si } n>0.05N \\ F \sim \mathcal{N}(p,\sqrt{\frac{pq}{n}\frac{N-n}{N-1}}) \end{cases}$$

$$E(F)=p.$$

2. Estimateur convergent en probabilité :
Cas : population infinie ou tirage non exhaustif :

$$Var(F)=\frac{pq}{n} \to 0 \text{ quand } n \to +\infty$$

Cas : population finie /de taille N/ et tirage exhaustif /taux de sondage $t=\frac{n}{N}>5\%$/ :

$$Var(F)=\frac{pq}{n}\frac{N-n}{N-1} \to 0 \text{ quand } n \to +\infty.$$

La proportion f observée sur l'échantillon est une estimation ponctuelle de la proportion p de la population $\Rightarrow \widehat{p}=f$.

Déterminer s_F, lorsque la proportion p de la population mère n'est pas connue.

$s_F=\sqrt{\frac{pq}{n}} \Rightarrow s_F^2=\frac{pq}{n}$. Si on ne connaît pas p et q, on les remplace par f et $(1-f)$ en tenant compte de la correction :

$$s_F^2=\frac{n}{n-1}\frac{f(1-f)}{n}=\frac{f(1-f)}{n-1} \Rightarrow s_F=\sqrt{\frac{f(1-f)}{n-1}} \quad \text{tirage avec remise.}$$

$$s_F^2=\frac{f(1-f)}{n-1}\frac{N-n}{N-1} \quad \text{tirage sans remise} \quad n>0.05N.$$

Exemple 2.1.1 /Feuille 2/ Supposons qu'une entreprise compte 200 employés et que l'échantillon de 50 employés a été prélevé au hasard parmi les deux cents.

Cat. salariale/mois	Nombre de salariés
Moins de 2 M.Euros	18
$[2-4[$	20
4 M.Euros et plus	12
Total	50

1. Donner une estimation de la proportion de l'ensemble des employés dont le salaire mensuel est de 2 M.Euros et plus.
2. Quel est le taux de sondage ?
3. Déterminer la probabilité qu'au moins 30 employés de cet échantillon possèdent un salaire mensuel de 2 M.Euros et plus lorsque la population échantillonnée en contient 64%.

Solution :

1.) $\hat{p} = f = \frac{20+12}{50} = \frac{32}{50} = 0.64 = 64\%$.

2.) $t = \frac{n}{N} = \frac{50}{200} = 0.25 > 0.05$.

3.) Soit F la v.a. proportion d'échantillon dans le cas de taux de sondage supérieur à 0.05 et proportion de la population $p = 64\%$ connue. On a $F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$.

On cherche la probabilité $P\left(F \geq \frac{30}{50}\right) = ?$

$$\begin{aligned} P\left(F \geq \frac{30}{50}\right) &= 1 - P\left(F \leq \frac{30}{50}\right) = 1 - \pi\left(\frac{30/50 - p}{\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}}\right) \\ &= 1 - \pi\left(\frac{30/50 - 0.64}{\sqrt{\frac{0.64*0.36}{50}}\sqrt{\frac{200-50}{200-1}}}\right) = 1 - \pi(-0.06781) = 1 - 1 + \pi(0.06781) \\ &= 0.52 = 52\% \end{aligned}$$

- **Estimation ponctuelle de la variance et de l'écart-type de la population**
- **Cas : μ connue**

Soient $X_1; X_2; \ldots; X_n$ n observations indépendantes de même loi de moyenne μ et de variance σ^2. Pour estimer σ^2, si la moyenne μ est connue, on peut construire l'estimateur :

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

1. Estimateur sans biais :

$$E(S^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n} E\left((X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n} V(X_i) = \sigma^2$$

$/V(X) = E\left((X - \mu)^2\right);\ V(X_i) = \sigma^2/$

2. Estimateur convergent :

$$\begin{aligned} V(S^2) &= V\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left((X_i - \mu)^2\right) = \frac{1}{n}\left(E\left((X - \mu)^4\right) - \left(E\left((X - \mu)^2\right)\right)^2\right) \\ &= \frac{1}{n}(\mu_4 - \sigma^4) \to 0, \quad \text{lorsque} \quad n \to +\infty, \end{aligned}$$

avec $\mu_k = E\left((X - \mu)^k\right)$.

La variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ observée sur l'échantillon est une estimation ponctuelle de la variance σ^2 de la population échantillonnée lorsque la moyenne μ de la population est connue.

- **Cas : μ inconnue**

Lorsque la moyenne μ est inconnue (cas le plus fréquent), pour estimer σ^2, on pourrait utiliser naturellement l'estimateur :

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

après avoir estimé μ.
Cependant, l'estimateur S^2 est biaisé : $E(S^2) = \frac{n-1}{n}\sigma^2$, on préfère alors d'utiliser l'estimateur :

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

appelé : carré de la déviation standard empirique.
1. Estimateur sans biais :

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

2. Estimateur convergent :

$$\begin{aligned} Var(S'^2) &= V\left(\frac{n}{n-1}S^2\right) = \frac{1}{(n-1)^2}\sum_{i=1}^{n} V\left((X_i - \bar{X})^2\right) \\ &= \frac{1}{n-1}\left(E\left((X-\bar{X})^4\right) - \left(E\left((X-\bar{X})^2\right)\right)^2\right) \\ &= \frac{1}{n-1}\left(\mu_4 - s^4\right) \xrightarrow[n\to\infty]{} 0, \text{ avec } \mu_4 = E\left((X-\bar{X})^4\right). \end{aligned}$$

La variance empirique

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

basée sur la variance observée s^2 sur l'échantillon est une estimation ponctuelle de la variance σ^2 de la population échantillonnée lorsque la moyenne μ de la population est inconnue.

$S' = S\sqrt{\frac{n}{n-1}}$ est un estimateur sans biais de σ.
$s' = s\sqrt{\frac{n}{n-1}}$ est une estimation ponctuelle de l'écart-type σ de la population.

Dans l'estimation ponctuelle de la moyenne μ, lorsqu'on ne connaît pas l'écart-type de la population mère, on détermine $s_{\bar{X}}$:

$s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Si on ne connaît pas σ on le remplace par s' et on a :

$$s_{\bar{X}} = \frac{s'}{\sqrt{n}} \quad \text{où} \quad s' = s\sqrt{\frac{n}{n-1}} \quad \text{et} \quad s = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \text{donc} \quad s_{\bar{X}} = \frac{s}{\sqrt{n-1}}$$

Exemple 2.1.2 /Feuille 2/ [8] Les prix d'un article en 5 différents marchés d'une région donnée sont :

i	1	2	3	4	5
x_i	75	82	83	78	80

Calculer les estimations ponctuelles de la moyenne et de l'écart-type.

Solution

L'effectif $n = 5$ de l'échantillon est inférieur à 30 et la moyenne μ et la variance σ^2 de la population sont inconnus. On utilise les expressions d'estimation ponctuelle les suivantes :
Moyenne : $\hat{\mu} = \bar{x} = \sum_{i=1}^{5} x_i = \frac{398}{5} = 79.6$
Ecart-type : $\hat{\sigma} = s\sqrt{\frac{n}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{5} x_i^2 - 5\bar{x}^2}{4}}$
On ajoute encore une ligne à la table :

i	1	2	3	4	5	Total
x_i	75	82	83	78	80	398
x_i^2	5625	6724	6889	6084	6400	31722

$$\hat{\sigma} = \sqrt{\frac{n}{n-1}}s = \sqrt{\frac{\sum_{i=1}^{5} x_i^2 - 5 * \bar{x}^2}{4}} = \sqrt{\frac{31722 - 5 * 6336.16}{4}} = 3.209361 \approx 3.21$$

Exemple 2.1.3 /Feuille 2/ [8] La table de distributions des salaires en € de 100 employés d'une entreprise est donnée ci-dessous :

Classe	Centre de la classe x_i^*	Effectif n_i
400 , 500	450	11
500 , 600	550	30
600 , 700	650	39
700 , 800	750	18
800 , 900	850	2

Calculer les estimations ponctuelles de la moyenne et de l'écart-type.

Solution

Comme les données sont groupées en classes, on utilise les expressions pour D.G.1.
On ajoute encore deux colonnes et une ligne à la table :

Classe	Centre de la classe x_i^*	Effectif n_i	$n_i x_i^*$	$n_i x_i^{*2}$
400 , 500	450	11	4950	2227500
500 , 600	550	30	16500	9075000
600 , 700	650	39	25350	16477500
700 , 800	750	18	13500	10125000
800 , 900	850	2	1700	1445000
Totale		100	62000	39350000

Moyenne : $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^* = \frac{1}{100}\sum_{i=1}^{5} n_i x_i^* = \frac{62000}{100} = 620$ €.
Ecart-type : $\hat{\sigma} = s\sqrt{\frac{n}{n-1}} = \sqrt{\frac{\sum_{i=1}^{k} n_i x_i^{*2} - n\bar{x}}{n-1}} = \sqrt{\frac{39350000-38440000}{99}} = 95.87$

2.2 Estimation par intervalle de confiance

Définition d'une région de confiance Le staticien fix à l'avance un petit nombre $\alpha \in (0,1)$ - un un niveau de risque, le seuil des probabilités significatives ou simplement le seuil. Les valeurs usuelles de α sont 1%, 5% ou 10%.

On cherche 2 statistiques $\Lambda_1 = f(X_1, \ldots, X_n)$ et $\Lambda_2 = f(X_1, \ldots, X_n)$ telles que l'on ait

$$P(\Lambda_1 \leq \theta \leq \Lambda_2) \geq 1 - \alpha \Longrightarrow$$

Il y a une probabilité forte (supérieure ou égale à $1-\alpha$) pour que l'intervalle aléatoire $[\Lambda_1, \Lambda_2]$ contient le nombre inconnu θ.

A la suite de prélèvement de l'échantillon Λ_1 prend la valeur $\widehat{\theta}_1$ et Λ_2 la valeur $\widehat{\theta}_2$.

L'intervalle $I.C._{1-\alpha} = [\widehat{\theta}_1, \widehat{\theta}_2]$ est un intervalle (unilatère ou bilatère) de confiance pour θ de seuil α ou de niveau de confiance $1-\alpha$

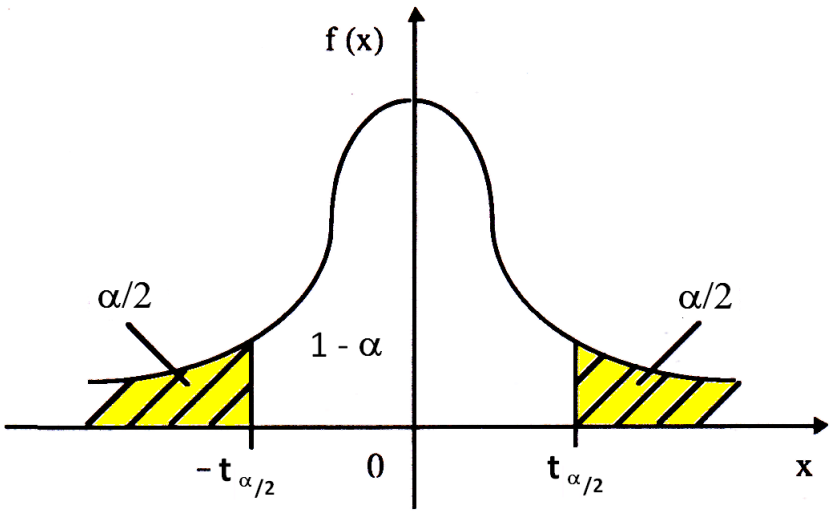
Un intervalle de confiance de niveau de confiance 95% a une probabilité au moins égale à 0.95 de contenir la vraie valeur inconnue θ. Par passage au complémentaire, le niveau de risque α correspondant à une majoration de la probabilité que la vraie valeur du paramètre θ ne soit pas dans $I.C._{1-\alpha}$. A niveau de confiance fixé, une région de confiance est d'autant meilleure qu'elle est de taille petite.

Obtention d'un intervalle de confiance

Soient $Y = f(X_1, \ldots, X_n)$ et $Z = g(X_1, \ldots, X_n)$ deux statistiques, telles que la v.a. $T = \frac{Y-\theta}{Z}$ obéisse à la loi normale centrée réduite ou à la loi de Studnet.

On cherche dans les tables un nombre $t_{\frac{\alpha}{2}}$ tel que :

$$P(|T| > t_{\frac{\alpha}{2}}) \leq \alpha, \text{ c'est-à-dire encore } P(|T| \leq t_{\frac{\alpha}{2}}) \geq 1 - \alpha$$



On aura donc

$$P\left(\left|\frac{Y-\theta}{Z}\right| \leq t_{\frac{\alpha}{2}}\right) \geq 1 - \alpha$$

ce qui est équivalent à

$$P(Y - t_{\frac{\alpha}{2}} Z \leq \theta \leq Y + t_{\frac{\alpha}{2}} Z) \geq 1 - \alpha.$$

L'intervalle $I.C._{1-\alpha} = [Y - t_{\frac{\alpha}{2}} Z, Y + t_{\frac{\alpha}{2}} Z]$ est, pour θ un intervalle de confiance de seuil α.

Choix du fractile $t_{\frac{\alpha}{2}}$

On choisie dans la table le fractile $t_{\frac{\alpha}{2}}$ qui vérifie l'égalité :

- pour un intervalle bilatéral $P\left(|T| > t_{\frac{\alpha}{2}}\right) = \alpha$, qui est équivalent aux

$$P(T > t'_{\frac{\alpha}{2}}) = \frac{\alpha}{2}, \text{ et } P(T < t''_{\frac{\alpha}{2}}) = \frac{\alpha}{2};$$

- pour un intervalle unilatéral à droite $P(T > t_\alpha) = \alpha$;
- pour un intervalle unilatéral à gauche $P(T < t_\alpha) = \alpha$.

Si on diminue α, c'est-à-dire augmente la confiance, on augmente $t_{\frac{\alpha}{2}}$ et, par suite augmente l'intervalle de confiance (plus un intervalle est grand, plus on peut avoir confiance en lui)

2.2.1 Intervalle de confiance de la moyenne d'une population : μ

μ n'est pas connu mais on sait que la moyenne de l'échantillon, $\bar{X}$, est un excellent estimateur de μ.

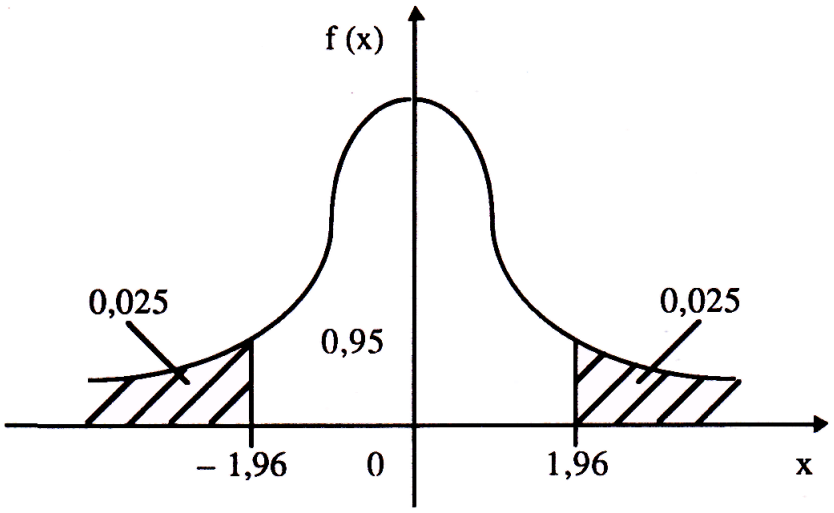
- **Cas : σ^2 connue et $n \geq 30$ ou $X \sim \mathcal{N}(\mu, \sigma)$**

Lorsque la variance de la population σ^2 est connue, la distribution d'échantillonnage de $\bar{X}$ est approximativement normale de moyenne $E(\bar{X}) = \mu$ et de variance connue $s_{\bar{X}} = \frac{\sigma^2}{n}$.

La statistique de test : $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1)$

On peut alors écrire : $P\left(-t_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq t_{\frac{\alpha}{2}}\right) = 1-\alpha$.

On détermine les fractiles $t_{\frac{\alpha}{2}}$ de la loi $\mathcal{N}(0,1)$: $P\left(-t_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq t_{\frac{\alpha}{2}}\right) = 1-\alpha$ [1]



Valeurs des fractiles $t_{\frac{\alpha}{2}}$ de la loi $\mathcal{N}(0,1)$ pour certains niveaux de risque α :

α	$1-\alpha$	$t_{\frac{\alpha}{2}}$
0,1	0,9	1,645
0,5	0,95	1,960
0,01	0,99	2,576

On en déduit l'intervalle de confiance de niveau $(1-\alpha)$ de μ :

$$\bar{x} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Marge d'erreur dans l'estimation de μ : $E = t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

L'intervalle $[\bar{x} - t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{x} + t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$ est "bilatéral symétrique" de niveau $1-\alpha$ de la moyenne μ centré en $\bar{x}$.

L'intervalle de confiance est l'intervalle de valeurs tel que l'on a une probabilité de $(1-\alpha)$ (fixée à l'avance) d'avoir la moyenne μ comprise entre les 2 bornes $\bar{x} - t_{\frac{\alpha}{2}} s_{\bar{x}}$ et $\bar{x} + t_{\frac{\alpha}{2}} s_{\bar{x}}$:

$$P(\bar{x} - t_{\frac{\alpha}{2}}\, s_{\bar{x}} < \mu < \bar{x} + t_{\frac{\alpha}{2}} s_{\bar{x}}) = (1-\alpha)$$

Ceci n'est strictement valable que si la population est distribuée normalement ou si $n \geq 30$.

- **Cas : σ^2 inconnue**

Lorsque la population est distribuée normalement, que σ n'est pas connu et que l'échantillon est de faible taille ($n < 30$), on se réfère à la loi de Student Fisher, similaire à la loi normale mais qui donne des valeurs de t différentes pour tenir compte de l'aléa plus grand engendré par un échantillon réduit.

La lecture de la table de Student (voir annexes) donne directement la valeur de t en fonction du nombre de degrés de liberté $(n-1)$ et du risque accepté α.

Par exemple, si $n = 16$ et l'intervalle de confiance est à $1 - \alpha = 95$:

le nombre de d.l. $= 16 - 1 = 15$
le risque accepté $\alpha = 1 - 0.95 = 0.05$ $\Rightarrow$ le t de Student $= 2{,}131$

Lorsque la variance σ^2 est inconnue on doit d'abord estimer la moyenne μ pour estimer σ^2 :

$$\text{Estimateur sans biais :} \quad S'^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

$$\text{Estimation :} \quad s'^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

Dans ce cas, la distribution d'échantillonnage de $\bar{X}$ a pour moyenne $E(\bar{X}) = \mu$ et de variance estimée $Var(\bar{X}) = \frac{s'^2}{n}$.

La statistique de test : $\frac{\bar{X}-\mu}{S'/\sqrt{n}} \rightarrow T_{n-1}$ *d.d.l.*

On peut alors écrire : $P\left(-t_{St\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{s'/\sqrt{n}} \leq t_{St\frac{\alpha}{2}}\right) = 1 - \alpha$

Les fractiles $t_{St\frac{\alpha}{2}}$ de la loi de Student à n d.d.l. (cf. table) :

$$P\left(-t_{St\frac{\alpha}{2}} \leq T_n \leq t_{St\frac{\alpha}{2}}\right) = P\left(|T_n| \leq t_{St\frac{\alpha}{n}}\right) = 1 - \alpha$$

On en déduit l'intervalle de confiance de niveau $(1-\alpha)$ de μ :

$$\bar{x} - t_{St\frac{\alpha}{2}} \frac{s'}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{St\frac{\alpha}{2}} \frac{s'}{\sqrt{n}}$$

Marge d'erreur dans l'estimation de μ : $E = t_{St\frac{\alpha}{2}} \frac{s'}{\sqrt{n}}$

Lorsque la population est distribuée normalement, que σ n'est pas connu et que l'échantillon est de faible taille ($n < 30$), on se réfère à la loi de Student Fisher.

Approximation : si la taille de l'échantillon est grande ($n \geq 30$) alors on peut remplacer la valeur du fractile $t_{St\frac{\alpha}{2}}$ de Student à $(n-1)$ *d.d.l.* par celle du fractile $t_{\frac{\alpha}{2}}$ de la loi normale centrée-réduite $\mathcal{N}(0,1)$.

Exemple 2.2.1 [2] /Feuille 2/

1. Soit X la v.a. «durée de vie du tube cathodique d'une marque de T.V.».

   On ne connaît pas la moyenne des durées de vie des tubes bien que l'on sache qu'elles sont distribuées normalement. L'écart-type de la distribution des durées de vie $\sigma = 450$.

   Dans un échantillon de 55 tubes on a calculé que la durée de vie moyenne était de 9 500 heures.

   Déterminer l'intervalle de confiance à 90 % de la durée de vie moyenne de la population des tubes.

   Solution :
   Comme la population est distribuée normalement, que σ est connu et que $n = 55 > 30$, on peut utiliser la loi normale.

   Pour $1-\alpha = 90\%$ on a $P(-t < T < t) = 0.90 \Rightarrow \pi(t) - \pi(-t) = 0.90 \Rightarrow 2\pi(t) - 1 = 0.90$
   $\Rightarrow \pi(t) = \frac{1.90}{2} = 0.95 \Rightarrow t = 1.645.$
   Donc $P(\bar{X} - 1.645\ s_{\bar{X}} < \mu < \bar{X} + 1.645\ s_{\bar{X}}) = 0.90$
   $s_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{450}{\sqrt{55}} = 60.678$
   L'intervalle de confiance à 90 % $= [9500 - 1.645 \times 60.678; 9500 + 1.645 \times 60.678]$
   $= [9400.18; 9599.81]$

   Remarque :

   Dans ce cas, même si la population n'était pas distribuée normalement, on aurait trouvé le même intervalle de confiance à 90 % en vertu du théorème central limite qui nous assure que, pour $n \geq 30$ (ici $n = 55$), la distribution d'échantillonnage de la moyenne peut être supposée normale même si la population ne l'est pas.

2. Reprenons le même exemple, mais cette fois l'échantillon est de taille $n = 25$. Déterminons l'intervalle de confiance à 99 % de la durée de vie moyenne des tubes, sachant que $\bar{x} = 9500$ heures.

   Solution :

   $$X \sim \mathcal{N}(\mu, 450);\quad n = 25,\quad \bar{X} = 9500,\quad 1-\alpha = 99\%$$

   On peut utiliser la loi normale car la population est normale et que σ est connu.

   Pour $1-\alpha = 99\% \Rightarrow P(-t < T < t) = 0.99 \Rightarrow \pi(t) - \pi(-t) = 2\pi(t) - 1 = 0.99$

$\Rightarrow \pi(t) = \frac{1.99}{2} = 0.995 \Rightarrow t = 2.575.$

$s_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{450}{\sqrt{25}} = 90$

Donc $P(9500 - 2.575 \times 90 < \mu < 9500 + 2.575 \times 90) = 0.99$

L'intervalle de confiance à 99% = [9 268.25 ; 9 731.75].

3. Supposons que la population soit distribuée normalement, mais que σ ne soit pas connu. A partir d'un échantillon de taille $n = 60$, nous avons $\bar{x} = 9450$ et $s = 446.234$.

   Estimons à l'aide d'un intervalle de confiance à 95 % la moyenne de la population.

   Solution :
   Comme $n = 60 > 30$, on peut utiliser la loi normale. De plus, comme la population est distribuée normalement, ce n'est pas la peine de faire appel au théorème central limite.

   Pour $1 - \alpha = 95\% \Rightarrow P(-t < T < t) = 0.95 \Rightarrow 2\pi(t) - 1 = 0.95 \Rightarrow \pi(t) = \frac{1.95}{2} = 0.975 \Rightarrow t = 1.96.$

   $s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Nous ne connaissons pas σ, il faut l'estimer.

   $s' = s\sqrt{\frac{n}{n-1}} = 446.234\sqrt{\frac{60}{59}} = 450$

   $s_{\bar{X}} = \frac{s'}{\sqrt{n}} = \frac{450}{\sqrt{60}} = 58.094$

   Donc $P(9450 - 1.96 \times 58.094 < \mu < 9450 + 1.96 \times 58.094) = 0,95$

   L'intervalle de confiance à 95 % = [9 336.13 ; 9 563.86]

   Remarque :

   Dans ce cas, même si la population n'était pas distribuée normalement, on aurait trouvé le même intervalle de confiance à 95 %, en vertu du théorème central limite (car $n = 60 > 30$).

4. Supposons que la distribution soit normale, que σ ne soit pas connu, et que l'écart type s d'un échantillon de taille $n = 25$ soit égal à 440,908, $\bar{x}$ étant égal à 9 500.

   Déterminons l'intervalle de confiance à 99 % et comparons le à celui de l'exemple 2.

   Solution :
   Comme on suppose que la population est distribuée normalement, que σ est inconnu, que $n = 25 < 30$, on peut utiliser ici la loi de Student pour calculer les $t_{St\frac{\alpha}{2}}$.

   nombre de d.l. $= n - 1 = 25 - 1 = 24$
   le risque accepté $= \alpha = 1 - 0.99 = 0.01$ $\Rightarrow t_{St\frac{\alpha}{2}} = 2.797$

   σ n'est pas connu $\Rightarrow s' = s\sqrt{\frac{n}{n-1}} = 440.908\sqrt{\frac{25}{24}} = 450$

   $s_{\bar{X}} = \frac{s'}{\sqrt{n}} = \frac{450}{\sqrt{25}} = 90$

   Donc $P(9500 - 2.797 \times 90 < \mu < 9500 + 2.797 \times 90) = 0.99$

   L'intervalle de confiance = [9 248.27 ; 9 751.73].

Cet intervalle de confiance est plus étendu que celui de l'exemple 2 (à conditions à peu près identiques, à l'utilisation de la loi de Student près), Ceci s'explique par l'aléa plus important dû à l'estimation de l'écart type de la population sur un échantillon de petite taille.

2.2.2 Intervalle de confiance de la proportion d'une population : p

p n'est pas connue et on cherche à l'estimer à partir de l'échantillon.

L'intervalle de confiance est l'intervalle de valeurs tel que l'on a une probabilité $1-\alpha$ % (fixée à l'avance) d'avoir la proportion p comprise entre les 2 bornes $f - ts_F$ et $f + ts_F$ Dans le cas de grande taille de l'échantillon prélevé ($n \geq 30$), l'estimation par intervalle de confiance de p (inconnue) de la population se déduit de la distribution d'échantillonnage de la proportion :

$$F = \frac{1}{n}\sum_{i}^{n} X_i \quad (X_1; \ldots; X_i; \ldots; X_n) \quad \text{i.i.d.} \quad X_i \sim \mathcal{B}(p)$$

La distribution d'échantillonnage de F est approximativement normale de moyenne $E(F) = p$ et de variance en fonction de p (inconnue) $Var(F) = \frac{pq}{n}$ estimée par son estimateur $\frac{f(1-f)}{n-1}$ ou en convergence par $\frac{f(1-f)}{n}$.

La statistique de test : $\frac{F-p}{\sqrt{\frac{f(1-f)}{n}}} \sim \mathcal{N}(0;1)$

On peut alors écrire : $P\left(-t_{\frac{\alpha}{2}} \leq \frac{F-p}{\sqrt{\frac{f(1-f)}{n}}} \leq t_{\frac{\alpha}{2}}\right) = 1-\alpha$

On en déduit l'intervalle de confiance de niveau $(1-\alpha)$ de p :

$$f - t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} \leq p \leq f + t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$$

L'intervalle asymptotique de confiance de niveau $(1-\alpha)$ de p est :

$$f - t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} \leq p \leq f + t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$$

Marge d'erreur dans l'estimation de p : $E = t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$.

Intervalle "bilatéral symétrique" de niveau $1-\alpha$ de la proportion p centré en f.

$$P(f - t_{\frac{\alpha}{2}} s_F < p < f + t_{\frac{\alpha}{2}} s_F) = 1 - \alpha\%$$

Cette approximation de la loi Binomiale par la loi Normale n'est valable que lorsque $n > 30$, $np > 5$, $nq > 5$.
$s_F = \frac{\sqrt{pq}}{n}$. Comme on ne connaît pas p on estime s_F par l'estimation en convergence $\sqrt{\frac{f(1-f)}{n}}$.
Au seuil de probabilité de $(1-\alpha)\%$, l'intervalle asymptotique de confiance pour p sera :

$$\left[f - t_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}; f + t_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}} \right]$$

Exemple 2.2.2 [2] /Feuille 2/

Les responsables d'une étude de marché ont choisi au hasard 500 femmes dans une grande ville et ont constaté que 35 % des femmes retenues dans l'échantillon préfèrent utiliser une marque de lessive A plutôt que les autres. Ils veulent déterminer l'intervalle de confiance à 95 % de la proportion des femmes de cette ville qui préfèrent la marque de lessive A.

Solution :

$f = 0.35 \Rightarrow s = \sqrt{\frac{0.35 \times 0.65}{500}} = 0.021331.$

$P(0.35 - 1.96 \times 0.02133 < p < 0.35 + 1.96 \times 0.02133) = 0.95$

L'intervalle de confiance est donc [0.3082 ; 0.3918]. Il y a donc entre 30.82 % et 39.18 % des femmes de cette ville qui préfèrent la marque de lessive A (avec un risque de 5 % de se tromper).

2.2.3 Précision - Taille d'échantillon - Risque d'erreur

1. La marge d'erreur ou niveau de précision recherché dans l'estimation par intervalle de confiance, lorsqu'on utilise l'estimation $\bar{\theta}$ de l'échantillon pour estimer la vraie valeur θ de la population, est l'écart (en valeur absolue), noté $E = |\bar{\theta} - \theta|$.

2 En pratique, on peut fixer la marge d'erreur qu'on ne veut pas excéder et déterminer la taille minimale de l'échantillon requise.

3 On peut déduire le risque d'erreur ou le niveau de confiance attribué à une estimation par intervalle.

Paramètre	Marge d'erreur	Taille d'échantillon	Risque d'erreur
Moyenne μ (σ^2 connue)	$E = t_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$	$n = \left(t_{\frac{\alpha}{2}} \frac{\sigma}{E}\right)^2$	$t_{\frac{\alpha}{2}} = \frac{\sqrt{n}}{\sigma} E$
Moyenne μ (σ^2 inconnue)	$E = t_{St\frac{\alpha}{2}} \frac{s'}{\sqrt{n}}$	$n = \left(t_{St\frac{\alpha}{2}} \frac{s'}{E}\right)^2$	$t_{St\frac{\alpha}{2}} = \frac{\sqrt{n}}{s'} E$
Proportion p	$E = t_{\frac{\alpha}{2}} \sqrt{\frac{pq}{n}}$	$n = \left(\frac{t_{\frac{\alpha}{2}}}{E}\right)^2 f(1-f)$	$t_{\frac{\alpha}{2}} = \sqrt{\frac{n}{f(1-f)}} E$

Exemple 2.2.3 /Feuille 2/

Les responsables d'une étude de marché ont choisi au hasard 500 femmes dans une grande ville et ont constaté que 35 % des femmes retenues dans l'échantillon préfèrent utiliser une marque de lessive A plutôt que les autres.

Supposons qu'avant de tirer l'échantillon, les responsables de l'étude aient décidé d'estimer la proportion p à $\pm 2\%$ près.

Quelle devrait être dans ce cas la taille minimale de l'échantillon à tirer, en désirant toujours avoir un intervalle de confiance à 95 % et en considérant que $f = 0.35$.

Solution :

Pour avoir la proportion à 2 % près, il faut que :

$$1.96\sqrt{\frac{0.35 \times 0.65}{n}} = 0.02$$
$$\Rightarrow (1.96)^2 \frac{0.35 \times 0.65}{n} = (0.02)^2$$
$$\Rightarrow n = \frac{(1.96)^2 \times 0.35 \times 0.65}{(0.02)^2} = 2184.91 = 2185.$$

2.2.4 Intervalle de confiance de la variance de la population : σ^2

- **Cas : μ connue**

σ^2 n'est pas connu mais on sait que la variance s^2 de l'échantillon, est un excellent estimateur de σ^2, lorsque la moyenne μ de la population est connue.

Lorsque la moyenne μ est connue, on peut montrer que :

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} U_i^2 \sim \chi^2_{n\ d.d.l.} \text{ avec } U_i \sim \mathcal{N}(0;1).$$

(cf. définition d'une variable aléatoire du khi-deux comme somme de carrés de variables aléatoires normales centrées réduites indépendantes).

La statistique de test : $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = n\frac{S^2}{\sigma^2} \sim \chi^2_{n\ d.d.l.}$

On peut alors écrire : $P(k_1 \leq n\frac{S^2}{\sigma^2} \leq k_2) = 1 - \alpha$. où, $k_1 = \chi^2_{\frac{\alpha}{2}}$ et $k_2 = \chi^2_{1-\frac{\alpha}{2}}$ sont les fractiles de la loi khi-deux à n degrés de liberté (cf. table du khi-deux). c'est-à-dire : $P(\chi^2_n \leq k_1) = \frac{\alpha}{2}$ et $P(\chi^2_n \leq k_2) = 1 - \frac{\alpha}{2}$.

On en déduit l'intervalle de confiance de niveau $(1 - \alpha)$ de σ^2 :

$$n\frac{s^2}{k_2} \leq \sigma^2 \leq n\frac{s^2}{k_1}$$

L'intervalle de confiance de la variance σ^2 est l'intervalle de valeurs tel que l'on a une probabilité de $(1-\alpha)$ (fixée à l'avance) d'avoir la variance σ^2 comprise entre les 2 bornes $n\frac{s^2}{k_2}$ et $n\frac{s^2}{k_1}$:

$$P\left(n\frac{s^2}{k_2} \leq \sigma^2 \leq n\frac{s^2}{k_1}\right) = (1 - \alpha)$$

Ceci n'est strictement valable que si la moyenne μ de la population est connue.

- **Cas : μ inconnue**

Lorsque la moyenne μ est inconnue, on estime σ^2 par l'estimateur

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

On peut également montrer que :

La statistique de test : $\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = (n-1)\frac{S'^2}{\sigma^2} \sim \chi^2_{(n-1)\,d.d.l.}$

On peut alors écrire : $P(k_1 \leq (n-1)\frac{S'^2}{\sigma^2} \leq k_2) = 1 - \alpha$. où, $k_1 = \chi^2_{\frac{\alpha}{2}}$ et $k_2 = \chi^2_{1-\frac{\alpha}{2}}$ sont les fractiles de la loi khi-deux à $n-1$ degrés de liberté (cf. table du khi-deux). c'est-à-dire : $P(\chi^2_{(n-1)} \leq k_1) = \frac{\alpha}{2}$ et $P(\chi^2_{(n-1)} \leq k_2) = 1 - \frac{\alpha}{2}$.

On en déduit l'intervalle de confiance de niveau $(1 - \alpha)$ de σ^2 :

$$(n-1)\frac{s'^2}{k_2} \leq \sigma^2 \leq (n-1)\frac{s'^2}{k_1}$$

Ou encore pour l'écart-type σ :

$$\sqrt{(n-1)\frac{s'^2}{k_2}} \leq \sigma \leq \sqrt{(n-1)\frac{s'^2}{k_1}}$$

L'intervalle de confiance de la variance σ^2, lorsque la moyenne μ de la population est inconnue, est l'intervalle de valeurs tel que l'on a une probabilité de $(1-\alpha)$ (fixée à l'avance) d'avoir la variance σ^2 comprise entre les 2 bornes $(n-1)\frac{s'^2}{k_2}$ et $(n-1)\frac{s'^2}{k_1}$:

$$P\left((n-1)\frac{s'^2}{k_2} \leq \sigma^2 \leq (n-1)\frac{s'^2}{k_1}\right) = (1-\alpha)$$

Exemple 2.2.4 /Feuille 2/ On suppose que le chiffre d'affaires mensuel d'une entreprise suit une loi normale de moyenne inconnue μ mais dont l'écart-type s a été estimé à 52 K.Euros. Sur les 16 derniers mois, la moyenne des chiffres d'affaires mensuels a été de 250 K.Euros.
1 Donner une estimation ponctuelle de l'écart-type σ du chiffre d'affaires mensuel de cette entreprise.
2 Établir un intervalle de confiance de niveau 95% de σ.

2.3 Comparaisons

Il existe de nombreuses applications qui consistent, par exemple, à comparer deux groupes d'individus en regard d'un caractère particulier (poids, taille, rendement,...), ou comparer deux procédés de fabrication selon une caractéristique (résistance, diamètre, longueur,...), ou encore comparer les proportions d'apparition d'un caractère de deux populations (proportion de défectueux, proportion de gens favorisant un parti politique,...).

Les distributions d'échantillonnage qui sont alors utilisées pour effectuer ces comparaisons 'Tests d'hypothèses' ou 'calcul d'intervalles de confiance' sont celles correspondant aux fluctuations d'échantillonnage de la différence de 2 moyennes, de 2 proportions ou encore le rapport de 2 variances observées.

2.3.1 Estimation ponctuelle de la différence de 2 moyennes

On prélève des échantillons $x_1; x_2; \ldots x_n$ et $y_1; y_2; \ldots; y_p$ dans deux populations distinctes. On considère que ces échantillons sont des réalisations de v.a.r. indépendantes $X_1; X_2; \ldots; X_n$ et $Y_1; Y_2; \ldots; Y_p$ les premières de loi de probabilité L_x, les secondes de loi de probabilité L_y telles que :

2 Populations, 2 échantillons indépendants

$$\begin{cases} x_1; x_2; \ldots; x_n \leftrightarrow \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \\ y_1; y_2; \ldots; y_p \leftrightarrow \bar{Y} = \frac{1}{p}\sum_{j=1}^{p} Y_j \end{cases}$$

On suppose normales les 2 populations, avec respectivement des moyennes μ_x et μ_y, et des variances σ_x^2 et σ_y^2.

$$\begin{cases} \forall i = 1; n \\ E(X_i) = \mu_x \text{ et } V(X_i) = \sigma_x^2 \end{cases} \qquad \begin{cases} \forall j = 1; p \\ E(Y_j) = \mu_y \text{ et } V(_Y j) = \sigma_y^2 \end{cases}$$

L'estimation de la différence $(\mu_x - \mu_y)$ s'effectue par la différence des distributions d'échantillonnages des moyennes : $\bar{X} - \bar{Y}$

Estimateur sans biais : $E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_x - \mu_y$.

Estimateur convergent : $V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}$.

La différence des moyennes $(\bar{x} - \bar{y})$ observée sur les échantillons est une estimation ponctuelle de la différence $(\mu_x - \mu_y)$ des moyennes des populations.

Pour l'estimation ponctuelle de la différence de 2 proportions, la différence $(f_x - f_y)$ observée sur les échantillons est une estimation ponctuelle de la différence des proportions $(p_x - p_y)$ des populations.

2.3.2 Intervalle de confiance de la différence de 2 moyennes

- **Cas : les variances σ_x^2 et σ_y^2 sont connues**

  On sait que :

$$\text{Moyennes} : E(\bar{X}) = \mu_x \text{ et } E(\bar{Y}) = \mu_y) \Rightarrow E(\bar{X} - \bar{Y}) = \mu_x - \mu_y.$$
$$\text{Variances} : V(\bar{X}) = \frac{\sigma_x^2}{n} \text{ et } V(\bar{Y}) = \frac{\sigma_y^2}{p} \Rightarrow V(\bar{X} - \bar{Y}) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}$$

La distribution d'échantillonnage de la différence $(\bar{X} - \bar{Y})$:

$$(\bar{X} - \bar{Y}) \rightarrow \mathcal{N}\left(\mu_x - \mu_y; \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}}\right)$$

La statistique de test : $\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{p}}} \rightarrow \mathcal{N}(0;1)$

Ce qui fournit aisément un intervalle de confiance de niveau $(1-\alpha)$ pour la différence $(\mu_x - \mu_y)$:

$$(\bar{x} - \bar{y}) - t_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}} \leq \mu_x - \mu_y \leq (\bar{x} - \bar{y}) + t_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}}$$

Marge d'erreur dans l'estimation de $(\mu_x - \mu_y)$:

$$E = t_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{p}}$$

Si l'intervalle de confiance à $(1 - \alpha)\%$ pour une différence de deux moyennes ou de deux proportions (différence de risque) contient zéro, les deux moyennes ou les deux proportions ne sont pas différentes. Si l'intervalle de confiance à $(1 - \alpha)\%$ ne contient pas zéro les différences sont significativement différentes.

Exemple 2.3.1 /Feuille 2/ Le temps mis par une machine pour fabriquer une pièce est supposé suivre une loi normale de paramètres μ et σ^2. Dans un atelier, deux machines A et B fabriquent la même pièce. Pour un échantillon de 9 pièces fabriquées, on a obtenu les résultats suivants :

	Machine A	Machine B
Nombre de pièces fabriquées	9	9
Temps moyen observé (mn)	50	45
Variances des populations	25	36

1. Déterminer un intervalle de confiance, de niveau $(1-\alpha) = 95\%$, de la différence des temps moyens des deux machines $\mu_a - \mu_b$.
2. Question : La machine A est-elle aussi performante que la machine B ?

Solution :

- **Remarques :** Petits échantillons $n_A = n_B = 9$ pièces mais le temps de fabrication est supposé normalement distribué. Les variances $\sigma_A^2 = 25$ et $\sigma_B^2 = 36$ sont connues.
- Statistique de test : $\frac{\left(\bar{X}_A - \bar{X}_B\right) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim \mathcal{N}(0,1)$.
- Les données : $n_A = n_B = n = 9$.
  Niveau de confiance : $1 - \alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.
  $t_{\frac{\alpha}{2}} = t_{2.5\%} = \pm 1.96$ cf. Table de la loi normale $\mathcal{N}(0,1)$
  Marge d'erreur dans l'estimation de $(\mu_A - \mu_B)$: $E = t_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_A^2 + \sigma_B^2}{n}} = 1.96\sqrt{\frac{25+36}{9}} = 5.10mn$
  Estimation ponctuelle de la différence $(\mu_A - \mu_B)$: $\bar{x}_A - \bar{x}_B = 50 - 45 = 5mn$.
- Intervalle de confiance de niveau 95% de $(\mu_A - \mu_B)$:

$$5 - 5.10 = -0.10 \leq (\mu_A - \mu_B) \leq 5 + 5.10 = 10.10$$
$$(\mu_A - \mu_B) \in [-0.10mn, 10.10mn]$$

- Conclusion : $0 \in I.C._{95\%}$, donc la différence de 5 mn observée sur les échantillons n'est pas significative (avec un risque d'erreur de 5%), on peut donc considérer que ces deux machines ont des performances identiques.
- Question : oui, la machine B est aussi performante que la machine A, l'écart observé de 5 mn n'est pas significatif, il est dû aux fluctuations d'échantillonnage.

- **Cas : les variances σ_x^2 et σ_y^2 sont inconnues - Grands échantillons : $n \geq 30$ et $p \geq 30$**

Le cas précédant est évidemment peu courant en pratique ; voyons à présent, dans les mêmes conditions que ci-dessus, les cas les plus fréquents.

Si les échantillons prélevés dans chaque population (quelconque, par forcément normale) sont de grandes tailles alors on peut remplacer les variances inconnues σ_x^2 et σ_y^2 par leur estimation respective $s_x'^2$ et $s_y'^2$. Dans ce cas :

La distribution d'échantillonnage de la différence $(\bar{X} - \bar{Y})$ est approximativement normale.

La statistique de test : $\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x'^2}{n}+\frac{s_y'^2}{p}}} \to \mathcal{N}(0;1)$

Ce qui fournit aisément un intervalle de confiance de niveau $(1-\alpha)$ pour la différence $(\mu_x - \mu_y)$:

$$(\bar{x} - \bar{y}) - t_{\frac{\alpha}{2}} \sqrt{\frac{s_x'^2}{n} + \frac{s_y'^2}{p}} \leq \mu_x - \mu_y \leq (\bar{x} - \bar{y}) + t_{\frac{\alpha}{2}} \sqrt{\frac{s_x'^2}{n} + \frac{s_y'^2}{p}}$$

Marge d'erreur dans l'estimation de $(\mu_x - \mu_y)$:

$$E = t_{\frac{\alpha}{2}} \sqrt{\frac{s_x'^2}{n} + \frac{s_y'^2}{p}}$$

Exemple 2.3.2 /Feuille 2/ On fait subir à des cadres intermédiaires de deux grandes entreprises (une œuvrant dans la fabrication d'équipement de transport et l'autre dans la fabrication de produits électriques) un test d'appréciation et d'évaluation. La compilation des résultats pour chaque groupe à l'issue de cette évaluation s'établit comme suit :

	1 Équipement	2 Produits Électriques
Nombre de cadres	34	32
Appréciation globale moyenne	184	178
Somme des Carrés des Écarts /SCE/	15774	9858

1. Déterminer un intervalle de confiance qui a 95 chances sur 100 de contenir la valeur vraie de la différence des moyennes $(\mu_1 - \mu_2)$ des deux groupes de cadres.
2. Question : Selon cet intervalle, que peut-on conclure quant à la performance des cadres de ces deux secteurs au test d'évaluation ? Est-ce qu'en moyenne, la performance est vraisemblablement identique ou semble-t-il une différence significative entre ces deux groupes ?

Solution :

- **Remarques :** Grands échantillons $n_1 = 34$ et $n_2 = 32$ indépendants. Les variances σ_1^2 et σ_2^2 sont inconnues.
- Statistique de test : $\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1'^2}{n_1}+\frac{s_2'^2}{n_2}}} \sim \mathcal{N}(0,1)$.
- Les données :

  $n_1 = 34$ et $n_2 = 32$.

Niveau de confiance : $1-\alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.

$t_{\frac{\alpha}{2}} = t_{2.5\%} = \pm 1.96$ cf. Table de la loi normale $\mathcal{N}(0,1)$

Estimation des variances : $s_1'^2 = \frac{SCE_1}{n_1-1} = \frac{15774}{33} = 478$ et $s_2'^2 = \frac{SCE_2}{n_2-1} = \frac{9858}{31} = 318$.

Marge d'erreur dans l'estimation de $(\mu_1-\mu_2)$: $E = t_{\frac{\alpha}{2}}\sqrt{\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}} = 1.96\sqrt{\frac{478}{34} + \frac{318}{32}} = 15.6$

Estimation ponctuelle de la différence $(\mu_1 - \mu_2)$: $\bar{x}_1 - \bar{x}_2 = 184 - 178 = 6$.

- Intervalle de confiance de niveau 95% de $(\mu_1 - \mu_2)$:

$$6 - 9.6 = -3.6 \leq (\mu_1 - \mu_2) \leq 6 + 3.6 = 15.6$$
$$(\mu_1 - \mu_2) \in [-3.60, 15.60]$$

- Conclusion : $0 \in I.C._{95\%}$, donc la différence de 6 points observée sur les appréciations moyennes n'est pas significative (avec un risque d'erreur de 5%), on peut donc considérer que deux groupes de cadres ont des appréciations globales identiques.

- Question : oui, en moyenne, la performance est identique entre ces deux groupes de cadres. L'écart observé de 6 points est attribuable aux fluctuations d'échantillonnage.

• Cas : les variances sont inconnues mais supposées égales $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Petits échantillons n (et/ou) $p < 30$. Populations normales

Dans le cas de petits échantillons issus de populations normales, on ne peut pas remplacer les variances inconnues σ_x^2 et σ_y^2 par leur estimation $s_x'^2$ et $s_y'^2$ calculées sur chacun des échantillons (elles seront peu précises).

Puisqu'on les suppose égales à une valeur inconnue σ^2, on se servira de l'information des deux échantillons pour obtenir une estimation unique s'^2, de la variance $\sigma^2 = \sigma_x^2 = \sigma_y^2$:

On montre que : $S'^2 = \frac{nS_x^2 + pS_y^2}{n+p-2}$ est un bon estimateur de σ^2.

$$\text{Moyennes} : E(\bar{X} - \bar{Y}) = \mu_x - \mu_y.$$
$$\text{Variances} : V(\bar{X} - \bar{Y}) = \frac{s'^2}{n} + \frac{s'^2}{p} = s'^2\left(\frac{1}{n} + \frac{1}{p}\right)$$

La statistique de test : $\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{s'\sqrt{\frac{1}{n}+\frac{1}{p}}} \rightarrow T_{(n+p-2)\ d.d.l.}$

D'où l'intervalle de confiance de niveau $(1-\alpha)$ pour la différence $(\mu_x - \mu_y)$:

$$(\bar{x} - \bar{y}) - t_{St\frac{\alpha}{2}} s'\sqrt{\frac{1}{n} + \frac{1}{p}} \leq \mu_x - \mu_y \leq (\bar{x} - \bar{y}) + t_{St\frac{\alpha}{2}} s'\sqrt{\frac{1}{n} + \frac{1}{p}}$$

Cas particulier

Si $n = p$ (échantillons indépendants de même taille), on a plus simplement : $S'^2 = \frac{n(S_x^2+S_y^2)}{2(n-1)} =$

$\frac{SCE_x+SCE_y}{2(n-1)}$

La statistique de test : $\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{S'\sqrt{\frac{2}{n}}} \sim T_{2(n-1)\ d.d.l.}$

Les limites de l'intervalle de confiance de $(\mu_x - \mu_y)$:

$$(\bar{x} - \bar{y}) \pm t_{St\frac{\alpha}{2}} s' \sqrt{\frac{2}{n}}$$

Exemple 2.3.3 /Feuille 2/ Un laboratoire indépendant a effectué, pour le compte d'une revue sur la protection du consommateur, un essai de durée de vie sur un type d'ampoules électriques d'usage courant (60 Watts , 120 Volts) fabriquées par deux entreprises concurrentielles, dans le secteur de produits d'éclairage. Les essais effectués dans les mêmes conditions, sur un échantillon de 21 lampes provenant de chaque fabricant, donnent les résultats suivants :

La durée de vie d'une ampoule est supposée normalement distribuée.(les variances des populations sont supposées égales).

	Fabricant 1	Fabricant 2
Nombre d'essais	21	21
Durée de vie moyenne observée (h)	1025	1070
Somme des Carrés des Écarts	2400	2800

1. Déterminer un intervalle de confiance de niveau 95% de la différence des durées de vie moyennes des ampoules de ces deux fabricants.
2. Question : Est-ce que la revue peut affirmer, qu'en moyenne, les durées de vie des ampoules des deux fabricants sont identiques (ou différentes) ?
En d'autres termes, est-ce que la différence observée lors des essais est significative ?

Solution :

- **Remarques :** petits échantillons $n_1 = n_2 = n = 21$ indépendants. Les variances σ_1^2 et σ_2^2 sont inconnues mais supposées égales $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Statistique de test : $\frac{\left(\bar{X}_1-\bar{X}_2\right)-(\mu_1-\mu_2)}{s'\sqrt{\frac{2}{n}}} \sim T_{2(n-1)=40\ d.d.l.}$.
- Les données :

  $n_1 = n_2 = n = 21$.

  Niveau de confiance : $1 - \alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.

  $t_{St\frac{\alpha}{2}} = t_{2.5\%} = \pm 2.021$ cf. Table de la loi de Student à 40 d.d.l.

  Estimation de la variance commune : $nS^2 = SCE$, $s'^2 = \frac{SCE_1+SCE_2}{2(n-1)} = \frac{2400+2800}{40} = 11.40^2$.

Marge d'erreur dans l'estimation de $(\mu_1 - \mu_2)$: $E = t_{St\frac{\alpha}{2}} s' \sqrt{\frac{2}{n}} = 2.021 \times 11.40 \sqrt{\frac{2}{21}} = 7.11\ h$

Estimation ponctuelle de la différence $(\mu_1 - \mu_2)$: $\bar{x}_1 - \bar{x}_2 = 1025 - 1070 = -45\ h$.

- Intervalle de confiance de niveau 95% de $(\mu_1 - \mu_2)$:

$$-45 - 7.11 = -52.11 \leq (\mu_1 - \mu_2) \leq -45 + 7.11 = -37.89$$
$$(\mu_1 - \mu_2) \in [-52.11, -37.89\ h]$$

- Conclusion : 0 n'appartient pas à $I.C._{95\%}$, l'écart de - 45 h observé sur les durées de vie moyennes est significatif (avec un risque d'erreur de 5%). Cet écart n'est donc pas attribuable aux fluctuations d'échantillonnage.

- Question : oui, la revue doit conclure, avec un risque d'erreur de 5%, que les durées de vie des ampoules de ces deux fabricants ne sont pas identiques.

- **Cas : Échantillons appariés**

Échantillons dépendants (Données associées par paires)

Exemple 1 : On compare 2 méthodes de mesures en soumettant à ces méthodes les mêmes individus. Les 2 échantillons sont issus de deux lois différentes, mais ne sont pas indépendants (en général !).

Exemple 2 : Lorsque nous avons, pour chaque élément de l'échantillon, deux valeurs obtenues à des périodes différentes (avant / après) ou selon des traitements différents.

Donc, dans ce cas, les deux séries de mesures ne sont pas indépendantes l'une de l'autre. Il serait alors (échantillons indépendants) incorrect de procéder à un test de comparaison de moyennes tel que décrit précédemment.

On doit alors procéder comme suit avec la condition suivante : $Z_1 = (X_1 - Y_1)$; $Z_2 = (X_2 - Y_2)$; ...; $Z_n = (X_n - Y_n)$ sont indépendantes de loi $\mathcal{N}(\mu_z = \mu_x - \mu_y; \sigma_z^2 = \sigma_{x-y}^2)$: les différences de chaque paire d'observations suivent des lois normales.

On revient ainsi à un seul échantillon différence $(z_1; z_2; \ldots; z_n)$.

σ_z^2 étant généralement inconnue, on l'estime à partir :

$$S'^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(Z_i - \bar{Z}\right)^2 = \frac{SCE}{n-1}$$

On obtient, comme au paragraphe sur l'estimation par intervalle de confiance d'une moyenne μ_z lorsque la variance σ_z^2 est inconnue :

La statistique de test : $\frac{\bar{Z} - \mu_z}{S'/\sqrt{n}} \sim T_{n-1\ d.d.l.}$

On en déduit l'intervalle de confiance de niveau $(1 - \alpha)$ de $\mu_z = (\mu_x - \mu_y)$:

$$\bar{z} - t_{St\frac{\alpha}{2}} \frac{s'}{\sqrt{n}} \leq \mu_z \leq \bar{z} + t_{St\frac{\alpha}{2}} \frac{s'}{\sqrt{n}}$$

Exemple 2.3.4 /Feuille 2/ On mesure 12 pièces avec des méthodes différentes. On a obtenu les résultats suivants :

$\bar{x} = 1$; $\bar{y} = 2,08$; SCE_x /somme des carrés des écarts/ $= s_x = 106,16$; $SCE_y = s_y = 118,19$ et $SCE_{x-y} = s_{x-y} = 14,58$.

Déterminer un intervalle de confiance de niveau 95% de la différence des deux méthodes de mesures.

Solution :

- **Remarques :** Échantillons appariés (dépendants). Conditions d'application : la mesure différence $Z = X - Y$ est supposée normalement distribuée.

- Statistique de test : $\frac{(\bar{Z}-\mu_z)}{S'/\sqrt{n}} \sim T_{n-1=11\ d.d.l.}$.

- Les données :

  $n = 12 \Rightarrow \nu = n - 1 = 11\ d.d.l.$

  $\bar{z}_{12} = \bar{x}_{12} - \bar{y}_{12} = 1 - 2.08 = -1.08$: moyenne calculée sur l'échantillon différence de taille $n = 12$ (estimation ponctuelle de μ_z)

  $$s'^2_{12} = \frac{SCE_{z=x-y}}{n-1} = \frac{14.58}{11} = 1.3254 = 1.151^2$$

  Seuil de signification : $\alpha = 5\%$.

  $t_{St\frac{\alpha}{2}} = t_{2.5\%} = \pm 2.201$ cf. Table de la loi de Student à $\nu = n - 1 = 11\ d.d.l.$

- Marge d'erreur dans l'estimation de μ : $E = t_{St\frac{\alpha}{2}} \frac{s'_{12}}{\sqrt{n}} = 2.201 \frac{1.151}{\sqrt{12}} = 0.7315$.

- Intervalle de confiance de niveau 95% de μ (variance σ^2 z inconnue) :

  $$-1.08 - 0.7311 = -1.811 \leq (\mu_z = \mu_x - \mu_y) \leq -1.08 + 0.7315 = -0.3485$$
  $$\mu_z = (\mu_x - \mu_y) \in [-1.811, -0.3485]$$

- Conclusion : 0 n'appartient pas à $I.C._{95\%}$, l'écart de - 1.08 observé est significatif (avec un risque d'erreur de 5%). On peut donc conclure que $\mu_z = (\mu_x - \mu_y) \neq 0 \Leftrightarrow \mu)x \neq \mu_y$; les deux méthodes de mesures sont différentes..

- Remarque importante : Si on fait l'erreur de considérer ces deux échantillons de mesures comme des échantillons indépendants, on trouve un intervalle de confiance de niveau 95% de $(\mu_x - \mu_y) \in [-9.72; 7.56]$. Dans ce cas, $0 \in I.C._{95\%}$ c'est-à-dire que $\mu_x \approx \mu_y$; les deux méthodes de mesures sont identiques.

2.3.3 Différence de 2 proportions

Cas : Grands échantillons : $n_1 \geq 30$ **et** $n_2 \geq 30$

Il y a de nombreuses applications où nous devons décider si l'écart observé entre deux proportions échantillonnales est significatif ou s'il est plutôt attribuable au hasard de l'échantillonnage.

Comme dans le cas de la comparaison de deux moyennes, on doit connaître la distribution d'échantillonnage de la différence $(P_1 - P_2)$ des deux proportions pour estimer, par intervalle de confiance, cette différence.

On traite uniquement le cas où nous sommes en présence de grands échantillons prélevés au hasard et indépendamment de deux populations. Dans ce cas :

La statistique de test : $\frac{(F_1-F_2)-(p_-p_2)}{\sqrt{\frac{f_1(1-f_1)}{n_1}+\frac{f_2(1-f_2)}{n_2}}} \to \mathcal{N}(0;1)$

D'où l'intervalle de confiance de niveau $(1-\alpha)$ de $(p_1 - p_2)$:

$$(f_1 - f_2) - t_{\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}} \leq p_1 - p_2 \leq (f_1 - f_2) + t_{\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}$$

On peut également supposer l'hypothèse d'égalité des proportions inconnues p_1 et p_2 à une valeur commune p $(p_1 = p_2 = p)$ que l'on estime par f en combinant les proportions observées dans chaque échantillon comme suit :

$$f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

On peut donc aussi utiliser la statistique de test :

$$\frac{(F_1 - F_2) - (p_1 - p_2)}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \to N(0;1)$$

D'où l'intervalle de confiance de niveau $(1-\alpha)$ de $(p_1 - p_2)$:

$$(f_1 - f_2) - t_{\frac{\alpha}{2}}\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \leq p_1 - p_2 \leq (f_1 - f_2) + t_{\frac{\alpha}{2}}\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Exemple 2.3.5 /Feuille 2/ Dans deux municipalités avoisinantes, on a effectué un sondage pour connaître l'opinion des contribuables sur un projet d'aménagement d'un site. Les résultats de l'enquête se résument comme suit :

	Municipalité 1	Municipalité 2
Nombre de personnes interrogées	250	250
En faveur du projet	110	118

1. Quelle est l'estimation ponctuelle de la différence de proportions des contribuables de chaque municipalité favorisant l'aménagement du site ?
2. Déterminer l'intervalle de confiance de niveau $(1-\alpha) = 95\%$ de contenir la valeur vraie de la différence des proportions, $(p_1 - p_2)$?
3. Question : Avec l'intervalle calculé en 2), est-ce que l'on rejetterait, au seuil de signification $\alpha = 5\%$, l'hypothèse selon laquelle les contribuables des deux municipalités favorisent dans la même proportion l'aménagement du site sur leur territoire ?

2.3.4 Rapport de 2 variances (comparaison de 2 variances)

La comparaison de 2 populations normales peut porter non seulement sur leur valeur centrale (moyenne), mais également sur leur dispersion. La caractéristique de dispersion la plus utilisée est la variance.

Rappelons qu'une des conditions d'application de la loi de Student dans le cas de comparaison de moyennes est que les échantillons proviennent de 2 populations normales de variances identiques : $\sigma_1^2 = \sigma_2^2$. Cette hypothèse peut être maintenant vérifiée à l'aide de l'intervalle de confiance du rapport des 2 variances : **Test d'égalité de 2 variances.**

On suppose que l'on a prélevé deux échantillons indépendants de tailles n_1 et n_2 de deux populations normales $\mathcal{N}(\mu_1; \sigma_1)$ et $\mathcal{N}(\mu_2; \sigma_2)$ de paramètres inconnus.

On sait déjà que :

$$\sum_{i=1}^{n_1} \frac{(X_i - \bar{X}_1)^2}{\sigma_1^2} = (n_1 - 1)\frac{S_1'^2}{\sigma_1^2} \to \chi^2_{(n_1-1)\ d.d.l.}$$
$$\sum_{i=1}^{n_2} \frac{(X_i - \bar{X}_2)^2}{\sigma_2^2} = (n_2 - 1)\frac{S_2'^2}{\sigma_2^2} \to \chi^2_{(n_2-1)\ d.d.l.}$$

On peut alors montrer que la statistique de test :

$$\frac{\sigma_2^2}{\sigma_1^2}\frac{S_1'^2}{S_2'^2} \to \mathcal{F}_{(n_1-1),(n_2-1)\ d.d.l.}$$

On en déduit, au niveau $(1-\alpha)$, un intervalle de confiance pour le rapport $\frac{\sigma_2^2}{\sigma_1^2}$: $f_{inf}\frac{S_2'^2}{S_1'^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_{sup}\frac{S_2'^2}{S_1'^2}$
où, $f_{inf} = f_{1-\frac{\alpha}{2}} = P\left(F(n_1 - 1, n_2 - 1) > f_1\right) = 1 - \frac{\alpha}{2}$
et $f_{sup} = f_{\frac{\alpha}{2}} = P\left(F(n_1 - 1, n_2 - 1) > f_2\right) = \frac{\alpha}{2}$
sont les fractiles de la loi de Fisher-Snédécor à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté (cf. table). On recherche des limites F_{sup} et F_{inf} dans les tableau du F à $\alpha/2$ (ie risque global de α %) : F_{sup} pour un échantillonne de n_1 et de n_2 est $F^{\frac{n_1-1}{n_2-1}}$. $F_{inf} = \frac{1}{F^{\frac{n_2-1}{n_1-1}}}$.

Exemple 2.3.6 /Feuille 2/ Reprenons l'exemple de la durée de vie moyenne de 2 types d'ampoules électriques d'usage courant (60 Watts , 120 Volts) fabriquées par deux entreprises concurrentielles, dans le secteur de produits d'éclairage. Les essais effectués dans les mêmes conditions,

sur un échantillon de 21 lampes provenant de chaque fabricant, donnent les résultats suivants : La durée de vie d'une ampoule est supposée normalement distribuée. **On ne dispose d'aucune information sur les variances des deux populations.**

	Fabricant 1	Fabricant 2
Nombre d'essais	21	21
Durée de vie moyenne observée (h)	1025	1070
Somme des Carrés des Écarts	2400	2800

1. Déterminer un intervalle de confiance de niveau 95% du rapport des variances des populations d'ampoules de ces deux fabricants.
2. Question : Peut-on considérer l'égalité des variances $\sigma_2^2 = \sigma_1^2$?

Solution :

- **Remarques :** petits échantillons $n_1 = n_2 = n = 21$ indépendants.

- Statistique de test : $\frac{\sigma_2^2}{\sigma_1^2}\frac{S_1'^2}{S_2'^2} \sim F_{(n_1-1=20;n_2-1=20)\ d.d.l.}$.

- Les données :

  $n_1 = n_2 = n = 21$.

  Niveau de confiance : $1 - \alpha = 95\% \Rightarrow$ risque d'erreur : $\alpha = 5\%$.

  $f_2 = f_{\frac{\alpha}{2}} = t_{2.5\%} = 2.464$ et $f_1 = f_{97.5\%} = \frac{1}{f_2} = \frac{1}{2.464} = 0.406$ cf. Table de la loi de Fisher $F_{(20;20)}$.

  Estimation des variances : $s_1'^2 = \frac{SCE_1}{(n-1)} = \frac{2400}{20} = 120$ et $s_2'^2 = \frac{SCE_2}{(n-1)} = \frac{2800}{20} = 140$.

- Intervalle de confiance de niveau 95% de $\frac{\sigma_2^2}{\sigma_2^2}$:

$$0.474 = 0.406\frac{140}{120} = f_1\frac{s_2'^2}{s_1'^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_2\frac{s_2'^2}{s_1'^2} = 2.464\frac{140}{120} = 2.875$$
$$\frac{\sigma_2^2}{\sigma_1^2} \in [0.474, 2.875]$$

- Conclusion : $1 \in I.C._{95\%}$, il n'y a pas de différence significative (avec un risque d'erreur de 5%) entre les deux variances. On peut donc les supposer égales : $\sigma_1^2 \approx \sigma_2^2$.

2.3.5 Synthèse sur l'estimation

Estimation ponctuelle **Table 3**

Population mère P	taille N	Paramètres du caractère observé		
		moyenne μ	proportion p	variance σ^2
Echantillon	taille	Caractéristiques du caractère observé		
		moyenne	fréquence	variance
E	n	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ Série stat. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i$ D.O.1 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^*$ D.G.1	$f = \frac{n_A}{n}$	observée : $s^2 = \frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{x})^2$ empirique : $s'^2 = \frac{n}{n-1}s^2$
Estimations ponctuelles		$\widehat{\mu} = \bar{x}$	$\widehat{p} = f$	μ connue - $\widehat{\sigma}^2 = s^2$ μ inconnue - $\widehat{\sigma}^2 = s'^2$

Estimation par intervalle de confiance

Intervalle de confiance — Table 4

Paramètre estimé	Conditions	Statistique de test	Marge d'erreur E	$I.C._{(1-\alpha)}$
Moyenne μ	σ connue, p. 38, 39	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1)$	$t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
	σ inconnue $n<30$ p. 40, 41	$\frac{\bar{X}-\mu}{S'/\sqrt{n}} \to T_{n-1}$ $d.d.l.$	$t_{St\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$	$\bar{x} \pm t_{St\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$
	σ inconnue $n>30$ p. 42	$\frac{\bar{X}-\mu}{S'/\sqrt{n}} \to \mathcal{N}(0,1)$	$t_{\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$	$\bar{x} \pm t_{\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$
Proportion p	$n \geq 30$ p. 53, 54	$\frac{F-p}{\sqrt{\frac{f(1-f)}{n}}} \to \mathcal{N}(0;1)$	$t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$	$f \pm t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$
Variance σ^2 écart-type σ	μ connue p. 61	$n\frac{S^2}{\sigma^2} \to \chi^2_{n\ d.d.l.}$	n $d.d.l.$ $k_1 = \chi^2_{\frac{\alpha}{2}}$ $k_2 = \chi^2_{1-\frac{\alpha}{2}}$	$n\frac{s^2}{k_2} \leq \sigma^2 \leq n\frac{s^2}{k_1}$ $\sqrt{n\frac{s^2}{k_2}} \leq \sigma \leq \sqrt{n\frac{s^2}{k_1}}$
	μ inconnue $X \sim \mathcal{N}(\mu,\sigma)$ p. 62	$(n-1)\frac{S'^2}{\sigma^2} \to \chi^2_{(n-1)\ d.d.l.}$	$n-1$ $d.d.l.$ $k_1 = \chi^2_{\frac{\alpha}{2}}$ $k_2 = \chi^2_{1-\frac{\alpha}{2}}$	$(n-1)\frac{s'^2}{k_2} \leq \sigma^2 \leq (n-1)\frac{s'^2}{k_1}$ $\sqrt{(n-1)\frac{s'^2}{k_2}} \leq \sigma \leq \sqrt{(n-1)\frac{s'^2}{k_1}}$
	μ inconnue $n>100$ p. 63	$n\frac{S'^2}{\sigma^2} \to \mathcal{N}(n,\sqrt{2n})$	$t_{\frac{\alpha}{2}}\frac{s'^2}{2n}$ $t_{\frac{\alpha}{2}}\frac{s'}{\sqrt{2n}}$	$s'^2 \pm t_{\frac{\alpha}{2}}\frac{s'^2}{2n}$ $s' \pm t_{\frac{\alpha}{2}}\frac{s'}{\sqrt{2n}}$

Intervalle de confiance du rapport de 2 variances — Table 5

Conditions	Statistique de test	Marge d'erreur E	$I.C._{(1-\alpha)}$
$X_1 \sim \mathcal{N}(\mu_2,\sigma_2)$ $X_2 \sim \mathcal{N}(\mu_2,\sigma_2)$ p. 93 - 95	$\frac{\sigma_2^2}{\sigma_1^2}\frac{S'^2_1}{S'^2_2} \to \mathcal{F}_{(n_1-1),(n_2-1)\ d.d.l.}$	$f_{inf} = f_{1-\frac{\alpha}{2}}$ $= P\left(F(n_1-1,n_2-1) > f_{inf}\right)$ $= 1-\frac{\alpha}{2}$ $f_{sup} = f_{\frac{\alpha}{2}} =$ $P\left(F(n_1-1,n_2-1) > f_{sup}\right) = \frac{\alpha}{2}$	$f_{inf}\frac{S'^2_2}{S'^2_1} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_{sup}\frac{S'^2_2}{S'^2_1}$
Conclusion : Si $1 \in I.C._{(1-\alpha)\%}$, il n'y a pas de différence significative (avec un risque d'erreur de $\alpha\%$) entre les deux variances. On peut donc les supposer égales : $\sigma_1^2 \approx \sigma_2^2$.			

Intervalle de confiance de la différence de 2 moyennes **Table 6**

Conditions	Statistique de test	Marge d'erreur E	$I.C._{(1-\alpha)}$
σ_X^2, σ_Y^2 connues p. 68, 69	$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{p}}} \to \mathcal{N}(0;1)$	$t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{p}}$	$(\bar{X}-\bar{Y}) \pm E$
σ_X^2, σ_Y^2 inconnues $n, p \geq 30$ p. 73, 74	$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{S_x'^2}{n}+\frac{S_y'^2}{p}}} \to \mathcal{N}(0;1)$	$t_{\frac{\alpha}{2}}\sqrt{\frac{S_x'^2}{n}+\frac{S_y'^2}{p}}$	$(\bar{X}-\bar{Y}) \pm E$
$\sigma_X^2 = \sigma_Y^2 = \sigma^2$ inconnues $n, p \leq 30$ p. 78, 79	$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{s'\sqrt{\frac{1}{n}+\frac{1}{p}}} \to T_{(n+p-2)\ d.d.l.}$ $S'^2 = \frac{nS_x^2+pS_y^2}{n+p-2}$	$t_{St\frac{\alpha}{2}}s'\sqrt{\frac{1}{n}+\frac{1}{p}}$	$(\bar{X}-\bar{Y}) \pm E$
$\sigma_X^2 = \sigma_Y^2 = \sigma^2$ inconnues $n = p \leq 30$, p. 80	$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{S'\sqrt{\frac{2}{n}}} \to T_{2(n-1)\ d.d.l.}$ $S'^2 = \frac{n(S_x^2+S_y^2)}{2(n-1)}$	$t_{St\frac{\alpha}{2}}s'\sqrt{\frac{2}{n}}$	$(\bar{X}-\bar{Y}) \pm E$
Echantillons appariés p. 84, 85 $Z = X - Y$ $Z \sim \mathcal{N}(\mu_Z, \sigma_Z)$	$\frac{\bar{Z}-\mu_z}{S'/\sqrt{n}} \to T_{n-1\ d.d.l.}$ $S'^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)^2$	$t_{St\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$	$\bar{Z} \pm E$
Conclusion : Si $0 \in I.C_{(1-\alpha)} \Longrightarrow$ les deux moyennes ne sont pas différentes ; Si $0 \notin I.C_{(1-\alpha)} \Longrightarrow$ les moyennes sont significativement différentes.			

Intervalle de confiance de la différence de 2 proportions **Table 7**

Conditions	Statistique de test	Marge d'erreur E	$I.C._{(1-\alpha)}$
$n, p \geq 30$ p. 89 - 91	$\frac{(F_1-F_2)-(p-p_2)}{\sqrt{\frac{f_1(1-f_1)}{n_1}+\frac{f_2(1-f_2)}{n_2}}} \to \mathcal{N}(0;1)$	$t_{\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n_1}+\frac{f_2(1-f_2)}{n_2}}$	$(f_1-f_2) \pm E$
$n_1, n_2 \geq 30$ $p_1 = p_2 = p$ p. 89 - 91	$\frac{(F_1-F_2)-(p_1-p_2)}{\sqrt{f(1-f)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \to N(0;1)$ $f = \frac{n_1f_1+n_2f_2}{n_1+n_2}$	$t_{\frac{\alpha}{2}}\sqrt{f(1-f)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$(f_1-f_2) \pm E$
Conclusion : Si $0 \in I.C_{(1-\alpha)} \Longrightarrow$ les deux proportions ne sont pas différentes ; Si $0 \notin I.C_{(1-\alpha)} \Longrightarrow$ les proportions sont significativement différentes.			

Chapitre 3

Les tests d'hypothèse

3.1 Généralités

3.1.1 Principe d'un test d'hypothèses

Les tests d'hypothèse constituent un autre aspect important de l'inférence statistique. Le principe général d'un test d'hypothèse peut s'énoncer comme suit :

- On étudie une population dont les éléments possèdent un caractère (mesurable ou qualitatif) et dont la valeur du paramètre relative au caractère étudié est inconnue.
- Une hypothèse est formulée sur la valeur du paramètre : cette formulation résulte de considérations théoriques, pratiques ou encore elle est simplement basée sur un pressentiment.
- On veut porter un jugement sur la base des résultats d'un échantillon prélevé de cette population.

On appelle *tests d'hypothèses, tests de signification*, ou *règles de décision*, les procédés qui permettent de décider si des hypothèses sont vraies ou fausses, ou de déterminer si des échantillons observés diffèrent significativement des résultats supposés.

Il est bien évident que la statistique (c'est-à-dire la variable d'échantillonnage) servant d'estimateur au paramètre de la population ne prendra pas une valeur rigoureusement égale à la valeur théorique proposée dans l'hypothèse. Cette variable aléatoire comporte des fluctuations d'échantillonnage qui sont régies par des distributions connues.

Pour décider si l'hypothèse formulée est supportée ou non par les observations, il faut une méthode qui permettra de conclure si l'écart observé entre la valeur de la statistique obtenue dans l'échantillon et celle du paramètre spécifiée dans l'hypothèse est trop important pour être uniquement imputable au hasard de l'échantillonnage.

La construction d'un test d'hypothèse consiste en fait à déterminer entre quelles valeurs peut varier la variable aléatoire, en supposant l'hypothèse vraie, sur la seule considération du hasard de l'échantillonnage.

Les distributions d'échantillonnage d'une moyenne, d'une variance et d'une proportion que nous avons traitées dans un chapitre précédent vont être particulièrement utiles dans l'élaboration des tests statistiques.

3.1.2 Définition des concepts utiles à l'élaboration des tests d'hypothèse

Hypothèse statistique.

Une **hypothèse statistique** est un énoncé (une affirmation) concernant les caractéristiques (valeurs des paramètres, forme de la distribution des observations) d'une population.

Test d'hypothèse.

Un **test d'hypothèse** (ou test statistique) est une démarche qui a pour but de fournir une règle de décision permettant, sur la base de résultats d'échantillon, de faire un choix entre deux hypothèses statistiques.

Hypothèse nulle (H_0) et hypothèse alternative (H_1).

L'hypothèse selon laquelle on fixe à priori un paramètre de la population à une valeur particulière s'appelle l'**hypothèse nulle** et est notée H_0.

N'importe quelle autre hypothèse qui diffère de l'hypothèse H_0 s'appelle l'**hypothèse alternative** (ou contre-hypothèse) et est notée H_1.

C'est l'hypothèse nulle qui est soumise au test et toute la démarche du test s'effectue en considérant cette hypothèse comme vraie.

Dans notre démarche, nous allons établir des règles de décision qui vont nous conduire à l'acceptation ou au rejet de l'hypothèse nulle H_0. Toutefois cette décision est fondée sur une information partielle, les résultats d'un échantillon. Il est donc statistiquement impossible de prendre la bonne décision à coup sûr. En pratique, on met en œuvre une démarche qui nous permettrait, à long terme de rejeter à tort une hypothèse nulle vraie dans une faible proportion de cas. La conclusion qui sera déduite des résultats de l'échantillon aura un caractère probabiliste : on ne pourra prendre une décision qu'en ayant conscience qu'il y a un certain

risque qu'elle soit erronée. Ce risque nous est donné par le seuil de signification du test.

Seuil de signification du test

Le risque, consenti à l'avance et que nous notons α, de rejeter à tort l'hypothèse nulle H_0 alors qu'elle est vraie (favoriser alors l'hypothèse H_1), s'appelle le **seuil de signification du test et s'énonce en probabilité ainsi,**

$$\alpha = P(\text{rejeter } H_0 \,|\, H_0 \text{ vraie}) = P(\text{choisir } H_1 \,|\, H_0 \text{ vraie}).$$

A ce seuil de signification, on fait correspondre sur la distribution d'échantillonnage de la statistique une **région de rejet** de l'hypothèse nulle (appelée également **région critique**). L'aire de cette région correspond à la probabilité α.

Si par exemple on choisit $\alpha = 0.05$, cela signifie que l'on admet d'avance que la variable d'échantillonnage peut prendre, dans 5% des cas, une valeur se situant dans la zone de rejet de H_0, bien que H_0 soit vraie et ceci uniquement d'après le hasard de l'échantillonnage.

Sur la distribution d'échantillonnage correspondra aussi une région complémentaire, dite **région d'acceptation** de H_0 (ou région de non-rejet) de probabilité $1 - \alpha$.

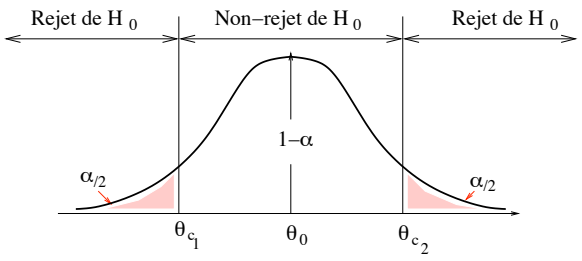
Remarque 2 1. Les seuils de signification les plus utilisés sont $\alpha = 0.05$ et $\alpha = 0.01$, dépendant des conséquences de rejeter à tort l'hypothèse H_0.

2. La valeur observée de la statistique sous l'hypothèse H_0 déduite des résultats de l'échantillon appartient, soit à la région de rejet de l'hypothèse nulle H_0 (on favorisera alors l'hypothèse H_1), soit à la région de non-rejet de H_0 (on favorisera alors l'hypothèse H_0).

Exemple 3.1.1 Supposons que nous affirmions que la valeur d'un paramètre θ d'une population est égale à la valeur θ_0. On s'intéresse au changement possible du paramètre θ dans l'une ou l'autre direction (soit $\theta > \theta_0$, soit $\theta < \theta_0$). On effectue un test bilatéral.

Les hypothèses H_0 et H_1 sont alors $\left\{ \begin{array}{ll} H_0 & \theta = \theta_0 \\ H_1 & \theta \neq \theta_0. \end{array} \right\}$

On peut schématiser les régions de rejet et de non-rejet de H_0 comme suit :



Si, suite aux résultats de l'échantillon, la valeur de la statistique utilisée se situe dans l'intervalle $[\theta_{c_1}, \theta_{c_2}]$, on acceptera H_0 au seuil de signification choisi. Si, au contraire, la valeur obtenue est supérieure à θ_{c_2} ou inférieure à θ_{c_1}, on rejette H_0 et on accepte H_1.

Remarque 3 Si on s'intéresse au changement du paramètre dans une seule direction, on opte pour un test unilatéral, en choisissant comme hypothèse H_1 soit $\theta > \theta_0$, soit $\theta < \theta_0$. La région critique est alors localisée uniquement à droite ou uniquement à gauche de la région d'acceptation.

Remarques importantes

1. Pour un test bilatéral, les 2 valeurs critiques (tables statistiques) sont des limites de la statistique qui conduisent au rejet de H_0, selon le seuil de signification α choisi.

2. Un test unilatéral "risque à droite" ou "risque à gauche" ne comporte qu'une seule valeur critique.

3. Quelle que soit le type de test, l'hypothèse nulle H_0 comporte toujours le signe égal $(=; \geq; \leq)$ et spécifie la valeur du paramètre.

4. L'hypothèse alternative H_1 est formulée en choisissant l'une ou l'autre des trois formes $(\neq; <; >)$. On choisira la plus pertinente à la situation pratique analysée.

5. Dans la plupart des tests d'hypothèses, l'inégalité dans l'hypothèse H_1 dénote dans quelle direction est localisée la région de rejet (critique) de l'hypothèse H_0.

Démarche d'un test statistique

Les principales étapes à suivre dans l'élaboration d'un test statistique sont :

1. Hypothèses statistiques,

2. Seuil de signification,

3. Condition d'application du test,

4. La statistique qui convient pour le test,

5. Règle de décision,

6. Calcul de la statistique sous H_0.

7. Décision et conclusion.

3.2 Tests permettant de déterminer si un échantillon appartient à une population donnée

3.2.1 Test sur une moyenne : comparaison d'une moyenne expérimentale à une moyenne théorique dans le cas d'un caractère quantitatif

Nous voulons déterminer si l'échantillon de taille n dont nous disposons appartient à une population de moyenne μ_0 au seuil de signification α. Nous allons dans tous les tests travailler de la même façon, en procédant en quatre étapes.

1ère étape : Formulation des hypothèses.

L'échantillon dont nous disposons provient d'une population de moyenne μ. Nous voulons savoir si $\mu = \mu_0$. On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 : $\begin{cases} H_0 & \mu = \mu_0 \\ H_1 & \mu \neq \mu_0. \end{cases}$.

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

- On détermine la statistique qui convient pour ce test. Ici, l'estimateur de la moyenne μ, c'est-à-dire $\bar{X}$, semble tout indiqué.
- On détermine la loi de probabilité de $\bar{X}$ en se plaçant sous l'hypothèse H_0. Deux cas peuvent se produire.

Premier cas : L'échantillon est de grande taille (ou bien la population est normale de variance σ^2 connue).

$\bar{X}$ suit alors une loi normale de moyenne μ_0 (puisqu'on se place sous H_0) et d'écart-type $\frac{\sigma}{\sqrt{n}}$, $\bar{X} \rightharpoonup \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$. On pose

$$T = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

T mesure un écart réduit. T est aussi appelée **fonction discriminante du test**. $T \rightharpoonup \mathcal{N}(0, 1)$.

Deuxième cas : L'échantillon est de petite taille (prélevé au hasard d'une population normale de variance σ^2 inconnue).

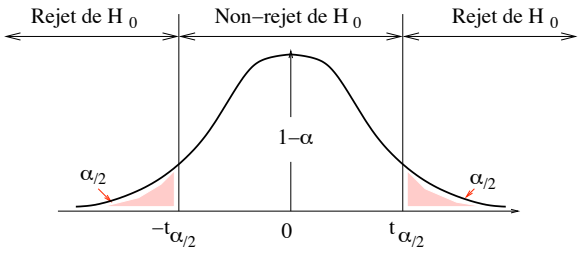
Dans ce cas la fonction discriminante du test sera

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n-1}}}.$$

Ici $T \rightharpoonup T_{n-1}$ (loi de Student à $n-1$ degrés de liberté).

3ème étape : Détermination des valeurs critiques de T délimitant les zones d'acceptation et de rejet.

On impose toujours à la zone d'acceptation de H_0 concernant l'écart réduit d'être centrée autour de 0.



Il nous faut donc déterminer dans la table la valeur maximale $t_{\alpha/2}$ de l'écart réduit imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$.

4ème étape : Calcul de la valeur de T prise dans l'échantillon et conclusion du test.

On calcule la valeur t_0 prise par T dans l'échantillon.

- Si la valeur t_0 se trouve dans la zone de rejet, on dira que l'écart-réduit observé est statistiquement significatif au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.
- Si la valeur t_0 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé n'est pas significatif au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

3.2.2 Tests sur une proportion

Nous nous proposons de tester si la proportion p d'éléments dans la population présentant un certain caractère qualitatif peut être ou non considérée comme égale à une valeur hypothétique p_0. Nous disposons pour ce faire de la proportion d'éléments possédant ce caractère dans un échantillon de taille n. Nous allons procéder comme au paragraphe précédent, en quatre étapes.

1ère étape : Formulation des hypothèses.

L'échantillon dont nous disposons provient d'une population dont la proportion d'éléments présentant le caractère qualitatif est p. Nous voulons savoir si $p = p_0$. On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 : $\begin{cases} H_0 & p = p_0 \\ H_1 & p \neq p_0. \end{cases}$

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

On détermine la statistique qui convient pour ce test. Ici, l'estimateur de la proportion p, c'est-à-dire F, semble tout indiquée.

On détermine la loi de probabilité de F **en se plaçant sous l'hypothèse** H_0. On suppose que l'on dispose d'un grand échantillon (et que "p n'est pas trop petit" (de manière que l'on ait $np \geq 15$ et $n(1-p) \geq 15$). F suit alors une loi normale de moyenne p_0 (puisqu'on se place sous H_0) et d'écart-type $\sqrt{\frac{p_0(1-p_0)}{n}}$, $F \rightharpoonup \mathcal{N}\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$.

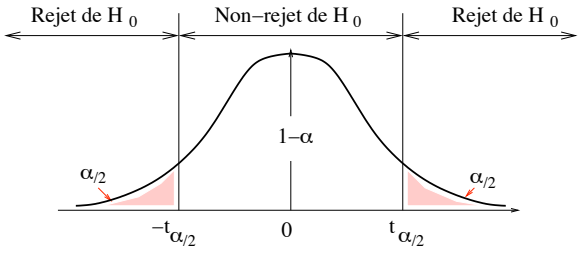
On pose

$$T = \frac{F - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

T mesure un écart réduit. T est aussi appelée fonction discriminante du test. $T \rightharpoonup \mathcal{N}(0,1)$.

3ème étape : Détermination des valeurs critiques de T délimitant les zones d'acceptation et de rejet.

On impose toujours à la zone d'acceptation de H_0 concernant l'écart réduit d'être centrée autour de 0.



Il nous faut donc déterminer dans la table la valeur maximale $t_{\alpha/2}$ de l'écart réduit imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$.

4ème étape : Calcul de la valeur de T prise dans l'échantillon et conclusion du test.

On calcule la valeur t_0 prise par T dans l'échantillon.

- Si la valeur t_0 se trouve dans la zone de rejet, on dira que l'écart-réduit observé est statistiquement significatif au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.

- Si la valeur t_0 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé n'est pas significatif au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

3.3 Risques de première et de deuxième espèce

3.3.1 Définitions

Tous les règles de décision que nous avons déterminées acceptaient un risque α qui était le risque de rejeter à tort l'hypothèse H_0, c'est-à-dire le risque de rejeter l'hypothèse H_0, alors que H_0 est vraie. Ce risque s'appelle aussi le **risque de première espèce**.

La règle de décision du test comporte également un deuxième risque, à savoir de celui de ne pas rejeter l'hypothèse nulle H_0 alors que c'est l'hypothèse H_1 qui est vraie. C'est le **risque de deuxième espèce**.

Les deux risques peuvent se définir ainsi :

α = $P(\text{rejeter } H_0 \,|\, H_0 \text{ vraie})$ = probabilité de commettre une erreur de première espèce.
β = $P(\text{ne pas rejeter } H_0 \,|\, H_1 \text{ vraie})$ = probabilité de commettre une erreur de deuxième espèce.

Risque de première espèce α **:** c'est le seuil de signification α ; risque de rejeter à tort l'hypothèse nulle H_0 lorsque celle-ci est vraie : $\alpha = P(\text{rejet } H_0 | H_0 \text{ vraie})$

Risque de deuxième espèce β **:** c'est le risque de ne pas rejeter l'hypothèse nulle H_0 alors que l'hypothèse H_1 vraie : $\beta = P(\text{non rejet } H_0 | H_1 \text{ vraie})$.

Le risque de première espèce α est choisi à priori. Toutefois le risque de deuxième espèce β dépend de l'hypothèse alternative H_1 et on ne peut le calculer que si on spécifie des valeurs particulières du paramètre dans l'hypothèse H_1 que l'on suppose vraie.

Le graphique de β en fonction des diverses valeurs du paramètre posées en H_1 s'appelle **la courbe d'efficacité du test.**

Les risques liés aux tests d'hypothèses peuvent se résumer ainsi :

		Conclusion du test	
		Accepter H_0	**Rejeter H_0**
H_0 est vraie	La décision est	**Bonne**	**Fausse**
	probabilité de prendre cette décision avant l'expérience	$1-\alpha$	α
H_0 est vraie	La décision est	**Fausse**	**Bonne**
	probabilité de prendre cette décision avant l'expérience	β	$1-\beta$

Remarque 4 La probabilité complémentaire $(1-\beta)$ du risque de deuxième espèce β définit la **puissance du test** à l'égard de la valeur du paramètre dans l'hypothèse alternative H_1. La puissance du test représente la probabilité de rejeter l'hypothèse nulle H_0 lorsque l'hypothèse vraie est H_1. Plus β est petit, plus le test est puissant.

$$1-\beta = P(\text{rejet } H_0 | H_1 \text{ vraie})$$

Le graphique de $(1-\beta)$ en fonction des diverses valeurs du paramètre posées en H_1 s'appelle **la courbe de puissance du test.**

Exemple : En contrôle industriel, le risque de 1ère espèce α correspond au risque pris par le producteur (ou fournisseur) alors que le risque de 2ème espèce β correspond au risque pris par le consommateur (ou client).

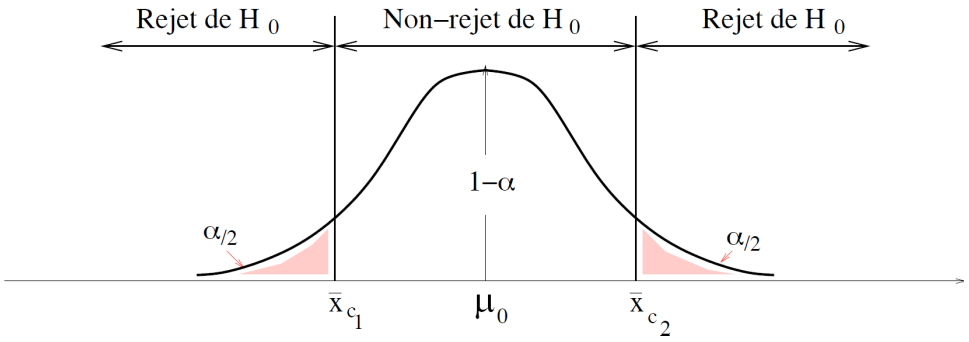
Les risques liés aux tests d'hypothèses peuvent se résumer comme suit :

Conclusion du test	Réalité : H_0 vraie	Réalité : H_1 vraie
Décision : Non-rejet H_0	bonne : $1-\alpha$	mauvaise : β
Décision : Rejet de H_0	mauvaise : α	bonne : $1-\beta$

3.3.2 Schématisation des deux risques d'erreur sur la distribution d'échantillonnage

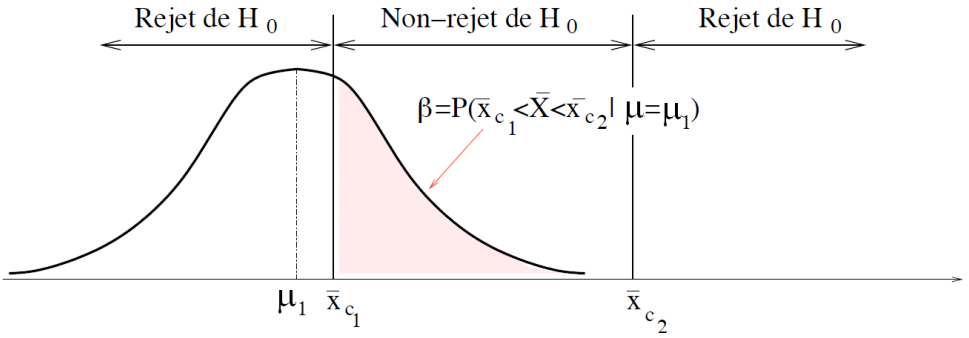
A titre d'exemple, regardons ce qu'il se passe à propos d'un test sur la moyenne. On peut visualiser sur la distribution d'échantillonnage de la moyenne comment sont reliés les deux risques d'erreur associés aux tests d'hypothèses.

Les zones d'acceptation de H_0 ($\mu = \mu_0$) et de rejet de H_0 se visualisent ainsi :

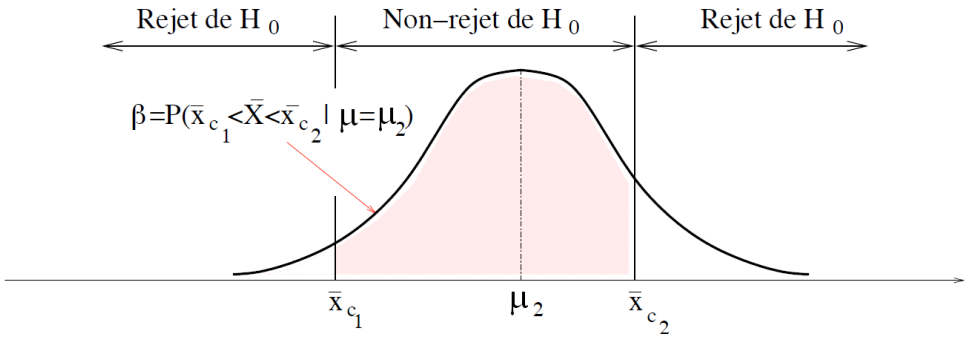


Donnons diverses valeurs à μ (autres que μ_0) que l'on suppose vraie et schématisons le risque de deuxième espèce β.

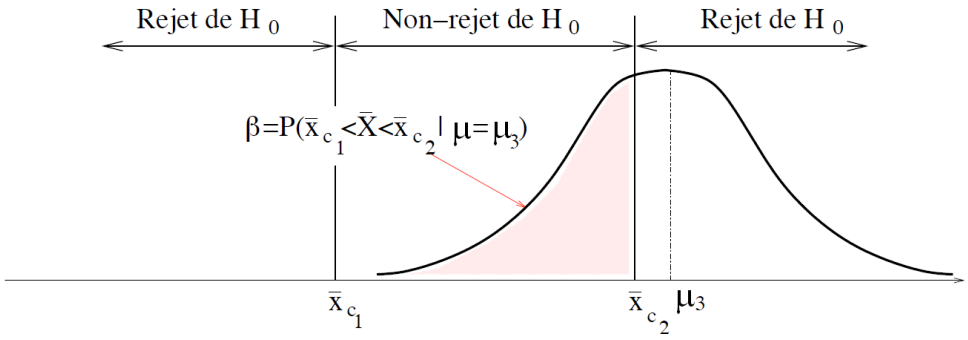
Hypothèse vraie : $\mu = \mu_1$ ($\mu_1 < \mu_0$) La distribution d'échantillonnage de $\bar{X}$ en supposant vraie $\mu = \mu_1$ est illustrée en pointillé et l'aire hachurée sur cette figure correspond à la région de non-rejet de H_0. Cette aire représente β par rapport à la valeur μ_1.



Hypothèse vraie : $\mu = \mu_2$ $(\mu_2 = \mu_0)$



Hypothèse vraie : $\mu = \mu_3$ $(\mu_3 > \mu_0)$



Cette schématisation permet d'énoncer quelques propriétés importantes concernant les deux risques d'erreur.

1. Pour un même risque α et une même taille d'échantillon, on constate que, si l'écart entre la valeur du paramètre posée en H_0 et celle supposée dans l'hypothèse vraie H_1 augmente, le risque β diminue.

2. Une réduction du risque de première espèce (de $\alpha = 0.05$ à $\alpha = 0.01$ par exemple) élargit la zone d'acceptation de H_0. Toutefois, le test est accompagné d'une augmentation du risque de deuxième espèce β. On ne peut donc diminuer l'un des risques qu'en consentant à augmenter l'autre.

3. Pour une valeur fixe de α et un σ déterminé, l'augmentation de la taille d'échantillon aura pour effet de donner une meilleure précision puisque $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ diminue. La zone d'acceptation de H_0 sera alors plus restreinte, conduisant à une diminution du risque β. Le test est alors plus puissant.

Exemple 3.3.1 Un procédé de remplissage est ajusté de telle sorte que les contenants pèsent en moyenne 400g. Le poids des contenants est supposé normalement distribué avec un écart-type de 8g. Pour vérifier si le procédé de remplissage se maintient à 400g, en moyenne, on opte

pour la règle décision suivante sur un échantillon prélevé de 16 contenants : Le processus opère correctement si : 396.08 g $\leq \bar{X} \leq$ 403.92 g Sinon arrêter le processus de remplissage.
a) Quelles sont les hypothèses statistiques que l'on veut tester avec cette méthode de contrôle ?
b) Déterminer la probabilité de commettre une erreur de première espèce.
c) Lors d'un récent contrôle, on a obtenu, pour un échantillon de 16 contenants, un poids moyen de 395g. Doit-on poursuivre ou arrêter la production ?
d) Quelle est la probabilité de commettre une erreur de deuxième espèce selon l'hypothèse alternative $H_1 : \mu = 394g$?
e) Avec ce plan de contrôle, quelle est la probabilité de rejeter l'hypothèse selon laquelle le procédé opère à 400g, alors qu'en réalité il opère à 394g ?
f) Faire de même pour les valeurs suivantes sous $H_1 : \mu =$ 395g, 396g, 397g, 398g, 399g et 400g. Tracer la courbe d'efficacité du test.

Solution :

a) Quelles sont les hypothèses statistiques que l'on veut tester avec cette méthode de contrôle ?

Hypothèses statistiques : $\begin{cases} H_0 : \bar{x} = \mu = 400 \text{ le processus est ajusté} \\ H_1 : \bar{x} \neq \mu = 400 \text{ le processus n'est pas ajusté} \end{cases}$

Seuil de signification : $\alpha = 5\%$.

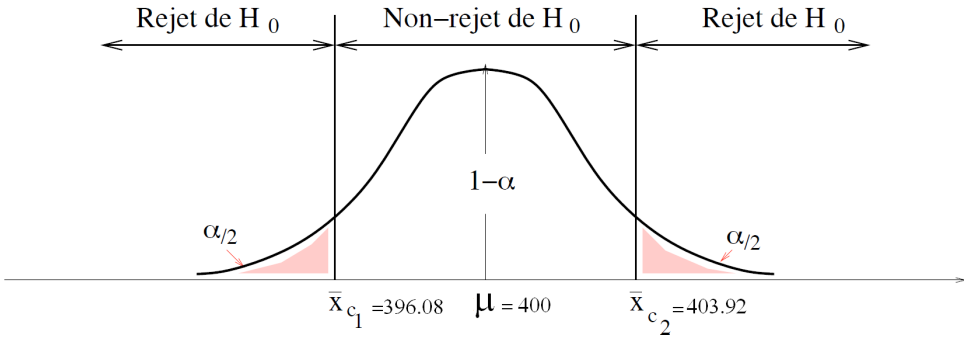
Conditions d'application du test : petit échantillon $n = 16$ provenant d'une population normale de moyenne $\mu = 400$ et écart-type $\sigma = 8$ connu. Test bilatéral du poids moyen à une moyenne connue $\mu = 400$ et un écart-type connu $\sigma = 8$.

Statistique du test : $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$

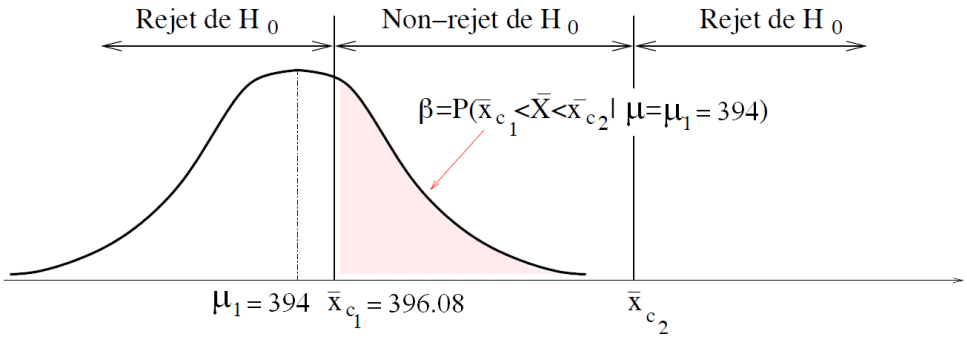
b) Déterminer la probabilité de commettre une erreur de première espèce.
Probabilité de commettre une erreur de première espèce $\alpha = 5\%$.

c) Lors d'un récent contrôle, on a obtenu, pour un échantillon de 16 contenants, un poids moyen de 395g. Doit-on poursuivre ou arrêter la production ?
$n = 16$, $\bar{x} = 395$ Comme $395 < 396.08$ on doit arrêter la production.

d) Quelle est la probabilité de commettre une erreur de deuxième espèce selon l'hypothèse alternative $H_1 : \mu = 394g$?



$\mu_1 = 394 < \mu_0 = 400$



$$\begin{aligned}
\beta &= P\left(\bar{x}_{c_1} \leq \bar{X} \leq \bar{x}_{c_2} | \mu = \mu_1 = 394\right) = P\left(\bar{X} \leq \bar{x}_{c_2}\right) - P\left(\bar{X} \leq \bar{x}_{c_1}\right) \\
&= P\left(Z \leq \frac{\bar{x}_{c_2} - \mu_1}{8}\right) - P\left(Z \leq \frac{\bar{x}_{c_1} - \mu_1}{8}\right) \\
&= P\left(Z \leq \frac{403.92 - 394}{8}\right) - P\left(Z \leq \frac{396.8 - 394}{8}\right) \\
&= P\left(Z \leq \frac{9.92}{8}\right) - P\left(Z \leq \frac{2.08}{8}\right) \\
&= P(Z \leq 1.24) - P(Z \leq 0.26) = 0.89251 - 0.60257 = 0.28994 = 28.99\%
\end{aligned}$$

$\beta = 28.99\%$

e) Avec ce plan de contrôle, quelle est la probabilité de rejeter l'hypothèse selon laquelle le procédé opère à 400g, alors qu'en réalité il opère à 394g ?

$$P(rejet \;\; H_0 | H_1 \;\; vraie) = 1 - \beta = 1 - 0.2899 = 0.7101 = 71\%$$

La puissance du test est 71%

f) Faire de même pour les valeurs suivantes sous H_1 : μ = 395g, 396g, 397g, 398g, 399g et 400g. Tracer la courbe d'efficacité du test.
La courbe d'efficacité du texte = la courbe de puissance du test : $(1 - \beta)/\mu_1$ en H_1

μ_1	$(\bar{x}_{c_1} - \mu_1)/\sigma$	$(\bar{x}_{c_2} - \mu_1)/\sigma$	$\beta = P(\bar{x}_{c_1} \leq \bar{x} \leq \bar{x}_{c_2} \mid \mu = \mu_1)$
395	0,135	1,115	0,313880650
396	0,010	0,990	0,334923560
397	-0,115	0,865	0,352258139
398	-0,240	0,740	0,365184901
399	-0,365	0,615	0,373166937
400	-0,490	0,490	0,375866103

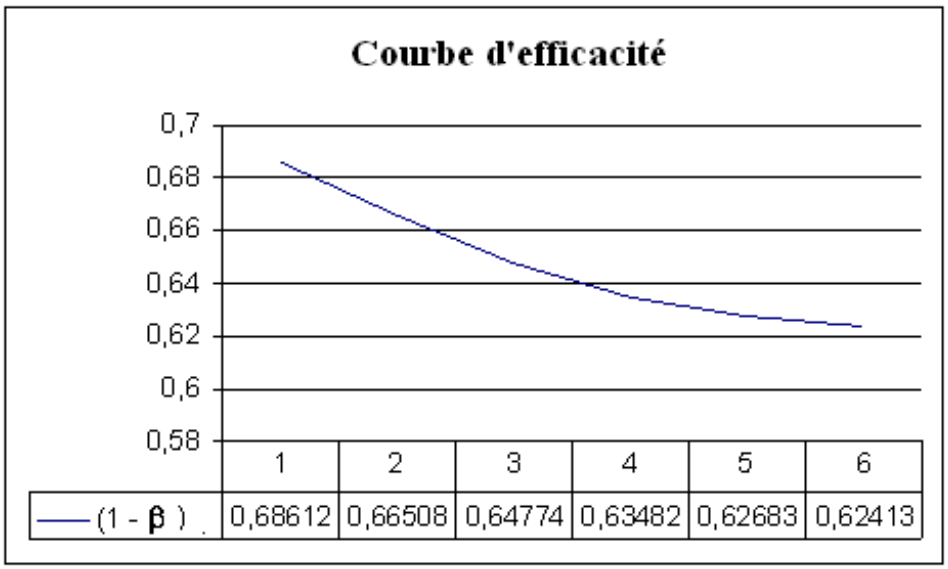


FIGURE 3.1 : Courbe d'efficacité du test

3.3.3 Exemples d'application

1. Test bilatéral

Lorsqu'on s'intéresse à l'égalité ou la différence spécifiée sous H_1 par le signe "$\neq$", on opte pour un test bilatéral.

Règle de décision pour accepter H_0 :

$$-t_{\frac{\alpha}{2}} < t_0 < t_{\frac{\alpha}{2}}$$

Exemple 3.3.2 Une entreprise fournit à un client des tiges d'acier. Le client exige que les tiges aient en moyenne, une longueur de 29 mm. On admet que la longueur des tiges est normalement distribuée. On veut vérifier si le procédé de fabrication opère bien à 29 mm. Un échantillon aléatoire de 12 tiges provenant de la fabrication donne une longueur moyenne de 27.25 mm et un écart-type empirique de 2.97 mm. Doit-on conclure, au seuil $\alpha = 5\%$, que la machine est déréglée ?
1. Hypothèses statistiques :
2. Seuil de signification :
3. Statistique de test :
4. Calcul de la statistique de test sous l'hypothèse nulle H_0 :
5. Règle de décision :

Solution

1. Hypothèses statistiques $\begin{cases} H_0 : \mu = \mu_0 = 29 \text{ (la machine n'est pas déréglée)} \\ H_1 : \mu \neq \mu_0 = 29 \text{ (la machine est déréglée)} \end{cases}$

2. Seuil de signification : $\alpha = 5\%$

3. Conditions d'application du test : petit échantillon $n = 12$ provenant d'une population normale. Test bilatéral de la longueur moyenne des tiges (variance inconnue) à une moyenne donnée $\mu_0 = 29$.

4. Statistique de test : $\frac{\bar{X}-\mu}{S'_n/\sqrt{n}} \sim T_{n-1=11 \ d.d.l.}$

5. Calcul de la statistique de test sous l'hypothèse nulle $H_0 : \mu = \mu_0 = 29$

$$t_0 = \frac{\bar{x}_{12} - \mu}{s'_{12}/\sqrt{12}} = \frac{27.25 - 29}{3.10/\sqrt{12}} = -1.954 \quad \text{avec } s'_{12} = \sqrt{\frac{12}{11}} s_{12} = \sqrt{\frac{12}{11}} 2.97 = 3.10$$

6. Règle de décision : fractiles de la loi de Sudent T_{11} (cf. table) :

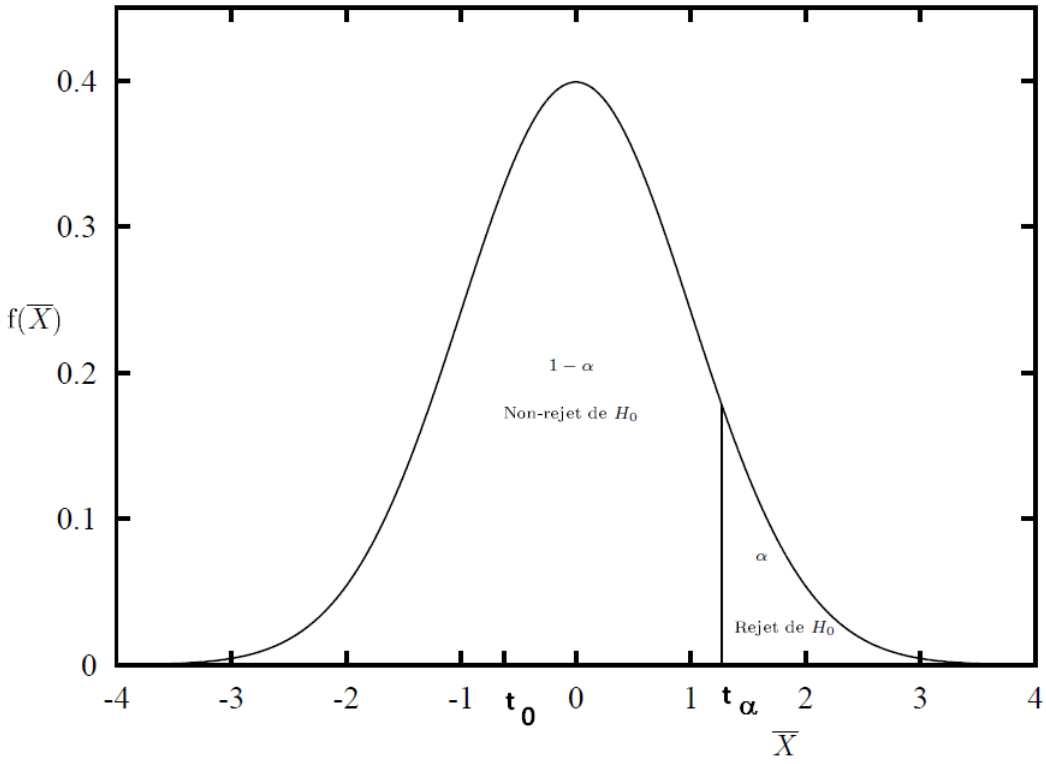
$$t_{\frac{\alpha}{2}} = 2.5\% = \pm 2.201$$

7. Décision et conclusion : t_0 appartient à la zone de non-rejet de H_0 ($-2.201 < t_0 = -1.954 < 2.201$), on peut donc conclure, avec risque d'erreur $\alpha = 5\%$ qu'il n'y a pas de différence significative. La machine semble bien réglée, il n'y a pas lieu d'intervenir.

2. Test unilatéral à droite

Lorsqu'on s'intéresse au changement d'un paramètre dans une direction spécifiée sous H_1 par le signe " $>$ ", on opte pour un test unilatéral "risque à droite".

Pour accepter l'hypothèse H_0 il faut que $t_\alpha > t_0$.



Exemple 3.3.3 Aux dernières élections, un parti politique a obtenu 42% des suffrages. Un récent sondage a révélé que, sur 1041 personnes interrogées en âge de voter, 458 accorderaient son appui à ce parti. Le secrétaire général du parti a déclaré que la popularité de son parti est en hausse.
Que penser de cette affirmation au seuil de signification $\alpha = 5\%$?
1. Hypothèses statistiques :
2. Seuil de signification :
3. Conditions d'application du test :
4. Statistique de test :
5. Calcul de la statistique de test sous l'hypothèse nulle H_0 :
6. Règle de décision :
7. Décision et conclusion :

Solution

1. Hypothèses statistiques $\begin{cases} H_0 : p = p_0 = 0.42 \ (p \leq p_0) \\ H_1 : p > p_0 = 0.42 \text{ (popularité en hausse)} \end{cases}$

2. Seuil de signification : $\alpha = 5\%$

3. Conditions d'application du test : grand échantillon ($n = 1041$) Test unilatéral "risque à droite" sur une proportion donnée $p_0 = 42\%$ au premier sondage.

Sachant que pour le second sondage, la proportion estimée est : $f = \hat{p} = \frac{458}{1041} = 44\%$

4. Statistique de test : $\frac{f-p}{\sqrt{\frac{pq}{n}}} \sim \mathcal{N}(0,1)$

5. Calcul de la statistique de test sous l'hypothèse nulle $H_0 : p = p_0 = 0.42$

$$t_0 = \frac{f - p_0}{\sqrt{\frac{p_0 - q_0}{n}}} = \frac{0.44 - 0.42}{\sqrt{\frac{0.42\,0.58}{1041}}} = 1.307$$

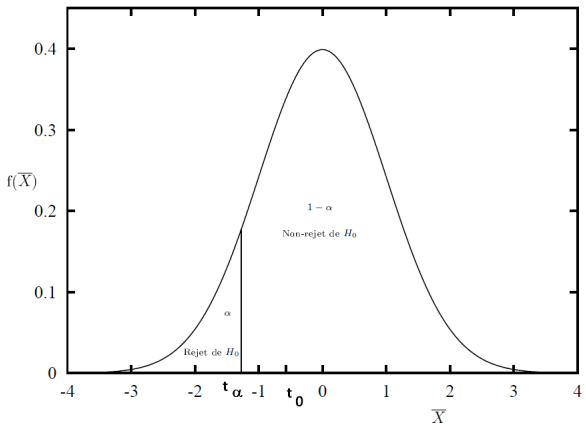
6. Règle de décision : fractile de la loi $\mathcal{N}(0,1)$ (cf. table) : $t_{5\%} = 1.645$

7. Décision et conclusion : t_0 appartient à la zone de non-rejet de H_0 ($t_{5\%} = 1.645 < t_0 = 1.307$), on peut donc conclure, avec risque d'erreur $\alpha = 5\%$ que la proportion du second sondage n'est pas significativement supérieure à celle du premier sondage. L'écart observé de 2% entre les deux sondages est dû aux fluctuations d'échantillonnage. L'affirmation du chef n'est pas justifiée statistiquement.

3. Test unilatéral à gauche

Lorsqu'on s'intéresse au changement d'un paramètre dans une direction spécifiée sous H_1 par le signe " $<$ ", on opte pour un test unilatéral "risque à gauche".

Pour accepter l'hypothèse H_0 il faut que $t_\alpha < t_0$.



Exemple 3.3.4 Le responsable de la production suggère au client des tiges d'acier avec un nouvel alliage. Il semble que ceci permettrait d'obtenir une résistance à la rupture plus élevée. Les résultats d'un test de résistance à la rupture de 50 tiges avec et sans le nouvel alliage se résument comme suit.

	Sans le nouvel alliage	Avec le nouvel alliage
Nombre de tiges	50	50
Résistance moyenne	600.50	605.00
Variance empirique	148.50	137.61

Au seuil de signification $\alpha = 5\%$, est-ce que l'hypothèse selon laquelle la résistance moyenne à la rupture sans l'alliage est moins élevée que celle avec l'alliage est confirmée ?
1. Hypothèses statistiques :
2. Seuil de signification :
3. Conditions d'application du test :
4. Statistique de test :
5. Calcul de la statistique de test sous l'hypothèse nulle H_0 :
6. Règle de décision :
7. Décision et conclusion :

Solution

1. Hypothèses statistiques $\begin{cases} H_0 : \mu_s = \mu_a \ \ (\mu_s \geq \mu_a) \\ H_1 : \mu_s < \mu_a \end{cases}$

2. Seuil de signification : $\alpha = 5\%$

3. Conditions d'application du test : grands échantillons ($n_s > 30$ et $n_a > 30$) (variances inconnues). Test unilatéral "risque à gauche".

4. Statistique de test : $\frac{\left(\bar{X}_n - \bar{Y}_p\right) - (\mu_s - \mu_a)}{\sqrt{\frac{s'^2_s}{n_s} + \frac{s'^2_a}{n_a}}} \sim \mathcal{N}(0,1)$

5. Calcul de la statistique de test sous l'hypothèse nulle $H_0 : \mu_s - \mu_a = 0$

$$t_0 = \frac{(600.5 - 605) - 0}{\sqrt{\frac{148.5}{49} + \frac{137.61}{49}}} = -1.864 \quad \text{sachant que } \frac{s'^2}{n} = \frac{s^2}{n-1}$$

6. Règle de décision : fractile de la loi $\mathcal{N}(0,1)$ (cf. table) : $t_{5\%} = -1.645$

7. Décision et conclusion : t_0 appartient à la zone de rejet de H_0 ($t_0 = -1.864 < t_{5\%} = 1.645$), on peut donc conclure, avec risque d'erreur $\alpha = 5\%$ qu'il y a une différence significative. La résistance moyenne à la rupture sans alliage est significativement plus petite que celle avec alliage.

3.4 Comparaisons. Tests permettant de déterminer si deux échantillons appartiennent à la même population

Introduction

Il existe de nombreuses applications qui consistent, par exemple, à comparer deux groupes d'individus en regard d'un caractère quantitatif particulier (poids, taille, rendement scolaire, quotient intellectuel,....) ou à comparer deux procédés de fabrication selon une caractéristique quantitative particulière (résistance à la rupture, poids, diamètre, longueur,...) ou encore de comparer les proportions d'apparition d'un caractère qualitatif de deux populations (proportion de défectueux, proportion de gens favorisant un parti politique,...). Les variables aléatoires qui sont alors utilisées pour effectuer des tests d'hypothèses (ou aussi calculer des intervalles de confiance) sont la **différence des moyennes** d'échantillon, le **quotient des variances** d'échantillon ou la **différence des proportions** d'échantillon.

3.4.1 Comparaison de deux moyennes d'échantillon : "test T"

Nous nous proposons de tester si la moyenne de la première population (μ_1) peut être ou non considérée comme égale à la moyenne de la deuxième population (μ_2). Nous allons alors comparer les deux moyennes d'échantillon $\bar{x}_1$ et $\bar{x}_2$. Il est évident que si $\bar{x}_1$ et $\bar{x}_2$ diffèrent beaucoup, les deux échantillons n'appartiennent pas la même population. Mais si $\bar{x}_1$ et $\bar{x}_2$ diffèrent peu, il se pose la question de savoir si l'écart $d = \bar{x}_1 - \bar{x}_2$ peut être attribué aux hasards de l'échantillonnage. Afin de donner une réponse rigoureuse à cette question, nous procéderons encore en quatre étapes.

1ère étape : Formulation des hypothèses.

Le premier échantillon dont nous disposons provient d'une population dont la moyenne est μ_1. Le deuxième échantillon dont nous disposons provient d'une population dont la moyenne est μ_2.

Nous voulons savoir si il s'agit de la même population en ce qui concerne les moyennes, c'est-à-dire si $\mu_1 = \mu_2$. On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 : $\begin{cases} H_0 & \mu_1 = \mu_2 \\ H_1 & \mu_1 \neq \mu_2. \end{cases}$

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

On détermine la statistique qui convient pour ce test. Ici, la différence $D = \bar{X}_1 - \bar{X}_2$ des deux moyennes d'échantillon, semble tout indiquée.

On détermine la loi de probabilité de D **en se plaçant sous l'hypothèse** H_0. On suppose que l'on dispose de grands échantillons ($n_1 \geq 30$ et $n_2 \geq 30$). $\bar{X}_1$ suit alors une loi normale de moyenne μ_1 et d'écart-type $\frac{\sigma_1}{\sqrt{n_1}}$ que l'on peut sans problème estimer par $\frac{s_1}{\sqrt{n_1 - 1}}$ (car $n_1 \geq 30$).

I.e. $\bar{X}_1 \rightharpoonup \mathcal{N}(\mu_1, \frac{s_1}{\sqrt{n_1-1}})$.

De même $\bar{X}_2$ suit alors une loi normale de moyenne μ_2 et d'écart-type $\frac{\sigma_2}{\sqrt{n_2}}$ que l'on peut sans problème estimer par $\frac{s_2}{\sqrt{n_2-1}}$ (car $n_2 \geq 30$). I.e. $\bar{X}_2 \rightharpoonup \mathcal{N}(\mu_2, \frac{s_2}{\sqrt{n_2-1}})$.

On en déduit, puisque $\bar{X}_1$ et $\bar{X}_2$ sont indépendantes que D suit également une loi normale.

$E(D) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2 = 0$ puisqu'on se place sous H_0.

$E(D) = V(\bar{X}_1) + V(\bar{X}_2) = \dfrac{s_1^2}{n_1 - 1} + \dfrac{s_2^2}{n_2 - 1}$ puisque les variables sont indépendantes.

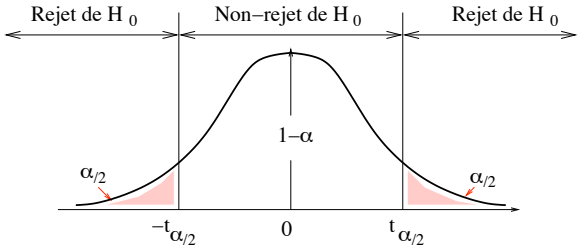
On pose

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}}}.$$

T mesure un écart réduit. T est la fonction discriminante du test $T \rightharpoonup \mathcal{N}(0, 1)$.

3ème étape : Détermination des valeurs critiques de T délimitant les zones d'acceptation et de rejet.

On impose toujours à la zone d'acceptation de H_0 concernant l'écart réduit d'être centrée autour de 0.



Il nous faut donc déterminer dans la table la valeur maximale $t_{\alpha/2}$ de l'écart réduit imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$.

4ème étape : Calcul de la valeur de T prise dans l'échantillon et conclusion du test.

On calcule la valeur t_0 prise par T dans l'échantillon.

- Si la valeur t_0 se trouve dans la zone de rejet, on dira que l'écart-réduit observé est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.
- Si la valeur t_0 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

Remarque 5 *Si on travaille sur de petits échantillons, si la loi suivie par la grandeur est une loi normale et si on ignore les écarts-type des populations, on doit utiliser la loi de Student.*

3.4.2 Comparaison de deux variances d'échantillon : "test F"

1ère étape : Formulation des hypothèses.

Le premier échantillon dont nous disposons provient d'une population dont l'écart-type est σ_1. Le deuxième échantillon dont nous disposons provient d'une population dont l'écart-type est σ_2. Nous voulons savoir si il s'agit de la même population en ce qui concerne les écarts-type, c'est-à-dire si $\sigma_1 = \sigma_2$. On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 :
$\begin{cases} H_0 & \sigma_1 = \sigma_2 \\ H_1 & \sigma_1 \neq \sigma_2. \end{cases}$

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

On détermine la statistique qui convient pour ce test. Ici, la variable aléatoire dont on connaît la loi est le rapport $F = \dfrac{S_1^2}{S_2^2}$ où S_1^2 et S_2^2 sont les variables aléatoires variances d'échantillon.

On détermine la loi de probabilité de F **en se plaçant sous l'hypothèse** H_0.

On suppose ici que les deux populations dont nous avons tiré les échantillons sont normales. Il en découle que

- $\dfrac{(n_1 - 1)S_1^2}{\sigma_1^2}$ suit la loi du khi-deux à $n_1 - 1$ degrés de liberté.
- De même, $\dfrac{(n_2 - 1)S_2^2}{\sigma_2^2}$ suit la loi du khi-deux à $n_2 - 1$ degrés de liberté.

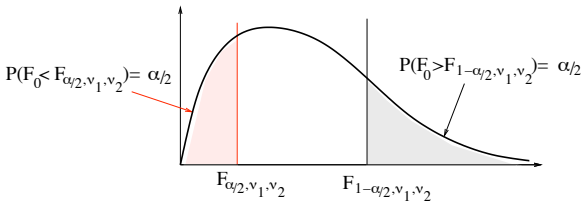
On considère alors le quotient

$$F_0 = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$$

qui est distribué suivant la loi de Fisher avec $\nu_1 = n_1 - 1$ et $\nu_2 = n_2 - 1$ degrés de liberté. Lorsqu'on se place sous l'hypothèse H_0, c'est le rapport $F_0 = \dfrac{S_1^2}{S_2^2}$ qui suit la loi de Fisher avec ν_1 et ν_2 degrés de liberté puisque $\sigma_1 = \sigma_2$. Ici la **fonction discriminante du test** est F_0.

3ème étape : Détermination des valeurs critiques de F_0 délimitant les zones d'acceptation et de rejet.

On impose maintenant à la zone d'acceptation de H_0 concernant le quotient des deux variances d'échantillon d'être centrée autour de 1.



On détermine dans les tables les deux valeurs $F_{\alpha/2,\nu_1,\nu_2}$ et $F_{1-\alpha/2,\nu_1,\nu_2}$ telles que $P(F_{\alpha/2,\nu_1,\nu_2} < F_0 < F_{1-\alpha/2,\nu_1,\nu_2} = 1-\alpha$.

On rejettera H_0 si la valeur f_0 prise par F_0 dans l'échantillon se trouve à l'extérieur de l'intervalle $[F_{\alpha/2,\nu_1,\nu_2}, F_{1-\alpha/2,\nu_1,\nu_2}]$.

Remarque 6 *On notera que pour obtenir la valeur critique inférieure de F_0, on doit utiliser la relation*

$$F_{1-\alpha/2,\nu_1,\nu_2} = \frac{1}{F_{\alpha/2,\nu_2,\nu_1}}.$$

4ème étape : Calcul de la valeur de F_0 prise dans l'échantillon et conclusion du test.

On calcule la valeur f_0 prise par F_0 dans l'échantillon.

- Si la valeur F_0 se trouve dans la zone de rejet, on dira que la valeur observée pour F est **statistiquement significative** au seuil α. Ce quotient est éloigné de 1 et ne permet pas d'accepter H_0. On rejette H_0.
- Si la valeur F_0 se trouve dans la zone d'acceptation, on dira que la valeur observée pour F **n'est pas significative** au seuil α. L'écart constaté par rapport à la valeur 1 attendue est imputable aux fluctuations d'échantillonnage. On accepte H_0.

3.4.3 Comparaison de deux proportions d'échantillon

Il y a de nombreuses applications (échéances électorales, expérimentations médicales...) où nous devons décider si l'écart observé entre deux proportions échantillonnales est significatif où s'il est attribuable au hasard de l'échantillonnage. Pour répondre à cette question, nous procéderons comme d'habitude en quatre étapes.

1ère étape : Formulation des hypothèses.

Le premier échantillon dont nous disposons provient d'une population 1 dont les éléments possèdent un caractère qualitatif dans une proportion inconnue p_1. Le deuxième échantillon dont nous disposons provient d'une population 2 dont les éléments possèdent le même caractère qualitatif dans une proportion inconnue p_2.

Nous voulons savoir si il s'agit de la même population en ce qui concerne les proportions, c'est-à-dire si $p_1 = p_2$. On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 : $\begin{cases} H_0 & p_1 = p_2 \\ H_1 & p_1 \neq p_2 \end{cases}$.

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

Nous traiterons uniquement le cas où nous sommes en présence de grands échantillons.

On détermine la statistique qui convient pour ce test. Ici, la différence $D = F_1 - F_2$ des deux proportions d'échantillon, semble tout indiquée, puisque F_1 est un estimateur sans biais de p_1 et F_2 un estimateur sans biais de p_2.

On détermine la loi de probabilité de D **en se plaçant sous l'hypothèse** H_0. F_1 suit alors une loi normale de moyenne p_1 et d'écart-type $\sqrt{\frac{p_1(1-p_1)}{n_1}}$.

De même, F_2 suit alors une loi normale de moyenne p_2 et d'écart-type $\sqrt{\frac{p_2(1-p_2)}{n_2}}$.

On en déduit, puisque F_1 et F_2 sont indépendantes que D suit également une loi normale.

$E(D) = E(F_1) - E(F_2) = p_1 - p_2 = 0$ puisqu'on se place sous H_0.

$V(D) = V(F_1) + V(F_2) = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}$ puisque les variables sont indépendantes. Ici, on a posé $p_1 = p_2 = p$ puisque l'on se place sous H_0.

Mais comment trouver p puisque c'est justement sur p que porte le test ? Puisque nous raisonnons en supposant l'hypothèse H_0 vraie, on peut considérer que les valeurs de F_1 et F_2 obtenues sur nos échantillons sont des approximations de p. De plus, plus la taille de l'échantillon est grande, meilleure est l'approximation (revoir le chapitre sur les intervalles de confiance). Nous allons donc pondérer les valeurs observées dans nos échantillons par la taille respective de ces échantillons. On approchera p dans notre calcul par $\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$.

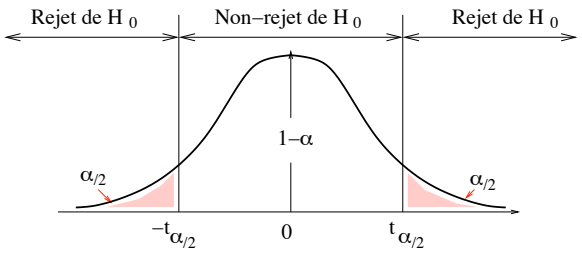
On pose

$$T = \frac{D}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}.$$

T mesure un écart réduit. T est la **fonction discriminante du test**. $T \rightharpoonup \mathcal{N}(0,1)$.

3ème étape : Détermination des valeurs critiques de T délimitant les zones d'acceptation et de rejet

On impose toujours à la zone d'acceptation de H_0 concernant l'écart réduit d'être centrée autour de 0.



Il nous faut donc déterminer dans la table la valeur maximale $t_{\alpha/2}$ de l'écart réduit imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$.

4ème étape : Calcul de la valeur de T prise dans l'échantillon et conclusion du test

On calcule la valeur t_0 prise par T dans l'échantillon.

- Si la valeur t_0 se trouve dans la zone de rejet, on dira que l'écart-réduit observé est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.
- Si la valeur t_0 se trouve dans la zone d'acceptation

$$-t_{\frac{\alpha}{2}} < t_0 < t_{\frac{\alpha}{2}},$$

on dira que l'écart-réduit observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

Exemple 3.4.1 Pour sa fabrication, un industriel utilise des pièces de deux constructeurs différents. Après six mois d'utilisation, il constate que sur les 80 pièces du constructeur 1, 50 ne sont jamais tombées en panne, alors que pour le constructeur 2 la proportion est de 40 sur 60. Au seuil de signification $\alpha = 5\%$, peut-on considérer que les proportions de pièces de ces deux constructeurs sont équivalentes ?

1. Hypothèses statistiques :
2. Seuil de signification :
3. Conditions d'application du test :
4. Statistique de test :
5. Calcul de la statistique de test sous l'hypothèse nulle H_0 :
6. Règle de décision :
7. Décision et conclusion :

Solution

1. Hypothèses statistiques $\begin{cases} H_0 : p_1 = p_2 \text{ (équivalentes)} \\ H_1 : p_1 \neq p_2 \text{ (différentes)} \end{cases}$

2. Seuil de signification : $\alpha = 5\%$

3. Conditions d'application du test : grands échantillons ($n_1 > 30$ et $n_2 > 30$). Test bilatéral symétrique.

4. Statistique de test : $\frac{(F_1 - F_2) - (p_1 - p_2)}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0,1)$
   avec $f_1 = 62.50\%$; $f_2 = 66.67\%$; $f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = 64.28\%$ et $(1 - f) = 35.71\%$.

5. Calcul de la statistique de test sous l'hypothèse nulle $H_0 : p_1 - p_2 = 0$

$$t_0 = \frac{(0.6250 - 0.6667) - 0}{\sqrt{0.6428 \times 0.3571\left(\frac{1}{80} + \frac{1}{60}\right)}} = -0.513$$

6. Règle de décision : fractile de la loi $\mathcal{N}(0,1)$ (cf. table) : $t_{2.5\%} = \pm 1.96$

7. Décision et conclusion : t_0 appartient à la zone de non-rejet de H_0 $(-1.96 < t_0 = -0.513 < -1.96)$, on peut conclure, avec risque d'erreur $\alpha = 5\%$, qu'il n'y a pas de différence significative entre ces deux proportions.
   On peut donc les considérer comme équivalentes.

3.5 Tests non-paramétriques

On qualifie de non-paramétriques "distribution free" les tests statistiques qui sont construits à partir d'une fonction des observations sur un échantillon aléatoire, fonction dont la loi de probabilité ne dépend pas de la connaissance de la distribution de la population-mère.

La validité des tests non-paramétriques dépend seulement d'un nombre très restreint de conditions d'application (échantillons considérés doivent être aléatoires et simples) beaucoup moins contraignantes que celles requises pour la mise en œuvre des tests paramétriques (distribution normale de la population-mère ou échantillon de grande taille).

Un test non-paramétrique présente quelques avantages :

1. son application est relativement facile et rapide,
2. s'applique à des échantillons de petites tailles,
3. s'applique à des caractères qualitatifs, à des grandeurs de mesure, à des ratios, à des rangs de classement, etc.

On distinguera principalement les deux familles suivantes :

1. Test du Khi-deux de Pearson :
   (a) Test d'ajustement ou d'adéquation entre deux distributions.
   (b) Test d'indépendance dans un tableau de contingence.
   (c) Test d'homogénéité de plusieurs populations.
2. Tests appliqués aux rangs et aux signes
   (a) Test de la somme des rangs (Wilcoxon et Mann-Withney)
   (b) Test de signes
   (c) Test de la somme des rangs des différences positives (Wilcoxon)
   (d) Test d'indépendance de rangs de Spearman

Une caractéristique essentielle des méthodes non-paramétriques est leur relative **simplicité** et **rapidité des calculs.**

Remplacer les valeurs observées par des indicateurs ou des rangs provoque évidemment une certaine perte d'information. De ce fait, **les tests non-paramétriques** sont généralement **moins puissants** que **les tests paramétriques**, beaucoup **plus robustes.**

3.5.1 Test d'ajustement de deux distributions : "test du khi-deux"

Introduction

Le **test de Pearson**, appelé aussi le **test du khi-deux** est un outil statistique qui permet de vérifier la concordance entre une distribution expérimentale et une distribution théorique.

On cherche donc à déterminer si un modèle théorique est susceptible de représenter adéquatement le comportement probabiliste de la variable observée, comportement fondé sur les fréquences des résultats obtenus sur l'échantillon.

Comment procéder ?

Répartitions expérimentales

On répartit les observations suivant k classes (si le caractère est continu) ou k valeurs (si le caractère est discret). On dispose alors des effectifs des k classes : n_1, $n_2, \ldots, n_k$. On a bien sûr la relation

$$\sum_{i=1}^{k} n_i = N,$$

où N est le nombre total d'observations effectuées.

Répartitions théoriques

En admettant comme plausible une distribution théorique particulière, on peut construire une répartition idéale des observations de l'échantillon de taille N en ayant recours aux probabilités tablées (ou calculées) du modèle théorique : p_1, $p_2, \ldots, p_k$. On obtient alors les effectifs théoriques $n_{t,i}$ en écrivant $n_{t,i} = Np_i$. On dispose automatiquement de la relation $\sum_{i=1}^{k} n_{t,i} = N$.

Remarque 7 *Dans la pratique, on se placera dans le cas où $N \geq 50$ et où chaque $n_{t,i}$ est supérieur ou égal à 5. Si cette condition n'est pas satisfaite, il y a lieu de regrouper deux ou plusieurs classes adjacentes. Il arrive fréquemment que ce regroupement s'effectue sur les classes aux extrémités de la distribution. k représente donc le nombre de classes après regroupement.*

Définition de l'écart entre les deux distributions

Pour évaluer l'écart entre les effectifs observés n_i et les effectifs théoriques $n_{t,i}$, on utilise la somme des écarts normalisés entre les deux distributions, à savoir

$$\chi^2 = \frac{(n_1 - n_{t,1})^2}{n_{t,1}} + \frac{(n_2 - n_{t,2})^2}{n_{t,2}} + \cdots + \frac{(n_k - n_{t,k})^2}{n_{t,k}}.$$

Plus le nombre χ^2 ainsi calculé est grand, plus la distribution étudiée différer de la distribution théorique.

Quelques considérations théoriques à propos de cet écart

Le nombre d'observations n_i parmi l'échantillon de taille N susceptible d'appartenir à la classe i est la réalisation d'une variable binomiale N_i de paramètres N et p_i (chacune des N observations appartient ou n'appartient pas à la classe i avec une probabilité p_i). Si N est suffisamment grand (on se place dans le cas d'échantillons de taille 50 minimum) et p_i pas trop petit (on a effectué des regroupements de classes pour qu'il en soit ainsi), on peut approcher la loi binomiale par la loi normale, c'est-à-dire $\mathcal{B}(N, p_i)$ par $\mathcal{N}(Np_i, \sqrt{Np_i(1-p_i)})$. Pour simplifier, on approxime $Np_i(1-p_i)$ par Np_i. Donc $\frac{N_i - Np_i}{Np_i}$ suit la loi $\mathcal{N}(0,1)$. Lorsqu'on élève au carré toutes ces quantités et qu'on en fait la somme, on obtient une somme de k lois normales centrées réduites (presque) indépendantes. Nous avons vu au chapitre 3 que cette somme suivait une loi du khi-deux.

Mais quel est le nombre de degrés de liberté de cette variable du khi-deux ?

Il y a k carrés, donc à priori k degrés de liberté. Mais on perd toujours un degré de liberté car on a fixé l'effectif total de l'échantillon,

$$\sum_{i=1}^{k} N_i = N.$$

On peut perdre d'autres degrés de liberté si certains paramètres de la loi théorique doivent être estimés à partir de l'échantillon.

1. Si la distribution théorique est entièrement spécifiée, c'est-à-dire si on cherche à déterminer si la distribution observée suit une loi dont les paramètres sont connus avant même de choisir l'échantillon, on a $k-1$ degrés de liberté (k carrés indépendants moins une relation entre les variables).

2. S'il faut d'abord estimer r paramètres de la loi à partir des observations de l'échantillon (par exemple on cherche si la distribution est normale mais on ne connaît d'avance ni sa moyenne ni son écart-type), il n'y a plus que $k - 1 - r$ degrés de liberté.

Dans le cas général, on dira que la loi du khi-deux suivie par l'écart entre les deux distributions a $k - 1 - r$ degrés de liberté lorsqu'on a estimé r paramètres de la loi théorique à partir des observations de l'échantillon (avec la possibilité pour r de valoir 0).

EXCEL : $CHITEST(y_{\text{observés}}; y_{\text{estimés}}) = p$

Si $p > \alpha$ on accepte l'hypothèse H_0.

Le test d'ajustement de Pearson

Il nous faut maintenant décider, à l'aide de cet indicateur qu'est le χ^2, si les écarts entre les effectifs théoriques et ceux qui résultent des observations sont significatifs d'une différence de distribution ou si ils sont dus aux fluctuations d'échantillonnage. Nous procéderons comme d'habitude en quatre étapes.

1ère étape : Formulation des hypothèses.

On va donc tester l'hypothèse H_0 contre l'hypothèse H_1 :

$$\begin{cases} H_0 & \text{Les observations suivent la distribution théorique spécifiée,} \\ H_1 & \text{Les observations ne suivent pas la distribution théorique spécifiée.} \end{cases}$$

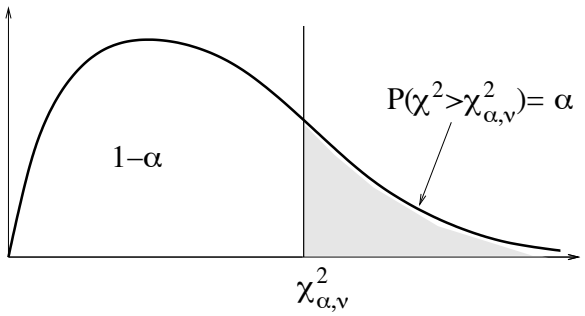
2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

On utilise la variable aléatoire

$$\chi^2 = \frac{(n_1 - n_{t,1})^2}{n_{t,1}} + \frac{(n_2 - n_{t,2})^2}{n_{t,2}} + \cdots + \frac{(n_k - n_{t,k})^2}{n_{t,k}}.$$

3ème étape : Détermination des valeurs critiques de χ^2 délimitant les zones d'acceptation et de rejet.

On impose à la zone d'acceptation de H_0 concernant la valeur du χ^2 d'être un intervalle dont 0 est la borne inférieure (car un χ^2 est toujours positif).



Il nous faut donc déterminer dans la table la valeur maximale $\chi^2_{\alpha,\nu}$ de l'écart entre les deux distributions imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant $P(\chi^2 > \chi^2_{\alpha,\nu}) = \alpha$. $\chi^2_{\alpha,\nu}$ représente donc la valeur critique pour un test sur la concordance entre deux distributions et le test sera toujours unilatéral à droite.

4ème étape : Calcul de la valeur de χ^2 prise dans l'échantillon et conclusion du test.

On calcule la valeur χ_0^2 prise par χ^2 dans l'échantillon.

- Si la valeur χ_0^2 se trouve dans la zone de rejet, on dira que l'écart observé entre les deux distributions est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.

- Si la valeur χ_0^2 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

EXCEL : $CHITEST(y_{\text{observés}}; y_{\text{estimés}}) = p$

Si $p > \alpha$ on accepte l'hypothèse H_0.

3.5.2 Test d'indépendance du khi-deux

Le test de khi-deux est fréquemment utilisé pour tester si deux caractères, qualitatifs ou quantitatifs (répartis en classes), observés dans une population sont indépendants ou si, au contraire, ils sont dépendants : présentent un certain degré d'association (liaison).

- **Principe général du test :**

1. Un échantillon aléatoire de taille n est prélevé d'une population et est observé selon deux caractères X à p modalités et Y à q modalités.

2. La répartition des n observations suivant les modalités croisées des deux caractères se présente sous la forme d'un tableau à double entrée appelé tableau de contingence.

3. Il s'agit par la suite de tester, à l'aide du khi-deux de Pearson, si les deux caractères sont indépendants ou non.

- **Tableau de contingence. Tableau des effectifs observés :**

	y_1	...	y_j	...	y_l	Total ligne
x_1	n_{11}	...	n_{1j}	...	n_{1l}	$n_{1.} = \sum_j n_{1j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	...	n_{ij}	...	n_{il}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	...	n_{kj}	...	n_{kl}	$n_{k.}$
Total colonne	$n_{.1} = \sum_i n_{ij}$	...	$n_{.j}$	...	$n_{.l}$	$n = n.. = \sum_i \sum_j n_{ij}$

- **Les hypothèses statistiques** peuvent s'énoncer ainsi :

$$\begin{cases} H_0 : \text{les caractères } X \text{ et } Y \text{ sont indépendants} \\ H_1 : \text{les caractères } X \text{ et } Y \text{ sont dépendants} \end{cases}$$

- **Sous l'hypothèse nulle H_0 :** indépendance des deux caractères, on a,

$$p_{ij} = p_{i.}p_{.j} \quad \forall (i = 1, k \text{ et } j = 1, l) \text{ (probabilités conjointes } p_{ij} = \frac{n_{ij}}{n}).$$

- l'estimation des effectifs théoriques s'obtient en répartissant la taille de l'échantillon n dans les proportions obtenues selon les estimations des probabilités conjointes

$$\text{(indépendance en probabilité) :} \quad f_{ij} = \frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{\hat{n}_{ij}}{n} \quad \text{d'où,} \quad \hat{n}_{ij} = \frac{n_{i.}n_{.j}}{n}$$

- Pour comparer les répartitions théorique et observée, on calcule, sous l'hypothèse nulle H_0 la quantité :

$$\chi^2_{calculé} = \sum_i^k \sum_j^l \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

laquelle sous H_0 est distribuée selon la loi du khi-deux $\chi^2_{(k-1)(l-1)\ d.d.l.}$: noté χ^2_{table} pour le risque d'erreur α choisi.

- Décision et conclusion du test statistique :

L'hypothèse nulle H_0 d'indépendance est rejetée, au niveau α, si $\chi^2_{calculé} \geq \chi^2_{table}$ (le test statistique est toujours unilatéral).

EXCEL : $CHITEST(y_{\text{observés}}; y_{\text{estimés}}) = p$

Si $p > \alpha \Longrightarrow$ on accepte l'hypothèse H_0.

Exemple 3.5.1 Test d'indépendance : taux de guérison et coût du médicament.
Pour comparer l'efficacité de 2 médicaments comparables, mais de prix très différents, la Sécurité sociale a effectué une enquête sur les guérisons obtenues avec ces deux traitements. Les résultats sont présentés dans le tableau suivant :

	Original	Générique	Total
Guérisons	156	44	200
Non-guérisons	44	6	50
Total	200	50	250

Tableau aux effectifs observés n_{ij}

Au seuil de signification $\alpha = 5\%$, peut-on conclure que ces deux médicaments ont la même efficacité ?
1. Hypothèses statistiques :
2. Seuil de signification :

3. Conditions d'application du test :
4. Degré de liberté :
5. Statistique de test :
6. Calcul de la statistique du $\chi^2_{calculé}$ sous l'hypothèse nulle H_0 :
7. Règle de décision et conclusion :

Solution

1. Hypothèses statistiques $\begin{cases} H_0 : \textit{indépendance du taux de guérison et du coût du médicament} \\ H_1 : \textit{dépendance} \end{cases}$

2. Seuil de signification : $\alpha = 5\%$

3. Conditions d'application du test : Un échantillon aléatoire de taille $n = 250$ observé selon deux caractères qualitatifs à $k = 2$ et $l = 2$ modalités.

4. Degré de liberté : $(k-1)(l-1) = 1$ *d.d.l.*

5. Statistique de test : $\sum_{i=1}^{k=2} \sum_{j=1}^{l=2} \frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} \sim \chi^2_{1\ d.d.l.}$

6. Calcul de la statistique du $\chi^2_{calculé}$ sous l'hypothèse nulle H_0 : Indépendance

	Original	Générique	Total
Guérisons	$\frac{200\times200}{250} = 160$	$\frac{200\times50}{250} = 40$	200
Non-guérisons	$\frac{50\times200}{250} = 16$	$\frac{50\times50}{250} = 10$	50
Total	200	50	250

Tableau aux effectifs théoriques $\hat{n}_{ij} = \frac{n_{i.}n_{.j}}{n}$

$$\chi^2_{calculé} = \sum_{i=1}^{k=2} \sum_{j=1}^{l=2} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = 2.5$$

7. Décision et conclusion : fractile de la loi du χ^2_1 (cf. table) : $\chi^2_{1;\alpha=5\%} = 3.84$.

   La valeur du $\chi^2_{calculé}$ appartient à la zone de non-rejet de H_0. En effet, $\chi^2_{calculé} = 2.5 < \chi^2_{1;5\%} = 3.84$. Il n'y a pas de dépendance significative entre les deux caractères : le taux de guérison et le coût du médicament sont indépendants. Au seuil de signification $\alpha = 5\%$, on peut conclure que ces deux médicaments ont la même efficacité

3.5.3 Test d'homogénéité de plusieurs populations

Introduction

On prélève au hasard k échantillons de tailles n_1, $n_2, \ldots, n_k$ de k populations. Les résultats du caractère observé dans chaque population sont ensuite classés selon r modalités. Dans ce cas, les totaux marginaux (les n_i) associés aux k échantillons sont fixés et ne dépendent pas du sondage. Il s'agit de savoir comparer les k populations entre elles et de savoir si elles ont un comportement semblable en regard du caractère étudié (qualitatif ou quantitatif). On rassemble les données dans un tableau à double entrée appelé **tableau de contingence**.

		Populations échantillonnées					
		$j=1$	$j=2$	...	j	...	$j=k$
Caractère observé selon r modalités	$i=1$	n_{11}	n_{12}		n_{1j}		n_{1k}
	$i=2$	n_{21}	n_{22}		n_{2j}		n_{2k}
	...						
	i	n_{i1}	n_{i2}		n_{ij}		n_{ik}
	...						
	$i=r$	n_{r1}	n_{r2}		n_{rj}		n_{rk}
		$n_1 = \sum_{i=1}^{r} n_{i1}$	$n_2 = \sum_{i=1}^{r} n_{i2}$		$n_j = \sum_{i=1}^{r} n_{ij}$		$n_k = \sum_{i=1}^{r} n_{ik}$

Test d'homogénéité

Il s'agit de comparer les effectifs observés pour chaque modalité du caractère avec les effectifs théoriques sous l'hypothèse d'une répartition équivalente entre les k populations et ceci pour chaque modalité du caractère. Si nous notons p_{ij} la probabilité théorique pour qu'une unité statistique choisie au hasard dans la population j présente la modalité i du caractère étudié, on peut alors préciser les hypothèses de la façon suivante :

1ère étape : Formulation des hypothèses.

H_0 : $p_{i1} = p_{i2} = \cdots = p_{ik}$ pour $i = 1, 2, \ldots, r$. Soit encore : les proportions d'individus présentant chaque modalité du caractère sont les mêmes dans les k populations.

H_1 : Les proportions des populations ne sont pas toutes égales.

2ème étape : Détermination de la fonction discriminante du test et de sa distribution de probabilité.

Sous l'hypothèse d'homogénéité des populations, on doit comparer les effectifs observés aux effectifs théoriques. Pour calculer les effectifs théoriques, il nous faut déterminer p_i, la proportion d'individus associée à la modalité i et que l'on suppose identique dans les k populations. On obtiendra une estimation de cette proportion en utilisant l'ensemble des données collectées. On

choisit donc

$$p_i = \frac{\sum_{j=1}^{k} n_{ij}}{\sum_{j=1}^{k} n_j}.$$

On en déduit les effectifs théoriques de chaque classe grâce à la relation

$$n_{t,ij} = p_i n_j.$$

Pour comparer les écarts entre ce qu'on observe et ce qui se passe sous l'hypothèse H_0, on considère la somme des écarts réduits de chaque classe, à savoir la quantité

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(N_{ij} - n_{t,ij})^2}{n_{t,ij}}.$$

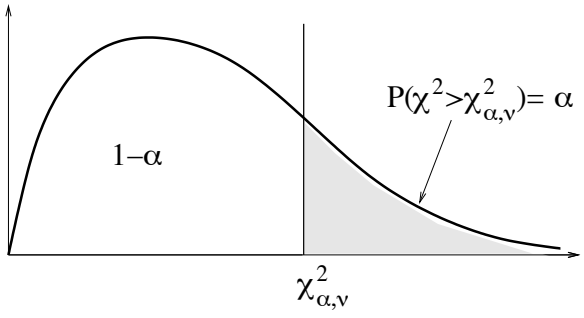
Cette variable aléatoire suit une loi du khi-deux (voir paragraphe précédent), mais quel est donc son nombre de degrés de liberté ?

Calcul du nombre de degrés de liberté du khi-deux.

- A priori, on a kr cases dans notre tableau donc kr degrés de liberté. Mais il faut retirer à cette valeur, le nombre de paramètres estimés ainsi que le nombre de relations entre les différents éléments des cases.
- On a estimé r probabilités théoriques à l'aide des valeurs du tableau$(p_1, p_2, \ldots, p_r)$, mais seulement $r-1$ sont indépendantes, puisqu'on impose la restriction $\sum_{i=1}^{r} p_i = 1$. Par ces estimations, on a donc supprimé $r-1$ degrés de liberté.
- Les effectifs de chaque colonne sont toujours liés par les relations $\sum_{i=1}^{r} N_{ij} = n_j$ (puisque les n_j sont imposés par l'expérience) et ces relations sont au nombre de k.
- Finalement, le nombre de degrés de liberté du khi-deux est $kr-(r-1)-k = (k-1)(r-1)$.

3ème étape : Détermination des valeurs critiques de délimitant les zones d'acceptation et de rejet.

On impose à la zone d'acceptation de H_0 concernant la valeur du χ^2 d'être un intervalle dont 0 est la borne inférieure (car un χ^2 est toujours positif).



Il nous faut donc déterminer dans la table la valeur maximale $\chi^2_{\alpha,\nu}$ de l'écart entre les deux distributions imputable aux variations d'échantillonnage au seuil de signification α, c'est-à-dire vérifiant $P(\chi^2 > \chi^2_{\alpha,\nu}) = \alpha$.

4ème étape : Calcul de la valeur de χ^2 prise dans l'échantillon et conclusion du test.

On calcule la valeur χ^2_0 prise par χ^2 dans l'échantillon.

- Si la valeur χ^2_0 se trouve dans la zone de rejet, on dira que l'écart observé entre les deux distributions est **statistiquement significatif** au seuil α. Cet écart est anormalement élevé et ne permet pas d'accepter H_0. On rejette H_0.
- Si la valeur χ^2_0 se trouve dans la zone d'acceptation, on dira que l'écart-réduit observé **n'est pas significatif** au seuil α. Cet écart est imputable aux fluctuations d'échantillonnage. On accepte H_0.

Bibliographie

[1] Belletante, B., B. Romier. *Mathématiques et Gestion. Les outils fondamentaux.* Enseignement Supérieur Tertiaire, Ellipses, 1991

[2] Dumoulin, D. *Mathématiques de gestion. Cours et applications* Collection D.E.C.S. dirigée par Th. Lamolette, Economica, Paris, 1987

[3] Jaffard, P. *Initiation aux méthodes de la statistique et du calcul des probabilités* Masson, Paris, 1990

[4] Rakotomalala, R. *Ouvrages* `http://eric.univ-lyon2.fr/ ricco/cours/ouvrages.html`

[5] Ramousse, R., Le Berre, M., Le Guelte, L. *Introduction aux statistiques*, chapitres 1 à 5, 1996 `http://www.cons-dev.org/elearning/stat/index.html`

[6] Ramousse, R., Le Berre, M., Le Guelte, L. *Une approche pragmatique de l'Analyse des données* `http://www.cons-dev.org/elearning/ando/index.html`

[7] Spiegel, M. *Théorie et application de la statistique* Serie Schaum, Ediscience, Paris, France, 1972

[8] Damgaliev, D., Tellalyan, . *Statistiques sur les entreprises*, NBU, Sofia, 2006 (*en bulgare)*

Annexe

Schémas

Synthèse sur les distributions d'échantillonnage

Table 1

Table 2

Estimation ponctuelle. Synthèse

Table 3

Intervalle de confiance. Synthèse

Table 4

Table 5

Table 6

Table 7

Tables statistiques

Table de Loi Normale

Fractiles de la Loi Normale

Fractiles de la loi du χ^2_ν

Table de la loi de Student

Table de la loi de Fisher-Snedecor $p = 0.05$

Table de la loi de Fisher-Snedecor $p = 0.025$

Table de la loi de Fisher-Snedecor $p = 0.01$

Feuilles

Feuille 1 : Échantillonnage

Feuille 2 : Estimation

Feuille 3 : Les tests d'hypothèse

Feuille 4 : Préparation pour les contrôles

Schémas

Synthèse sur les distributions d'échantillonnage

Table 1

Table 2

Estimation ponctuelle. Synthèse

Table 3

Intervalle de confiance. Synthèse

Table 4

Table 5

Table 6

Table 7

Synthèse sur les distributions d'échantillonnage **Table 1**

<table>
<tr><th>Variable aléatoire</th><th>Définition</th><th>Paramètres descriptifs</th><th colspan="4">Loi</th></tr>
<tr><td rowspan="4">$\bar{X}$ Moyenne d'échantillon</td><td rowspan="4">$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$
$= \frac{1}{n}\sum_{i=1}^{n} X_i$</td><td rowspan="4">$E(\bar{X}) = \mu$
$Var(\bar{X}) = \frac{\sigma^2}{n}$</td><td colspan="2">$n \geq 30$</td><td colspan="2">$n < 30$, $X \sim \mathcal{N}(\mu, \sigma)$</td></tr>
<tr><td>σ connu</td><td>σ inconnu estimation fiable
$\hat{\sigma}^2 = \frac{n}{n-1}s^2$</td><td>$\sigma$ connu</td><td rowspan="3">σ inconnu estimation fiable
$\hat{\sigma}^2 = \frac{n}{n-1}s^2$
$T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n-1}}}$
$= \frac{\bar{X}-\mu}{\frac{s'}{\sqrt{n}}}$
$T \sim T_{n-1}$</td></tr>
<tr><td colspan="3">tirage avec remise ;
tirage sans remise et $n < 0,05N$
$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$</td></tr>
<tr><td colspan="3">tirage sans remise et $n > 0,05N$
$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right)$</td></tr>
<tr><td rowspan="2">$X_1 : n_1, \mu_1, \sigma_1$
$X_2 : n_2, \mu_2, \sigma_2$</td><td rowspan="2">$\bar{X}_1 - \bar{X}_2$</td><td rowspan="2">$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$;
$Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$</td><td colspan="2">$n_1, n_2 \geq 30$; $n_i < 0,05N$</td><td colspan="2">$n_1, n_2 < 30$ et
$X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$,
$X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$</td></tr>
<tr><td colspan="4">$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
$n_i > 0,05N_i \rightarrow$ facteur d'exhaustivité</td></tr>
<tr><td rowspan="2">$S'^2_{\bar{X}}$ Variance d'échantillon - estimation de σ^2</td><td>$S^2_{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$</td><td>$E(S^2_{\bar{X}}) = \frac{n-1}{n}\sigma^2$,</td><td colspan="2">$n \geq 30$</td><td colspan="2">$n < 30$</td></tr>
<tr><td>$S'^2_{\bar{X}} = \frac{n}{n-1}S^2$
$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$</td><td>$E(S'^2_{\bar{X}}) = \sigma^2$</td><td colspan="2">$S'^2_{\bar{X}} \sim \mathcal{N}\left(\sigma^2, \sigma^2\sqrt{\frac{2}{(n-1)}}\right)$</td><td colspan="2">$\frac{(n-1)S'^2_{\bar{X}}}{\sigma^2} \sim \chi^2_{n-1}$</td></tr>
</table>

Table 2

Variable aléatoire	Définition	Paramètres descriptifs	Loi
F Proportion d'échantillon	$F = X/n$, $X \sim \mathcal{B}(n,p)$ $E(X) = np$ $Var(X) = npq$	$E(F) = p$ $Var(F) = \frac{pq}{n}$	$n \geq 30$, $np > 15$, $nq > 15$ $\mathcal{B}(n, \frac{p}{n}) \to \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$ tirage avec remise ; sans remise et $n < 0,05N$ $F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$ tirage sans remise et $n > 0,05N$ $F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$
$F_1 - F_2$ $F_1 \sim \mathcal{N}\left(p_1, \sqrt{\frac{p_1q_1}{n_1}}\right)$ $F_2 \sim \mathcal{N}\left(p_2, \sqrt{\frac{p_2q_2}{n_2}}\right)$	$F_1 - F_2$	$E(F_1 - F_2) = p_1 - p_2$ $Var(F_1 - F_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$	$n_1 \geq 30$; $n_2 \geq 30$ $F_1 - F_2 \sim$ $\mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}\right)$

Estimation ponctuelle. Synthèse

Table 3

Population mère P	taille N	Paramètres du caractère observé		
		moyenne μ	proportion p	variance σ^2
Echantillon	taille	Caractéristiques du caractère observé moyenne	fréquence	variance
E	n	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ Série stat. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i$ D.O.1 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^*$ D.G.1	$f = \frac{n_A}{n}$	observée : $s^2 = \frac{1}{n}\sum_{i=1}^{n}(s_i - \bar{x})^2$ empirique : $s'^2 = \frac{n}{n-1}s^2$
Estimations ponctuelles		$\widehat{\mu} = \bar{x}$	$\widehat{p} = f$	μ connue - $\widehat{\sigma}^2 = s^2$ μ inconnue - $\widehat{\sigma}^2 = s'^2$

Intervalle de confiance. Synthèse

Population P : taille N; moyenne $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$

Échantillon E : taille n; moyenne $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; variance $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$

Table 4 variance empirique $s'^2 = \frac{n}{n-1}s^2$

Paramètre estimé	Conditions	Statistique de test	Marge d'erreur E	$I.C._{(1-\alpha)}$
Moyenne μ	σ connue, p. 38, 39	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1)$	$t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
	σ inconnue $n < 30$ p. 40, 41	$\frac{\bar{X}-\mu}{S'/\sqrt{n}} \to T_{n-1}$ *d.d.l.*	$t_{St\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$	$\bar{x} \pm t_{St\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$
	σ inconnue $n \geq 30$ p. 42	$\frac{\bar{X}-\mu}{S'/\sqrt{n}} \to \mathcal{N}(0,1)$	$t_{\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$	$\bar{x} \pm t_{\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$
Proportion p	$n \geq 30$ p. 53, 54	$\frac{F-p}{\sqrt{\frac{f(1-f)}{n}}} \to \mathcal{N}(0;1)$	$t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$	$f \pm t_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$
Variance σ^2 écart-type σ	μ connue p. 61	$n\frac{S^2}{\sigma^2} \to \chi^2_{n\ d.d.l.}$	n *d.d.l.* $k_1 = \chi^2_{\frac{\alpha}{2}}$ $k_2 = \chi^2_{1-\frac{\alpha}{2}}$	$n\frac{s^2}{k_2} \leq \sigma^2 \leq n\frac{s^2}{k_1}$ $\sqrt{n\frac{s^2}{k_2}} \leq \sigma \leq \sqrt{n\frac{s^2}{k_1}}$
	μ inconnue $X \sim \mathcal{N}(\mu,\sigma)$ p. 62	$(n-1)\frac{S'^2}{\sigma^2} \to \chi^2_{(n-1)\ d.d.l.}$	$n-1$ *d.d.l.* $k_1 = \chi^2_{\frac{\alpha}{2}}$ $k_2 = \chi^2_{1-\frac{\alpha}{2}}$	$(n-1)\frac{s'^2}{k_2} \leq \sigma^2 \leq (n-1)\frac{s'^2}{k_1}$ $\sqrt{(n-1)\frac{s'^2}{k_2}} \leq \sigma \leq \sqrt{(n-1)\frac{s'^2}{k_1}}$
	μ inconnue $n > 100$ p. 63	$n\frac{S'^2}{\sigma^2} \to \mathcal{N}(n,\sqrt{2n})$	$t_{\frac{\alpha}{2}}\frac{s'^2}{2n}$ $t_{\frac{\alpha}{2}}\frac{s'}{\sqrt{2n}}$	$s'^2 \pm t_{\frac{\alpha}{2}}\frac{s'^2}{2n}$ $s' \pm t_{\frac{\alpha}{2}}\frac{s'}{\sqrt{2n}}$

Intervalle de confiance du rapport de 2 variances **Table 5**

Conditions	Statistique de test	Marge d'erreur E	$I.C._{(1-\alpha)}$
$X_1 \sim \mathcal{N}(\mu_1,\sigma_1)$ $X_2 \sim \mathcal{N}(\mu_2,\sigma_2)$ p. 93 - 95	$\frac{\sigma_2^2}{\sigma_1^2}\frac{S_1'^2}{S_2'^2} \to \mathcal{F}_{(n_1-1),(n_2-1)\ d.d.l.}$	$f_1 = f_{1-\frac{\alpha}{2}} = F_{1-\frac{\alpha}{2},n_1-1,n_2-1} = 1/F_{\frac{\alpha}{2},n_2-1,n_1-1}$ $= P(F(n_1-1,n_2-1) > f_1)$ $= 1-\frac{\alpha}{2}$ $f_2 = f_{\frac{\alpha}{2}} = F_{\frac{\alpha}{2},n_1-1,n_2-1}$ $P\left(F(n_1-1,n_2-1) > f_2\right) = \frac{\alpha}{2}$;	$f_1\frac{S_2'^2}{S_1'^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq f_2\frac{S_2'^2}{S_1'^2}$
Conclusion : Si $1 \in I.C._{(1-\alpha)\%}$, il n'y a pas de différence significative (avec un risque d'erreur de $\alpha\%$) entre les deux variances. On peut donc les supposer égales : $\sigma_1^2 \approx \sigma_2^2$.			

Intervalle de confiance de la différence de 2 moyennes **Table 6**

Conditions	Statistique de test	Marge d'erreur E	$I.C._{(1-\alpha)}$
σ_X^2, σ_Y^2 connues p. 68, 69	$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{p}}} \to \mathcal{N}(0;1)$	$t_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{p}}$	$(\bar{X}-\bar{Y}) \pm E$
σ_X^2, σ_Y^2 inconnues $n, p \geq 30$ p. 73, 74	$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{S_x'^2}{n}+\frac{S_y'^2}{p}}} \to \mathcal{N}(0;1)$	$t_{\frac{\alpha}{2}}\sqrt{\frac{S_x'^2}{n}+\frac{S_y'^2}{p}}$	$(\bar{X}-\bar{Y}) \pm E$
$\sigma_X^2 = \sigma_Y^2 = \sigma^2$ inconnues $n, p \leq 30$ p. 78, 79	$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{s'\sqrt{\frac{1}{n}+\frac{1}{p}}} \to T_{(n+p-2)\ d.d.l.}$ $S'^2 = \frac{nS_x^2+pS_y^2}{n+p-2}$	$t_{St\frac{\alpha}{2}}s'\sqrt{\frac{1}{n}+\frac{1}{p}}$	$(\bar{X}-\bar{Y}) \pm E$
$\sigma_X^2 = \sigma_Y^2 = \sigma^2$ inconnues $n = p \leq 30$, p. 80	$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{S'\sqrt{\frac{2}{n}}} \to T_{2(n-1)\ d.d.l.}$ $S'^2 = \frac{n(S_x^2+S_y^2)}{2(n-1)}$	$t_{St\frac{\alpha}{2}}s'\sqrt{\frac{2}{n}}$	$(\bar{X}-\bar{Y}) \pm E$
Echantillons appariés p. 84, 85 $Z = X - Y$ $Z \sim \mathcal{N}(\mu_Z, \sigma_Z)$	$\frac{\bar{Z}-\mu_z}{S'/\sqrt{n}} \to T_{n-1\ d.d.l.}$ $S'^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(Z_i-\bar{Z}\right)^2$	$t_{St\frac{\alpha}{2}}\frac{s'}{\sqrt{n}}$	$\bar{Z} \pm E$
Conclusion : Si $0 \in I.C_{(1-\alpha)} \Longrightarrow$ les deux moyennes ne sont pas différentes ; Si $0 \notin I.C_{(1-\alpha)} \Longrightarrow$ les moyennes sont significativement différentes.			

Intervalle de confiance de la différence de 2 proportions **Table 7**

Conditions	Statistique de test	Marge d'erreur E	$I.C._{(1-\alpha)}$
$n, p \geq 30$ p. 89 - 91	$\frac{(F_1-F_2)-(p-p_2)}{\sqrt{\frac{f_1(1-f_1)}{n_1}+\frac{f_2(1-f_2)}{n_2}}} \to \mathcal{N}(0;1)$	$t_{\frac{\alpha}{2}}\sqrt{\frac{f_1(1-f_1)}{n_1}+\frac{f_2(1-f_2)}{n_2}}$	$(f_1-f_2) \pm E$
$n_1, n_2 \geq 30$ $p_1 = p_2 = p$ p. 89 - 91	$\frac{(F_1-F_2)-(p_1-p_2)}{\sqrt{f(1-f)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \to N(0;1)$ $f = \frac{n_1f_1+n_2f_2}{n_1+n_2}$	$t_{\frac{\alpha}{2}}\sqrt{f(1-f)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$(f_1-f_2) \pm E$
Conclusion : Si $0 \in I.C_{(1-\alpha)} \Longrightarrow$ les deux proportions ne sont pas différentes ; Si $0 \notin I.C_{(1-\alpha)} \Longrightarrow$ les proportions sont significativement différentes.			

Tables statistiques

Table de Loi Normale

Fractiles de la Loi Normale

Fractiles de la loi du χ^2_ν

Table de la loi de Student

Table de la loi de Fisher-Snedecor $p = 0.05$

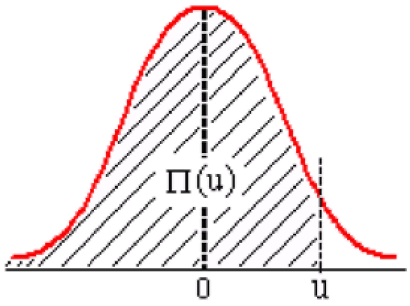
Table de la loi de Fisher-Snedecor $p = 0.025$

Table de la loi de Fisher-Snedecor $p = 0.01$

Table de la loi Normale

Fonction de répartition Π de la loi normale centrée réduite : $U \to \mathcal{N}(0,1)$
Probabilité de trouver une valeur inférieure à u

$$\Pi(u) = P(U \leq u); \Pi(-u) = P(U \leq -u) = 1 - \Pi(u)$$



u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	**0.89617**	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992

Exemple : $\Pi(1.26) = P(U \leq 1.26) = 0.89617 = 89.62\%$

Fractiles de la loi normale

$$U \to \mathcal{N}(0,1)$$

Pour $P < 0.5$ (colonne de gauche et ligne supérieure). Les fractiles sont négatifs.
Pour $P > 0.5$ (colonne de droite et ligne inférieure). Les fractiles sont positifs.

P	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01	
0	infini	3.0902	2.8782	2.7478	2.6521	2.5758	2.5121	2.4573	2.4089	2.3656	2.3263	**0.99**
0.01	2.3263	2.2904	2.2571	2.2262	2.1973	2.1701	2.1444	2.1201	2.0969	2.0748	2.0537	**0.98**
0.02	2.0537	2.0335	2.0141	1.9954	1.9774	1.9600	1.9431	1.9268	1.9110	1.8957	1.8808	**0.97**
0.03	1.8808	1.8663	1.8522	1.8384	1.8250	1.8119	1.7991	1.7866	1.7744	1.7624	1.7507	**0.96**
0.04	1.7507	1.7392	1.7279	1.7169	1.7060	1.6954	1.6849	1.6747	1.6646	1.6546	1.6449	**0.95**
0.05	1.6449	1.6352	1.6258	1.6164	1.6072	1.5982	1.5893	1.5805	1.5718	1.5632	1.5548	**0.94**
0.06	1.5548	1.5464	1.5382	1.5301	1.5220	1.5141	1.5063	1.4985	1.4909	1.4833	1.4758	**0.93**
0.07	1.4758	1.4684	1.4611	1.4538	1.4466	1.4395	1.4325	1.4255	1.4187	1.4118	1.4051	**0.92**
0.08	1.4051	1.3984	1.3917	1.3852	1.3787	1.3722	1.3658	1.3595	1.3532	1.3469	1.3408	**0.91**
0.09	1.3408	1.3346	1.3285	1.3225	1.3165	1.3106	1.3047	1.2988	1.2930	1.2873	1.2816	**0.90**
0.10	1.2816	1.2759	1.2702	1.2646	1.2591	1.2536	1.2481	1.2426	1.2372	1.2319	1.2265	**0.89**
0.11	1.2265	1.2212	1.2160	1.2107	1.2055	1.2004	1.1952	1.1901	1.1850	1.1800	1.1750	**0.88**
0.12	1.1750	1.1700	1.1650	1.1601	1.1552	1.1503	1.1455	1.1407	1.1359	1.1311	1.1264	**0.87**
0.13	1.1264	1.1217	1.1170	1.1123	1.1077	1.1031	1.0985	1.0939	1.0893	1.0848	1.0803	**0.86**
0.14	1.0803	1.0758	1.0714	1.0669	1.0625	1.0581	1.0537	1.0494	1.0451	1.0407	1.0364	**0.85**
0.15	1.0364	1.0322	1.0279	1.0237	1.0194	1.0152	1.0110	1.0069	1.0027	0.9986	0.9945	**0.84**
0.16	0.9945	0.9904	0.9863	0.9822	0.9782	0.9741	0.9701	0.9661	0.9621	0.9581	0.9542	**0.83**
0.17	0.9542	0.9502	0.9463	0.9424	0.9385	0.9346	0.9307	0.9269	0.9230	0.9192	0.9154	**0.82**
0.18	0.9154	0.9116	0.9078	0.9040	0.9002	0.8965	0.8927	0.8890	0.8853	0.8816	0.8779	**0.81**
0.19	0.8779	0.8742	0.8706	0.8669	0.8632	0.8596	0.8560	0.8524	0.8488	0.8452	0.8416	**0.80**
0.20	0.8416	0.8381	0.8345	0.8310	0.8274	0.8239	0.8204	0.8169	0.8134	0.8099	0.8064	**0.79**
0.21	0.8064	0.8030	0.7995	0.7961	0.7926	0.7892	0.7858	0.7824	0.7790	0.7756	0.7722	**0.78**
0.22	0.7722	0.7688	0.7655	0.7621	0.7588	0.7554	0.7521	0.7488	0.7454	0.7421	0.7388	**0.77**
0.23	0.7388	0.7356	0.7323	0.7290	0.7257	0.7225	0.7192	0.7160	0.7128	0.7095	0.7063	**0.76**
0.24	0.7063	0.7031	0.6999	0.6967	0.6935	0.6903	0.6871	0.6840	0.6808	0.6776	0.6745	**0.75**
0.25	0.6745	0.6713	0.6682	0.6651	0.6620	0.6588	0.6557	0.6526	0.6495	0.6464	0.6433	**0.74**
0.26	0.6433	0.6403	0.6372	0.6341	0.6311	0.6280	0.6250	0.6219	0.6189	0.6158	0.6128	**0.73**
0.27	0.6128	0.6098	0.6068	0.6038	0.6008	0.5978	0.5948	0.5918	0.5888	0.5858	0.5828	**0.72**
0.28	0.5828	0.5799	0.5769	0.5740	0.5710	0.5681	0.5651	0.5622	0.5592	0.5563	0.5534	**0.71**
0.29	0.5534	0.5505	0.5476	0.5446	0.5417	0.5388	0.5359	0.5330	0.5302	0.5273	0.5244	**0.70**
0.30	0.5244	0.5215	0.5187	0.5158	0.5129	0.5101	0.5072	0.5044	0.5015	0.4987	0.4958	**0.69**
0.31	0.4958	0.4930	0.4902	0.4874	0.4845	0.4817	0.4789	0.4761	0.4733	0.4705	0.4677	**0.68**
0.32	0.4677	0.4649	0.4621	0.4593	0.4565	0.4538	0.4510	0.4482	0.4454	0.4427	0.4399	**0.67**
0.33	0.4399	0.4372	0.4344	0.4316	0.4289	0.4261	0.4234	0.4207	0.4179	0.4152	0.4125	**0.66**
0.34	0.4125	0.4097	0.4070	0.4043	0.4016	0.3989	0.3961	0.3934	0.3907	0.3880	0.3853	**0.65**
0.35	0.3853	0.3826	0.3799	0.3772	0.3745	0.3719	0.3692	0.3665	0.3638	0.3611	0.3585	**0.64**
0.36	0.3585	0.3558	0.3531	0.3505	0.3478	0.3451	**0.3425**	0.3398	0.3372	0.3345	0.3319	**0.63**
0.37	0.3319	0.3292	0.3266	0.3239	0.3213	0.3186	0.3160	0.3134	0.3107	0.3081	0.3055	**0.62**
0.38	0.3055	0.3029	0.3002	0.2976	0.2950	0.2924	0.2898	0.2871	0.2845	0.2819	0.2793	**0.61**
0.39	0.2793	0.2767	0.2741	0.2715	0.2689	0.2663	0.2637	0.2611	0.2585	0.2559	0.2533	**0.60**
0.40	0.2533	0.2508	**0.2482**	0.2456	0.2430	0.2404	0.2378	0.2353	0.2327	0.2301	0.2275	**0.59**
0.41	0.2275	0.2250	0.2224	0.2198	0.2173	0.2147	0.2121	0.2096	0.2070	0.2045	0.2019	**0.58**
0.42	0.2019	0.1993	0.1968	0.1942	0.1917	0.1891	0.1866	0.1840	0.1815	0.1789	0.1764	**0.57**
0.43	0.1764	0.1738	0.1713	0.1687	0.1662	0.1637	0.1611	0.1586	0.1560	0.1535	0.1510	**0.56**
0.44	0.1510	0.1484	0.1459	0.1434	0.1408	0.1383	0.1358	0.1332	0.1307	0.1282	0.1257	**0.55**
0.45	0.1257	0.1231	0.1206	0.1181	0.1156	0.1130	0.1105	0.1080	0.1055	0.1030	0.1004	**0.54**
0.46	0.1004	0.0979	0.0954	0.0929	0.0904	0.0878	0.0853	0.0828	0.0803	0.0778	0.0753	**0.53**
0.47	0.0753	0.0728	0.0702	0.0677	0.0652	0.0627	0.0602	0.0577	0.0552	0.0527	0.0502	**0.52**
0.48	0.0502	0.0476	0.0451	0.0426	0.0401	0.0376	0.0351	0.0326	0.0301	0.0276	0.0251	**0.51**
0.49	0.0251	0.0226	0.0201	0.0175	0.0150	0.0125	0.0100	0.0075	0.0050	0.0025	0.0000	**0.50**
	0.01	**0.009**	**0.008**	**0.007**	**0.006**	**0.005**	**0.004**	**0.003**	**0.002**	**0.001**	**0**	**P**

Exemple : $\Pi(u) = P(U \leq u) = P = 0.6340 \Rightarrow u = 0.3425$;
$\Pi(u) = P(U \leq u) = P = 0.4020 \Rightarrow u = -0.2482$

Fractiles de la loi du χ^2_ν

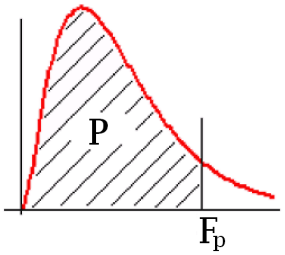
Pour $S \sim \chi^2_\nu$ à ν degrés de liberté le fractile χ^2_p d'ordre P est tel que :

$$P(X \leq \chi^2_p) = p$$

La table donne les fractiles χ^2_p, en fonction de ν, pour certaines valeurs de P.
Pour les valeurs de ν ne figurant pas dans la table, on pourra procéder par interpolation.
Par exemple, pour $\nu = 10$ et $P = 0,975$, on lit $\chi^2_p = 20,5$
et pour $P = 0,025$, on lit $\chi^2_p = 3,25$.
Pour $\nu = 75$ et $P = 0,975$, on lit $\chi^2_p = \frac{1}{2}(95,0 + 106,6) = 100,8$.

Fractiles de la loi de χ^2

Cette table donne les fractiles F_P de la loi de khi-deux à ν degrés de liberté : $P = P(\chi^2_\nu \leq F_P)$



ν P	0.010	0.020	0.025	0.050	0.100	0.150	0.200	0.800	0.900	0.950	0.975	0.980	0.990
1	0.000	0.001	0.001	0.004	0.016	0.036	0.064	1.642	2.706	3.841	5.024	5.412	6.64
2	0.020	0.040	0.051	0.103	0.211	0.325	0.446	3.219	4.605	5.991	7.378	7.824	9.21
3	0.115	0.185	0.216	0.352	0.584	0.798	1.005	4.642	6.251	7.815	9.348	9.837	11.35
4	0.297	0.429	0.484	0.711	1.064	1.366	1.649	5.989	7.779	9.488	11.143	11.668	13.28
5	0.554	0.752	0.831	1.145	1.610	1.994	2.343	7.289	9.236	11.070	12.833	13.388	15.09
6	0.872	1.134	1.237	1.635	2.204	2.661	3.070	8.558	10.645	12.592	14.449	15.033	16.81
7	1.239	1.564	1.690	2.167	2.833	3.358	3.822	9.803	12.017	14.067	16.013	16.622	18.48
8	1.646	2.032	2.180	2.733	3.490	4.078	4.594	11.030	13.362	15.507	17.535	18.168	20.09
9	2.088	2.532	2.700	3.325	4.168	4.817	5.380	12.242	14.684	16.919	19.023	19.679	21.67
10	2.558	3.059	3.247	3.940	4.865	5.570	6.179	13.442	15.987	**18.307**	20.483	21.161	23.21
11	3.053	3.609	3.816	4.575	5.578	6.336	6.989	14.631	17.275	19.675	21.920	22.618	24.73
12	3.571	4.178	4.404	5.226	6.304	7.114	7.807	15.812	18.549	21.026	23.337	24.054	26.22
13	4.107	4.765	5.009	5.892	7.042	7.901	8.634	16.985	19.812	22.362	24.736	25.472	27.69
14	4.660	5.368	5.629	6.571	7.790	8.696	9.467	18.151	21.064	23.685	26.119	26.873	29.14
15	5.229	5.985	6.262	7.261	8.547	9.499	10.307	19.311	22.307	24.996	27.488	28.259	30.58
16	5.812	6.614	6.908	7.962	9.312	10.309	11.152	20.465	23.542	26.296	28.845	29.633	32.00
17	6.408	7.255	7.564	8.672	10.085	11.125	12.002	21.615	24.769	27.587	30.191	30.995	33.41
18	7.015	7.906	8.231	9.390	10.865	11.946	12.857	22.760	25.989	28.869	31.526	32.346	34.81
19	7.633	8.567	8.907	10.117	11.651	12.773	13.716	23.900	27.204	30.144	32.852	33.687	36.19
20	8.260	9.237	9.591	10.851	12.443	13.604	14.578	25.038	28.412	31.410	34.170	35.020	37.57
21	8.897	9.915	10.283	11.591	13.240	14.439	15.445	26.171	29.615	32.671	35.479	36.343	38.93
22	9.542	10.600	10.982	12.338	14.041	15.279	16.314	27.301	30.813	33.924	36.781	37.659	40.29
23	10.196	11.293	11.689	13.091	14.848	16.122	17.187	28.429	32.007	35.172	38.076	38.968	41.64
24	10.856	11.992	12.401	13.848	15.659	16.969	18.062	29.553	33.196	36.415	39.364	40.270	42.98
25	11.524	12.697	13.120	14.611	16.473	17.818	18.940	30.675	34.382	37.652	40.646	41.566	44.31
26	12.198	13.409	13.844	15.379	17.292	18.671	19.820	31.795	35.563	38.885	41.923	42.856	45.64
27	12.879	14.125	14.573	16.151	18.114	19.527	20.703	32.912	36.741	40.113	43.195	44.140	46.96
28	13.565	14.847	15.308	16.928	18.939	20.386	21.588	34.027	37.916	41.337	44.461	45.419	48.28
29	14.256	15.574	16.047	17.708	19.768	21.247	22.475	35.139	39.087	42.557	45.722	46.693	49.59
30	14.953	16.306	16.791	18.493	20.599	22.110	23.364	36.250	40.256	43.773	46.979	47.962	50.89
40	22.164	23.838	24.433	26.509	29.051	30.856	32.345	47.269	51.805	55.758	59.342	60.436	63.69
50	29.707	31.664	32.357	34.764	37.689	39.754	41.449	58.164	63.167	67.505	71.420	72.613	76.15
60	37.485	39.699	40.482	43.188	46.459	48.759	50.641	68.972	74.397	79.082	83.298	84.580	88.38
70	45.442	47.893	48.758	51.739	55.329	57.844	59.898	79.715	85.527	90.531	95.023	96.388	100.42
80	53.540	56.213	57.153	60.391	64.278	66.994	69.207	90.405	96.578	101.88	106.63	108.07	112.33

Exemple : $\nu = 10 d.d.l.$ $P = P(\chi^2_{10} \leq F_P) = 0.95 \Rightarrow F_P = 18.307$
Approximation : Pour $\nu > 100 d.l.l.$ $\chi^2(\nu) \approx \mathcal{N}(\nu; \sqrt{2\nu})$ ou $\sqrt{2\chi^2} - \sqrt{2\nu - 1} \approx \mathcal{N}(0, 1)$

Table de la loi de Student

Soit une v.a. T ayant une densité de Student à ν degrés de liberté.
Le fractile t_p d'ordre P est tel que :

$$P(T \leq t_p) = \int_{-\infty}^{t_p} f(t)f(t)dt = P$$

Pour les valeurs de $P \leq 0,40$ on a $t_p = -t_{1-p}$.

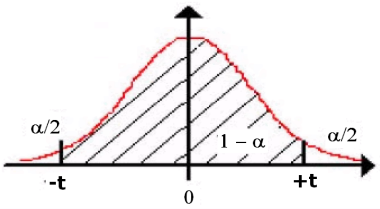
Pour les valeurs de ν ne figurant pas dans la table, on pourra :
- procéder par interpolation - utiliser l'approximation par la loi normale réduite ($\nu > 100$).

Par exemple, pour $\nu = 9$ et $P = 0,975$, on lit $t_p = 2,262$
et pour $P = 0,025$, on déduit $t_p = -2,262$.
Pour $\nu = 75$ et $P = 0,975$, on lit $t_p = \frac{1}{2}(1,994 + 1,990) = 1,992$.

Table de la loi de Student

Cette table donne les fractiles de la loi de Student à ν degrés de liberté : valeur t ayant la probabilité α d'être dépassée en valeur absolue :

$$P(|T_\nu| \leq t) = P(-t \leq T_\nu \leq t) = 1 - \alpha$$
$$P(|T_\nu| > t) = 1 - P(|T_\nu| \leq t) = \alpha$$



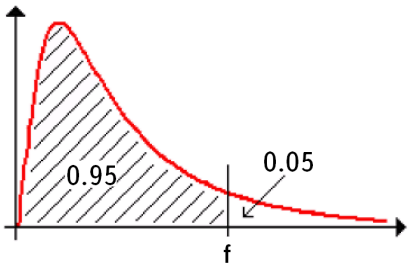
ν α	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.005	0.001
1	0.1584	0.3249	0.5095	0.7265	1	1.3764	1.9626	3.0777	6.3137	12.706	31.821	63.656	127.32	636.58
2	0.1421	0.2887	0.4447	0.6172	0.8165	1.0607	1.3862	1.8856	2.92	4.3027	6.9645	9.925	14.089	31.6
3	0.1366	0.2767	0.4242	0.5844	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8408	7.4532	12.924
4	0.1338	0.2707	0.4142	0.5686	0.7407	0.941	1.1896	1.5332	2.1318	2.7765	3.7469	4.6041	5.5975	8.6101
5	0.1322	0.2672	0.4082	0.5594	0.7267	0.9195	1.1558	1.4759	2.015	2.5706	3.3649	4.0321	4.7733	6.8685
6	0.1311	0.2648	0.4043	0.5534	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074	4.3168	5.9587
7	0.1303	0.2632	0.4015	0.5491	0.7111	0.896	1.1192	1.4149	1.8946	2.3646	2.9979	3.4995	4.0294	5.4081
8	0.1297	0.2619	0.3995	0.5459	0.7064	0.8889	1.1081	1.3968	1.8595	2.306	2.8965	3.3554	3.8325	5.0414
9	0.1293	0.261	0.3979	0.5435	0.7027	0.8834	1.0997	1.383	1.8331	2.2622	2.8214	3.2498	3.6896	4.7809
10	0.1289	0.2602	0.3966	0.5415	0.6998	0.8791	1.0931	1.3722	**1.8125**	**2.2281**	2.7638	3.1693	3.5814	4.5868
11	0.1286	0.2596	0.3956	0.5399	0.6974	0.8755	1.0877	1.3634	1.7959	2.201	2.7181	3.1058	3.4966	4.4369
12	0.1283	0.259	0.3947	0.5386	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.681	3.0545	3.4284	4.3178
13	0.1281	0.2586	0.394	0.5375	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123	3.3725	4.2209
14	0.128	0.2582	0.3933	0.5366	0.6924	0.8681	1.0763	1.345	1.7613	2.1448	2.6245	2.9768	3.3257	4.1403
15	0.1278	0.2579	0.3928	0.5357	0.6912	0.8662	1.0735	1.3406	1.7531	2.1315	2.6025	2.9467	3.286	4.0728
16	0.1277	0.2576	0.3923	0.535	0.6901	0.8647	1.0711	1.3368	1.7459	2.1199	2.5835	2.9208	3.252	4.0149
17	0.1276	0.2573	0.3919	0.5344	0.6892	0.8633	1.069	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.9651
18	0.1274	0.2571	0.3915	0.5338	0.6884	0.862	1.0672	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.9217
19	0.1274	0.2569	0.3912	0.5333	0.6876	0.861	1.0655	1.3277	1.7291	2.093	2.5395	2.8609	3.1737	3.8833
20	0.1273	0.2567	0.3909	0.5329	0.687	0.86	1.064	1.3253	1.7247	2.086	2.528	2.8453	3.1534	3.8496
21	0.1272	0.2566	0.3906	0.5325	0.6864	0.8591	1.0627	1.3232	1.7207	2.0796	2.5176	2.8314	3.1352	3.8193
22	0.1271	0.2564	0.3904	0.5321	0.6858	0.8583	1.0614	1.3212	1.7171	2.0739	2.5083	2.8188	3.1188	3.7922
23	0.1271	0.2563	0.3902	0.5317	0.6853	0.8575	1.0603	1.3195	1.7139	2.0687	2.4999	2.8073	3.104	3.7676
24	0.127	0.2562	0.39	0.5314	0.6848	0.8569	1.0593	1.3178	1.7109	2.0639	2.4922	2.797	3.0905	3.7454
25	0.1269	0.2561	0.3898	0.5312	0.6844	0.8562	1.0584	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782	3.7251
26	0.1269	0.256	0.3896	0.5309	0.684	0.8557	1.0575	1.315	1.7056	2.0555	2.4786	2.7787	3.0669	3.7067
27	0.1268	0.2559	0.3894	0.5306	0.6837	0.8551	1.0567	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565	3.6895
28	0.1268	0.2558	0.3893	0.5304	0.6834	0.8546	1.056	1.3125	1.7011	2.0484	2.4671	2.7633	3.047	3.6739
29	0.1268	0.2557	0.3892	0.5302	0.683	0.8542	1.0553	1.3114	1.6991	2.0452	2.462	2.7564	3.038	3.6595
30	0.1267	0.2556	0.389	0.53	0.6828	0.8538	1.0547	1.3104	1.6973	2.0423	2.4573	2.75	3.0298	3.646
50	0.1263	0.2547	0.3875	0.5278	0.6794	0.8489	1.0473	1.2987	1.6759	2.0086	2.4033	2.6778	2.937	3.496
60	0.1262	0.2545	0.3872	0.5272	0.6786	0.8477	1.0455	1.2958	1.6706	2.0003	2.3901	2.6603	2.9146	3.4602
70	0.1261	0.2543	0.3869	0.5268	0.678	0.8468	1.0442	1.2938	1.6669	1.9944	2.3808	2.6479	2.8987	3.435
80	0.1261	0.2542	0.3867	0.5265	0.6776	0.8461	1.0432	1.2922	1.6641	1.9901	2.3739	2.6387	2.887	3.4164
infini (loi normale)	0.1257	0.2533	0.3853	0.5244	0.6744	0.8416	1.0364	1.2816	1.6449	1.96	2.3264	2.5759	2.8072	3.2908

Exemple : $\nu = 10 d.d.l.$ $P = P(|T_{10}| \leq t) = 0.95 \Rightarrow t = \pm 2.2281$
$P = P(T_{10} \leq t) = 0.95 \Rightarrow t = +1.8125$

Table de la loi de Fisher-Snedecor

Valeur f de la variable de Fisher-Snedecor $F(\nu_1;\nu_2)$ ayant la probabilité 0.05 d'être dépassée

ν_1 : degrés de liberté du numérateur
ν_2 : degrés de liberté du dénominateur



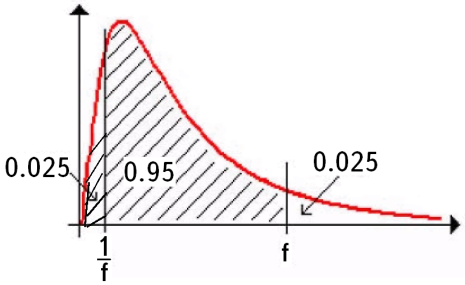
ν_2 ν_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.90	244.69	245.36	245.95	246.47
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	**4.74**	4.70	4.68	4.66	4.64	4.62	4.60
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99
10	4.96	4.10	3.71	3.48	**3.33**	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.16
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.13
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13	2.11
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.09
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.07
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	2.05
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08	2.06	2.04
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	2.02
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05	2.03	2.01
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.99
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07	2.04	2.01	1.99	1.97
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.05	2.02	1.99	1.97	1.95
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.07	2.03	2.00	1.98	1.95	1.93
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.99	1.96	1.94	1.92
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.90

Exemple : $\nu_1 = 5$ d.d.l. et $\nu_2 = 10$ d.d.l. $P = P(F_{5.10} \leq f) = 0.95 \Rightarrow f = 3.33$

Table de la loi de Fisher-Snedecor

Valeur f de la variable de Fisher-Snedecor $F(\nu_1;\nu_2)$ ayant la probabilité 0.025 d'être dépassée

ν_1 : degrés de liberté du numérateur
ν_2 : degrés de liberté du dénominateur



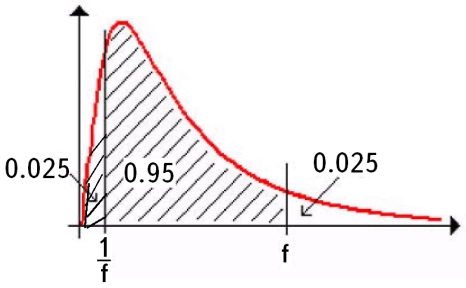
ν_2 ν_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.90	244.69	245.36	245.95	246.47
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	**4.74**	4.70	4.68	4.66	4.64	4.62	4.60
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99
10	4.96	4.10	3.71	3.48	**3.33**	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.16
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.13
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13	2.11
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.09
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.07
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	2.05
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08	2.06	2.04
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	2.02
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05	2.03	2.01
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.99
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07	2.04	2.01	1.99	1.97
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.05	2.02	1.99	1.97	1.95
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.07	2.03	2.00	1.98	1.95	1.93
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.99	1.96	1.94	1.92
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.90

Exemples : $\nu_1 = 5$ d.d.l. et $\nu_2 = 10$ d.d.l. $P = P(F_{97,5\%;5.10} \leq f') = 0.025$
$P(F_{97.5\%;10.5} \leq f) = 0.975 \Rightarrow f = 6.62 \Rightarrow f' = 1/f = 1/6.62 = 0.151$

Table de la loi de Fisher-Snedecor

Valeur f de la variable de Fisher-Snedecor $F(\nu_1;\nu_2)$ ayant la probabilité 0.01 d'être dépassée

ν_1 : degrés de liberté du numérateur
ν_2 : degrés de liberté du dénominateur



ν_2 ν_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4052	4999	5403	5624	5763	5858	5928	5980	6022	6055	6083	6106	6125	6143	6156
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.41	99.42	99.42	99.43	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.98	26.92	26.87
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25	14.20
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77	9.72
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60	7.56
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36	6.31
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56	5.52
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01	4.96
10	10.04	7.56	6.55	5.99	**5.64**	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29	4.25
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86	3.82
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70	3.66
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56	3.52
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.50	3.45	3.41
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.40	3.35	3.31
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.32	3.27	3.23
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.24	3.19	3.15
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.18	3.13	3.09
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.12	3.07	3.03
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.07	3.02	2.98
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	3.02	2.97	2.93
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.98	2.93	2.89
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.94	2.89	2.85
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96	2.90	2.86	2.81
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.99	2.93	2.87	2.82	2.78
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90	2.84	2.79	2.75
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.93	2.87	2.81	2.77	2.73
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.79	2.74	2.70
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.61	2.56	2.52

Exemple : $\nu_1 = 5$ d.d.l. et $\nu_2 = 10$ d.d.l. $P = P(F_{5.10} \leq f) = 0.95 \Rightarrow f = 2.64$

Feuilles

Feuille 1 : Échantillonnage

Feuille 2 : Estimation

Feuille 3 : Les tests d'hypothèse

Feuille 4 : Préparation pour les contrôles

Feuille 1 : Échantillonnage

Exercices

1. [7] On suppose que les poids de 3000 étudiants d'une université suivent une loi normale de moyenne 68,0 kilogrammes et d'écart-type 3,0 kilogrammes. Si l'on extrait un échantillon de 25 étudiants, quelle est la moyenne et l'écart-type théoriques de la distribution d'échantillonnage des moyennes pour
a) un échantillonnage non exhaustif,
b) un échantillonnage exhaustif,
c) un échantillonnage exhaustif, dont la taille de l'échantillon est $n = 300$?

2. Le magazine Barron's a rapporté que le nombre moyen de semaines passées au chômage par un individu est égale à 17,5. Supposez que pour la population de tous les chômeurs, la durée moyenne de chômage de la population soit de 17,5 semaines et que l'écart-type de la population soit de 4 semaines. Supposez que vous vouliez sélectionner un échantillon aléatoire de 50 chômeurs pour effectuer une étude.
a) Représenter la distribution d'échantillonnage de $\bar{x}$, la moyenne d'échantillon pour un échantillon de 50 chômeurs.
b) Quelle est la probabilité qu'un échantillon aléatoire simple de 50 chômeurs fournisse une moyenne d'échantillon qui s'écarte au plus de ± 1 semaine de la moyenne de la population ?
c) Quelle est la probabilité qu'un échantillon aléatoire simple de 50 chômeurs fournisse une moyenne d'échantillon qui s'écarte de $\pm 1/2$ semaine de la moyenne de la population ?

3. Pour estimer l'age moyen d'une population de 4000 employés, un échantillon aléatoire simple de 40 employés est sélectionné.
a) Utilisez-vous le facteur de correction pour population finie pour calculer l'écart-type de la moyenne de l'échantillon ? Expliquer.
b) Si l'écart-type de la population est $\sigma = 8,2$ ans, calculer l'écart-type de la moyenne de l'échantillon avec et sans le facteur de correction pour population finie. Quel est le raisonnement pour expliquer l'abandon du facteur de correction pour population finie lorsque $n/N \leq 0,05$?
c) Quelle est la probabilité que l'age moyen des employés de l'échantillon s'écarte au plus de ± 2 ans de l'age moyen de la population ?

4. Les producteurs de biens d'épicerie américains ont indiqué que 76% des consommateurs lisent les étiquettes indiquant la composition des produits. Supposez que la proportion de la population soit $p = 0,76$ est qu'un échantillon de 400 consommateurs soit issu de cette population.
a) Déterminer la distribution d'échantillonnage de la proportion d'échantillon f correspondant à la proportion des consommateurs de l'échantillon qui lisent l'étiquette de composition des produits.
b) Quelle est la probabilité que la proportion d'échantillon s'écarte d'au plus $\pm 0,03$ de proportion de la population ?
c) Répondre à la question (b) pour un échantillon de 750 clients.

5. [7] Cinq cents pignons ont un poids moyen de 502 grammes et un écart-type de 0,3 grammes. Trouver la probabilité pour qu'un échantillon de 100 pignons choisis au hasard ait un poids total
a) compris entre 469 et 500 grammes
b) plus grand que 510 grammes.

6. [7] A et B jouent tous deux à pile ou face en jetant chacun 50 pièces de monnaie. A gagne au jeu s'il réussit à avoir 5 faces ou davantage de plus que B, sinon c'est B qui gagne. Déterminer la probabilité pour que A ne gagne pas lors d'un jeu particulier.

7. [7] Les ampoules électriques d'un fabricant A ont une durée de vie moyenne de 1400 heures avec un écart-type de 200 heures, et celle d'un fabricant B ont une durée de vie moyenne de 1200 heures avec un écart-type de 100 heures. Si l'on teste des échantillons de 125 ampoules pour chaque marque, quelle est la probabilité pour que la marque d'ampoules A ait une durée de vie moyenne qui soit au moins supérieure de
a) 16 heures
b) 250 heures
à celle de la marque d'ampoules B ?

Feuille 2 : Estimation

Exemple 2.1.1 Supposons qu'une entreprise compte 200 employés et que l'échantillon de 50 employés a été prélevé au hasard parmi les deux cents.

Cat. salariale/mois	Nombre de salariés
Moins de 2 M.Euros	18
$[2-4[$	20
4 M.Euros et plus	12
Total	50

1. Donner une estimation de la proportion de l'ensemble des employés dont le salaire mensuel est de 2 M.Euros et plus.
2. Quel est le taux de sondage ?
3. Déterminer la probabilité qu'au moins 30 employés de cet échantillon possèdent un salaire mensuel de 2 M.Euros et plus lorsque la population échantillonnée en contient 64%.

Exemple 2.1.2 [8] Les prix d'un article en 5 différents marchés d'une région donnée sont :

i	1	2	3	4	5
x_i	75	82	83	78	80

Calculer les estimations ponctuelles de la moyenne et de l'écart-type.

Exemple 2.1.3 La table de distributions des salaires en € de 100 employés d'une entreprise est donnée ci-dessous :

Classe	Centre de la classe x_i^*	Effectif n_i
400 , 500	450	11
500 , 600	550	30
600 , 700	650	39
700 , 800	750	18
800 , 900	850	2

Calculer les estimations ponctuelles de la moyenne et de l'écart-type.

Exemple 2.2.1 [2]

1. Soit X la v.a. «durée de vie du tube cathodique d'une marque de T.V.».

   On ne connaît pas la moyenne des durées de vie des tubes bien que l'on sache qu'elles sont distribuées normalement. L'écart-type de la distribution des durées de vie $\sigma = 450$.

   Dans un échantillon de 55 tubes on a calculé que la durée de vie moyenne était de 9 500 heures.

   Déterminer l'intervalle de confiance à 90 % de la durée de vie moyenne de la population des tubes.

2. Reprenons le même exemple, mais cette fois l'échantillon est de taille $n = 25$. Déterminons l'intervalle de confiance à 99 % de la durée de vie moyenne des tubes, sachant que $\bar{x} = 9500$ heures.

3. Supposons que la population soit distribuée normalement, mais que σ ne soit pas connu. A partir d'un échantillon de taille $n = 60$, nous avons $\bar{x} = 9450$ et $s = 446.234$.
   Estimons à l'aide d'un intervalle de confiance à 95 % la moyenne de la population.

4. Supposons que la distribution soit normale, que σ ne soit pas connu, et que l'écart type s d'un échantillon de taille $n = 25$ soit égal à 440,908, $\bar{x}$ étant égal à 9 500.
   Déterminons l'intervalle de confiance à 99 % et comparons le à celui de l'exemple 2.

Exemple 2.2.2 [2] Les responsables d'une étude de marché ont choisi au hasard 500 femmes dans une grande ville et ont constaté que 35 % des femmes retenues dans l'échantillon préfèrent utiliser une marque de lessive A plutôt que les autres. Ils veulent déterminer l'intervalle de confiance à 95 % de la proportion des femmes de cette ville qui préfèrent la marque de lessive A.

Exemple 2.2.3 Les responsables d'une étude de marché ont choisi au hasard 500 femmes dans une grande ville et ont constaté que 35 % des femmes retenues dans l'échantillon préfèrent utiliser une marque de lessive A plutôt que les autres.

Supposons qu'avant de tirer l'échantillon, les responsables de l'étude aient décidé d'estimer la proportion p à $\pm 2\%$ près.

Quelle devrait être dans ce cas la taille minimale de l'échantillon à tirer, en désirant toujours avoir un intervalle de confiance à 95 % et en considérant que $f = 0.35$.

Exemple 2.2.4 On suppose que le chiffre d'affaires mensuel d'une entreprise suit une loi normale de moyenne inconnue μ mais dont l'écart-type s a été estimé à 52 K.Euros. Sur les 16 derniers mois, la moyenne des chiffres d'affaires mensuels a été de 250 K.Euros.
1 Donner une estimation ponctuelle de l'écart-type σ du chiffre d'affaires mensuel cette entreprise.
2 Établir un intervalle de confiance de niveau 95% de σ.

Exemple 2.3.1 Le temps mis par une machine pour fabriquer une pièce est supposé suivre une loi normale de paramètres μ et σ^2. Dans un atelier, deux machines A et B fabriquent la même pièce. Pour un échantillon de 9 pièces fabriquées, on a obtenu les résultats suivants :

	Machine A	Machine B
Nombre de pièces fabriquées	9	9
Temps moyen observé (mn)	50	45
Variances des populations	25	36

1. Déterminer un intervalle de confiance, de niveau $(1 - \alpha) = 95\%$, de la différence des temps moyens des deux machines $\mu_a - \mu_b$.
2. Question : La machine A est-elle aussi performante que la machine B ?

Exemple 2.3.2 On fait subir à des cadres intermédiaires de deux grandes entreprises (une œuvrant dans la fabrication d'équipement de transport et l'autre dans la fabrication de produits électriques) un test d'appréciation et d'évaluation. La compilation des résultats pour chaque groupe à l'issue de cette évaluation s'établit comme suit :

	1 Équipement	2 Produits Électriques
Nombre de cadres	34	32
Appréciation globale moyenne	184	178
Somme des Carrés des Écarts /SCDE/	15774	9858

1. Déterminer un intervalle de confiance qui a 95 chances sur 100 de contenir la valeur vraie de la différence des moyennes $(\mu_1 - \mu_2)$ des deux groupes de cadres.
2. Question : Selon cet intervalle, que peut-on conclure quant à la performance des cadres de ces deux secteurs au test d'évaluation ? Est-ce qu'en moyenne, la performance est vraisemblablement identique ou semble-t-il une différence significative entre ces deux groupes ?

Exemple 2.3.3 Un laboratoire indépendant a effectué, pour le compte d'une revue sur la protection du consommateur, un essai de durée de vie sur un type d'ampoules électriques d'usage courant (60 Watts , 120 Volts) fabriquées par deux entreprises concurrentielles, dans le secteur

de produits d'éclairage. Les essais effectués dans les mêmes conditions, sur un échantillon de 21 lampes provenant de chaque fabricant, donnent les résultats suivants :

La durée de vie d'une ampoule est supposée normalement distribuée.(les variances des populations sont supposées égales).

	Fabricant 1	Fabricant 2
Nombre d'essais	21	21
Durée de vie moyenne observée (h)	1025	1070
Somme des Carrés des Écarts	2400	2800

1. Déterminer un intervalle de confiance de niveau 95% de la différence des durées de vie moyennes des ampoules de ces deux fabricants.
2. Question : Est-ce que la revue peut affirmer, qu'en moyenne, les durées de vie des ampoules des deux fabricants sont identiques (ou différentes) ?
En d'autres termes, est-ce que la différence observée lors des essais est significative ?

Exemple 2.3.4 On mesure 12 pièces avec des méthodes différentes. On a obtenu les résultats suivants :

$\bar{x} = 1$; $\bar{y} = 2:08$; $SCE_x = 106.16$; $SCE_y = 118.19$ et $SCE_{x-y} = 14.58$.

Déterminer un intervalle de confiance de niveau 95% de la différence des deux méthodes de mesures.

Exemple 2.3.5 Dans deux municipalités avoisinantes, on a effectué un sondage pour connaître l'opinion des contribuables sur un projet d'aménagement d'un site. Les résultats de l'enquête se résument comme suit :

	Municipalité 1	Municipalité 2
Nombre de personnes interrogées	250	250
En faveur du projet	110	118

1. Quelle est l'estimation ponctuelle de la différence de proportions des contribuables de chaque municipalité favorisant l'aménagement du site ?
2. Déterminer l'intervalle de confiance de niveau $(1 - \alpha) = 95\%$ de contenir la valeur vraie de la différence des proportions, $(p_1 - p_2)$?
3. Question : Avec l'intervalle calculé en 2), est-ce que l'on rejetterait, au seuil de signification $\alpha = 5\%$, l'hypothèse selon laquelle les contribuables des deux municipalités favorisent dans la même proportion l'aménagement du site sur leur territoire ?

Exemple 2.3.6 Reprenons l'exemple de la durée de vie moyenne de 2 types d'ampoules électriques d'usage courant (60 Watts , 120 Volts) fabriquées par deux entreprises concurrentielles, dans le secteur de produits d'éclairage. Les essais effectués dans les mêmes conditions, sur un échantillon de 21 lampes provenant de chaque fabricant, donnent les résultats suivants : La durée de vie d'une ampoule est supposée normalement distribuée. **On ne dispose d'aucune information sur les variances des deux populations.**

	Fabricant 1	Fabricant 2
Nombre d'essais	21	21
Durée de vie moyenne observée (h)	1025	1070
Somme des Carrés des Écarts	2400	2800

1. Déterminer un intervalle de confiance de niveau 95% du rapport des variances des populations d'ampoules de ces deux fabricants.

2. Question : Peut-on considérer l'égalité des variances $\sigma_2^2 = \sigma_1^2$?

Estimation ponctuelle

Exercices

8. [1] Dans une ville comportant 20 000 salariés, un institut fait un sondage portant sur 100 salariés et trouve comme moyenne des salaires mensuels 7 100 € avec un écart-type de 700 €. Cet institut désire estimer la moyenne et l'écart-type de l'ensemble des salariés.

9. [1] Pour connaitre le nombre de garages qu'il fallait construire dans un immeuble en 1968 afin que les locataires puissent y garer leurs voitures, une enquête avait été faite : sur 100 ménages consultés, 40 avaient une voiture (on suppose, pour simplifier, une seule voiture par ménage).
a) Estimer la proportion p de manages qui avaient une voiture. On donnera une estimation ponctuelle puis une estimation par intervalle de confiance (à 95 %).
b) On prévoyait que 10 ans plus tard, le nombre de voitures par ménage serait de 0,6. Un ensemble de 600 appartements devrait être édifié. Quel nombre minimum de garages fallait-il construire pour être assuré avec une probabilité de 0,95 que tous les locataires puissent y ranger leurs voitures.

10. [7] On a effectué cinq mesures du diamètre d'une sphère qui ont respectivement donné 6,33 ; 6,37 ; 6,36 ; 6,32 et 6,37 cm. Déterminer des estimateurs sans biais et efficaces
a) de la moyenne vraie,
b) de la variance vraie.
Rep. $\hat{\mu} = 6.35$ cm ; $\hat{\sigma}^2 = 0.00055\ cm^2$

11. [7] Supposons que les poids de 100 étudiants de l'université X représentent un échantillon aléatoire des poids des étudiants de cette université de moyenne $\bar{x} = 67.45$ kg et variance $s^2 = 8.5275$. Déterminer des estimateurs non biaisés et efficaces
a) de la moyenne vraie,
b) de la variance vraie.
Rep. $\hat{\mu} = 67.45$ kg ; $\hat{\sigma}^2 = 8.6136$

12. [7] Donner un estimateur sans biais et inefficace de la moyenne du diamètre de la sphère de l'exercice 10.
Rep. $\hat{\mu} = m_e = 6.36$

Intervalle de confiance de la moyenne d'une population

13. [7] Déterminer un intervalle de confiance
a) à 95 %,
b) à 99 % pour estimer le poids moyen des étudiants de l'université X de l'exercice 11.

Rep. $I.C._{0.95} = [66.88, 68.02]$, $I.C._{0.99} = [66, 69, 68, 21]$

14. [3] Une firme a 2342 employés. Pour faire une évaluation rapide du nombre total a des enfants de tous ces employés, on fait un sondage au cours duquel on interroge 150 employés et on obtient les résultats suivants, en notant n_i le nombre des employés interrogés ayant $x_i = i$, $i = 0, 1, 2, \ldots$ enfants :

x_i	0	1	2	3
n_i	78	48	19	5

a) Donner un estimation de a.
b) Donner pour a un intervalle de confiance de seuil 0,05.

Rep. $a \approx 1577$; $I.C._{95\%}(a) = [1267 : 1884]$

15. [3] Une ville a 15 020 logements. Un sondage effectué sur 40 logements choisis au hasard a donné les nombres suivants d'habitants par logement :

4 - 3 - 3 - 3 - 2 - 3 - 3 - 6 - 5 - 4 - 4 - 5 - 3 - 4 - 7 - 2 - 3 - 4 - 2 - 3
4 - 2 - 4 - 3 - 4 - 2 - 1 - 3 - 3 - 4 - 3 - 3 - 6 - 2 - 5 - 4 - 3 - 2 - 1 - 4

Estimer le nombre total des habitant de la ville et donner un intervalle de confiance de seuil 0,05.

Rep. $\approx$ 54159 habitants ; $I.C._{95\%}$(nmbr habitants) = [47603 : 60690]

16. [7] Les mesures des diamètres de 200 roues dentées issues d'un échantillon aléatoire, fabriquées pendant une journée par une certaine machine, ont montré que la moyenne du diamètre était 0,854 cm et l'écart-type 0,042 cm. Déterminer les limites de confiance
a) à 95 %
b) à 99 % du diamètre moyen de toutes les roues dentées.

17. [7] En mesurant un temps de réaction, un psychologue estime que l'écart-type est de 0,05 seconde. Quelle doit être la taille de son échantillon de mesures pour que l'erreur de son estimation n'excède pas 0,01 seconde
a) à 95 %
b) à 99 % ?

Intervalle de confiance de la fréquence d'une population

18. [7] Un échantillon de 100 votants choisis au hasard parmi tous les votants d'une circonscription donnée a montré que 55 % d'entre eux étaient favorables à un certain candidat. Déterminer les limites de confiance
a) à 95%
b) à 99%
c)à 99.73 % de la proportion de tous les votants favorables à ce candidat.

19. [7] De quelle taille doit être l'échantillon de votants de l'exercice 18 si l'on veut être sur
a) à 95%
b) à 99.73 % que le candidat sera élu ?

20. [7] En jetant 40 fois une pièce, on obtient 24 fois face. Déterminer les limites de confiance
a) à 95%
b) à 99,73 % de la fréquence des faces que l'on aurait obtenue pour un nombre de jets illimité.

21. [2] Le directeur financier d'une société sait par expérience que 12 % des factures émises ne sont pas réglées dans les 10 jours ouvrables suivant l'échéance. Le chiffre d'affaires s'étant accru sensiblement, il veut vérifier si la situation a évolué.
Il fait prélever un échantillon aléatoire de 500 factures à partir duquel il constate que 14 % des factures ne sont pas réglées dans les délais.
Déterminer l'intervalle de confiance à 95 % et commenter ce résultat sachant que l'ensemble des factures pouvant être étudiées est de plusieurs dizaines de milliers.

Intervalle de confiance d'un écart-type

22. [7] On a calculé que l'écart-type des durées de vie d'un échantillon de 200 ampoules électriques valait 100 heures.
a) Déterminer les limites de confiance à 95 % de l'écart-type de l'ensemble des ampoules de ce type.

b) Déterminer les limites de confiance à 95 % de l'écart-type de l'ensemble des ampoules de ce type à la base d'un échantillon de 25 ampoules dont l'écart-type vaut 110 heures.

23. [7] L'écart-type de la résistance de rupture de 100 câbles testés par une usine est de 180 kg. Calculer les limites de confiance à 99 % de l'écart-type de tous les câbles fabriqués par l'usine.

Intervalle de confiance de différence

24. [7] Un échantillon de 150 lampes de qualité A a donné une durée de vie moyenne de 1400 heures et un écart-type de 120 heures. Un échantillon de 200 lampes de qualité B a donné une durée de vie moyenne de 1200 heures et un écart-type de 80 heures. Déterminer les limites de confiance
a) à 95 %
b) à 99 % de la différence des durées de vie moyenne des variétés A et B.
c) Est-ce-que les deux variétés possèdent les mêmes performances ?

25. [7] Sur un échantillon de 400 adultes et de 600 adolescents ayant regardé un certain programme de télévision, 100 adultes et 300 adolescents l'ont apprécié. Calculer les limites de confiance à 95 % de la différence des fréquences des adultes et des adolescents qui ont regardé et apprécié le programme.

26. [7] Un échantillon de 200 pièces fabriquées par une machine a donné 15 pièces défectueuses tandis qu'un échantillon de 100 autres pièces prélevé dans la production d'une autre machine a donné 12 pièces défectueuses.
a) Calculer les limites de confiance à 95 % de la différence des fréquences des pièces défectueuses sur les deux machines.
b) Les deux machines sont-elles de performances égales ?

27. [7] On administre des somnifères sous forme de piles à deux groupes de malades, A et B, comprenant respectivement 50 et 100 individus. On a donné au groupe A des piles d'un type nouveau et au groupe B des piles classiques. Les patients du groupe A ont dormi 7,82 heures en moyenne, ceux du groupe B 6,75 heures.
a) L'écart-type étant pour le groupe A 0,24 heures, pour le groupe B 0,30 heures, calculer les limites de confiance à 95 % pour la différence des moyennes d'heures de sommeil provoquées par les deux types de somnifères.
b) L'écart-type étant estimé pour le groupe A 0,20 heures, pour le groupe B 0,28 heures, calculer les limites de confiance à 99 % pour la différence des moyennes d'heures de sommeil provoquées par les deux types de somnifères.
c) Soit le groupe A composé de 10 individus et le groupe B de 15 individus, dont le sommeil moyen des individus du groupe A fut 7,55 heures, celui du groupe B fut 6,65 heures avec un écart-type observé de 0,22 heures et 0,28 heures respectivement. Calculer l'intervalle de confiance de la différence à 90 % des moyennes d'heures de sommeil.
d) On dispose seulement d'un groupe de 51 individus pour le test de l'efficacité des deux types de somnifères. On a donné une semaine des piles du type nouveau et les patients ont dormi $\bar{x} = 7.55$ heures en moyenne. Après deux semaines de repos, on a administré les piles du type classique et cette fois-ci les patients ont dormi $\bar{y} = 6,28$ heures en moyenne. La somme des carrés des écarts est $SCE_{x-y} = 12.25$ heurs. Déterminer un intervalle de confiance à 99 % de la différence des moyennes de sommeil en résultats des deux somnifères.

Indications et résultats :

a) $A : n_A = 50;\ \bar{x}_A = 7,,82$ h. ; $\sigma_A = 0,24$ h.
$B : n_B = 100;\ \bar{x}_B = 6,75$ h. ; $\sigma_B = 0,30$ h.
σ_A^2, σ_B^2 connues $\quad I.C._{.95\%}(\mu_A - \mu_B) = ?$ $\quad$ Table 6

$$\text{Statistique de test :} \quad \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \to \mathcal{N}(0,1)$$

$$\text{Marge d'erreur :} \quad E = t_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}};\ \alpha = 0,05;\ 1-\alpha = 0,95;\ t_{\frac{\alpha}{2}} = 1,96$$

$$\begin{aligned} E &= 1,96\sqrt{\frac{0,24^2}{50} + \frac{0,3^2}{100}} = 0,0088786 \\ \hat{\mu}_A - \hat{\mu}_B &= \bar{x}_A - \bar{x}_B = 7,82 - 6,75 = 1,07\ h. \\ I.C._{.0.95}(\mu_A - \mu_B) &= (\bar{x}_A - \bar{x}_B) \pm E = 1,07 \pm 0,09 \\ I.C._{.0.95}(\mu_A - \mu_B) &= [0,98 \quad 1,16] \end{aligned}$$

Comme $0 \notin I.C._{.0.95}(\mu_A - \mu_B) = [0,98 \quad 1,16] \Longrightarrow$ les heures moyennes de sommeil sont significativement différentes. Les deux types de somnifères influencent de façons différentes les patients.

b) $A : n_A = 50;\ \bar{x}_A = 7,82;\ s_A = 0,20$ h.
$B : n_B = 100;\ \bar{x}_B = 6,75;\ s_B = 0,28$ h.
σ_A, σ_B inconnus ; $\quad I.C._{.0,99}(\mu_A - \mu_B) = ?,\ n_A, n_B > 30$ $\quad$ Table 6

$$\text{Statistique de test :} \quad \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A'^2}{n_A} + \frac{s_B'^2}{n_B}}} \to \mathcal{N}(0,1)$$

$$\text{Fractile } t_{\frac{\alpha}{2}} : \quad \alpha = 0,01;\ \frac{\alpha}{2} = 0,005;\ 1 - \alpha/2 = 0,995;\ t_{\frac{\alpha}{2}} = 2,576$$

$$\text{Marge d'erreur :} \quad E = t_{\frac{\alpha}{2}} \sqrt{\frac{s_A'^2}{n_A} + \frac{s_B'^2}{n_B}} = t_{\frac{\alpha}{2}} \sqrt{\frac{s_A^2}{n_A - 1} + \frac{s_B^2}{n_B - 1}}$$

$$\begin{aligned} E &= 2,576\sqrt{\frac{0,20^2}{49} + \frac{0,28^2}{99}} = 0,103 \\ \hat{\mu}_A - \hat{\mu}_B &= \bar{x}_A - \bar{x}_B = 7,82 - 6,75 = 1,07\ h. \\ I.C._{.0,99}(\mu_A - \mu_B) &= (\bar{x}_A - \bar{x}_B) \pm E = 1,07 \pm 0,103 \\ I.C._{.0,99}(\mu_A - \mu_B) &= [0,967 \quad 1,173] \end{aligned}$$

Comme $0 \notin I.C._{.0,99}(\mu_A - \mu_B) = [0,967 \quad 1,173] \Longrightarrow$ les heures moyennes de sommeil sont significativement différentes. Les deux types de somnifères influencent de façons différentes les patients.

c) $A : n_A = 10;\ \bar{x}_A = 7,55$ h. ; $s_A = 0,22$ h.
$B : n_B = 15;\ \bar{x}_B = 6,65$ h. ; $s_B = 0,28$ h.
σ_A^2, σ_B^2 inconnues $\quad n_A < 30,\ n_B < 30$ $\quad I.C._{.95\%}(\mu_A - \mu_B) = ?$ Table 6

$\sigma_A = \sigma_B = \sigma$? Table 5

Statistique de test : $\dfrac{\sigma_B^2}{\sigma_A^2}\dfrac{s_A'^2}{s_B'^2} \to \mathcal{F}_{(n_A-1),(n_B-1)\ d.d.l.}$

Fractiles : $f_{Sup} = F^{n_A-1}_{n_B-1} = F^{9}_{14} = 2,65$

Table de la loi de Fisher-Snedecor $p = 0.05$ (risque global de 0,1)

$$f_{Inf} = \frac{1}{F^{14}_{9}} = \frac{1}{3,03} = 0,33$$

$$\begin{aligned}
\text{Marges d'erreur : } f_{Sup}\frac{s_B'^2}{s_A'^2} &= 2,65\frac{0,28^2 \times 15 \times 9}{0,22^2 \times 14 \times 10} = 2,65 \times 1,56 = 4,14 \\
f_{Inf}\frac{s_B'^2}{s_A'^2} &= 0,33\frac{0,28^2 \times 15 \times 9}{0,22^2 \times 14 \times 10} = 0,33 \times 1,56 = 0,515 \\
I.C._{0.90}\left(\frac{\sigma_A^2}{\sigma_B^2}\right) &= [0.515 \quad 4,14]
\end{aligned}$$

Comme $1 \in I.C._{0.90}\left(\frac{\sigma_A^2}{\sigma_B^2}\right) = [0.515 \quad 4,14] \Longrightarrow \sigma_A \approx \sigma_B$

σ_A^2, σ_B^2 inconnues et supposée égales $\sigma_A = \sigma_B \quad n_A < 30,\ n_B < 30$

$I.C._{95\%}(\mu_1 - \mu_2)$ =? (Table 6, p. 76)

Statistique de test : $\dfrac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{s'^2\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \to T_{(n_A+n_B-2)d.d.l.}$

$$s'^2 = \frac{n_A\, s_A^2 + n_B\, s_B^2}{n_a + n_B - 2} = \frac{10 \times 0,22^2 + 15 \times 0,28^2}{10 + 15 - 2} = 0,072$$

Fractile $t_{St\frac{\alpha}{2}}$: $t_{0,1;(10+15-2)} = t_{0,1;\,(23)} = 1,7139$

Marge d'erreur : $E = t_{St\frac{\alpha}{2}}\, s' \sqrt{\dfrac{1}{n_a} + \dfrac{1}{n_B}} = 1,7139 \times 0,072 \sqrt{\dfrac{1}{10} + \dfrac{1}{15}} = 1,74$

$$\begin{aligned}
I.C._{0,90}(\mu_A - \mu_B) &= (\bar{x}_A - \bar{x}_B) \pm E = 0,9 \pm 1,74 \\
I.C._{0,90}(\mu_A - \mu_B) &= [-0,84 \quad 2,64]
\end{aligned}$$

Comme $0 \in I.C._{0,90}(\mu_A - \mu_B) = [-0,84 \quad 2,64] \Longrightarrow \mu_A \approx \mu_B$

d) échantillons appariés

$n = 51\,;\ \bar{x}_A = 7,55\,;\ \bar{x}_B = 6,28\,;\ SCE_{X_A - X_B} = 12,25 \quad I.C._{99\%}(\mu_A - \mu_B)$ =? Table 6

$$Z = X_A - X_B; \quad \bar{Z} = \bar{X}_A - \bar{X}_B = 7,55 - 6,28 = 1,27h.$$

$$S' = \sqrt{\frac{SCE}{n}} = \sqrt{\frac{12,25}{51}} = \sqrt{0,24} = 0,49h.$$

Statistique de test : $\dfrac{(\bar{Z} - \mu_Z)}{s'\sqrt{n}} \to T_{(n-1)d.d.l.}$

Feractile $t_{St\frac{\alpha}{2}}$: $t_{St\frac{\alpha}{2}} = t_{[0,01;50]} = 2,6778$

Marge d'erreur : $E = t_{St\frac{\alpha}{2}}\dfrac{s'}{\sqrt{n}} = 2,6778\sqrt{\dfrac{SCE}{n}}$

$$= 2,6778\sqrt{\frac{12,25}{51}} = 2,6778 \times 0,49 = 1,31$$

$$\begin{aligned} I.C._{0,99}(\mu_A - \mu_B) &= \bar{z} \pm E = 0,9 \pm 1,74 \\ I.C._{0,99}(\mu_A - \mu_B) &= [1,27 - 1,31 \quad 1,27 + 1,31] = [-0,04 \quad 2,58] \end{aligned}$$

Comme $0 \in I.C._{0,99}(\mu_A - \mu_B) = [-0,04 \quad 2,58] \Longrightarrow \mu_A \approx \mu_B$. La différence des deux moyennes n'est pas significative. Elle est due aux fluctuations d'échantillonnage.

Feuille 3 : Les tests d'hypothèse

Exemple 3.3.1 Un procédé de remplissage est ajusté de telle sorte que les contenants pèsent en moyenne 400g. Le poids des contenants est supposé normalement distribué avec un écart-type de 8g. Pour vérifier si le procédé de remplissage se maintient à 400g, en moyenne, on opte pour la règle décision suivante sur un échantillon prélevé de 16 contenants : Le processus opère correctement si : 396.08 g $\leq \bar{X} \leq$ 403.92 g Sinon arrêter le processus de remplissage.
a) Quelles sont les hypothèses statistiques que l'on veut tester avec cette méthode de contrôle ?
b) Déterminer la probabilité de commettre une erreur de première espèce.
c) Lors d'un récent contrôle, on a obtenu, pour un échantillon de 16 contenants, un poids moyen de 395g. Doit-on poursuivre ou arrêter la production ?
d) Quelle est la probabilité de commettre une erreur de deuxième espèce selon l'hypothèse alternative $H_1 : \mu = 394g$?
e) Avec ce plan de contrôle, quelle est la probabilité de rejeter l'hypothèse selon laquelle le procédé opère à 400g, alors qu'en réalité il opère à 394g ?
f) Faire de même pour les valeurs suivantes sous $H_1 : \mu =$ 395g, 396g, 397g, 398g, 399g et 400g. Tracer la courbe d'efficacité du test.

Exemple 3.3.2 Une entreprise fournit à un client des tiges d'acier. Le client exige que les tiges aient en moyenne, une longueur de 29 mm. On admet que la longueur des tiges est normalement distribuée. On veut vérifier si le procédé de fabrication opère bien à 29 mm. Un échantillon aléatoire de 12 tiges provenant de la fabrication donne une longueur moyenne de 27.25 mm et un écart-type empirique de 2.97 mm. Doit-on conclure, au seuil $\alpha = 5\%$, que la machine est déréglée ?
1. Hypothèses statistiques :
2. Seuil de signification :
3. Statistique de test :
4. Calcul de la statistique de test sous l'hypothèse nulle H_0 :
5. Règle de décision :

Exemple 3.3.3 Aux dernières élections, un parti politique a obtenu 42% des suffrages. Un récent sondage a révélé que, sur 1041 personnes interrogées en âge de voter, 458 accorderaient son appui à ce parti. Le secrétaire général du parti a déclaré que la popularité de son parti est en hausse.
Que penser de cette affirmation au seuil de signification $\alpha = 5\%$?
1. Hypothèses statistiques :
2. Seuil de signification :
3. Conditions d'application du test :
4. Statistique de test :
5. Calcul de la statistique de test sous l'hypothèse nulle H_0 :
6. Règle de décision :
7. Décision et conclusion :

Exemple 3.3.4 Le responsable de la production suggère au client des tiges d'acier avec un nouvel alliage. Il semble que ceci permettrait d'obtenir une résistance à la rupture plus élevée. Les résultats d'un test de résistance à la rupture de 50 tiges avec et sans le nouvel alliage se résument comme suit.

	Sans le nouvel alliage	Avec le nouvel alliage
Nombre de tiges	50	50
Résistance moyenne	600.50	605.00
Variance empirique	148.50	137.61

Au seuil de signification $\alpha = 5\%$, est-ce que l'hypothèse selon laquelle la résistance moyenne à

la rupture sans l'alliage est moins élevée que celle avec l'alliage est confirmée ?
1. Hypothèses statistiques :
2. Seuil de signification :
3. Conditions d'application du test :
4. Statistique de test :
5. Calcul de la statistique de test sous l'hypothèse nulle H_0 :
6. Règle de décision :
7. Décision et conclusion :

Exemple 3.4.1 Pour sa fabrication, un industriel utilise des pièces de deux constructeurs différents. Après six mois d'utilisation, il constate que sur les 80 pièces du constructeur 1, 50 ne sont jamais tombées en panne, alors que pour le constructeur 2 la proportion est de 40 sur 60. Au seuil de signification $\alpha = 5\%$, peut-on considérer que les proportions de pièces de ces deux constructeurs sont équivalentes ?
1. Hypothèses statistiques :
2. Seuil de signification :
3. Conditions d'application du test :
4. Statistique de test :
5. Calcul de la statistique de test sous l'hypothèse nulle H_0 :
6. Règle de décision :
7. Décision et conclusion :

Exemple 3.5.1 Test d'indépendance : taux de guérison et coût du médicament.
Pour comparer l'efficacité de 2 médicaments comparables, mais de prix très différents, la Sécurité sociale a effectué une enquête sur les guérisons obtenues avec ces deux traitements. Les résultats sont présentés dans le tableau suivant :

	Original	Générique	Total
Guérisons	156	44	200
Non-guérisons	44	6	50
Total	200	50	250

Tableau aux effectifs observés n_{ij}

Au seuil de signification $\alpha = 5\%$, peut-on conclure que ces deux médicaments ont la même efficacité ?
1. Hypothèses statistiques :
2. Seuil de signification :
3. Conditions d'application du test :
4. Degré de liberté :
5. Statistique de test :
6. Calcul de la statistique du $\chi^2_{calculé}$ sous l'hypothèse nulle H_0 :
7. Règle de décision et conclusion :

Tests paramétriques

Exercices

28. [7] Le fabricant d'un médicament breveté affirmait qu'il était efficace à 90 % pour guérir une allergie en 8 heures. Dans un échantillon de 200 personnes atteintes par cette allergie, on en a gueri 160 par le médicament. Déterminer si l'affirmation du fabricant est légitime.

29. [7] La durée de vie moyenne d'un échantillon de 100 ampoules fluorescentes fabriquées par une usine est estimee à 1750 heures avec un écart-type de 120 heures. Si μ est la durée

de vie moyenne de toutes les ampoules produites par l'usine,
a) tester l'hypothèse $\mu = 1600$ heures avec l'hypothèse $\mu \neq 1600$ heures en choisissant un niveau de signification de 0.05.
b) tester l'hypothèse $\mu = 1600$ heures avec l'hypothèse $\mu \neq 1600$ heures en choisissant un niveau de signification de 0.01 si l'échantillon est composé de 20 ampoules
c) tester l'hypothèse $H_0 : \mu = 1600$ heures contre l'hypothèse $H_1 : \mu < 1600$ heures, en choisissant un seuil de signification de 0.05 et si l'échantillon est composé de 100 ampoules
d) tester l'hypothèse $H_0 : \mu = 1600$ heures contre l'hypothèse $H_1 : \mu > 1600$ heures, en choisissant un seuil de signification de 0.01 et si l'échantillon est composé de 20 ampoules

30. [7] Une machine a produit dans le passé des rondelles ayant une épaisseur de 0.05 cm. Pour déterminer si la machine est encore en état de marche, on choisit un échantillon de 10 rondelles dont les épaisseurs ont une moyenne de 0.053 cm et un écart-type de 0.003 cm. Tester l'hypothèse qui affirme que la machine est en état de marche au seuil de signification de
a) 0.05
b) 0.01

31. [7] L'écart-type de la charge d'une balance correspondant à des colis de 40.0 kilogrammes a été dans le passé de 0.25 kg. Un échantillon de 20 colis tiré au hasard indique un écart-type de 0.32 kg. L'accroissement apparent de la variabilité est-il significatif aux seuils de signification
a) de 0.05
b) de 0.01 ?

Test unilatéral à gauche

32. Un ciment est fabriqué pour présenter une résistance de 30 MPa (valeur de design). On suppose que la distribution de la résistance du ciment est $X \sim \mathcal{N}(30, \sqrt{20})$. Un échantillon de 5 éprouvettes a fourni les valeurs suivantes : 30.1, 29.5, 29.6, 28.4, 28.9.
a) Peu-t-on rejeter l'hypothèse avec $\alpha = 0,05$ que le ciment dans son ensemble a une résistance de 30 MPa sur la seule foi de ces 5 échantillons ?
b) La même question si l'échantillon contenait 20 observations au lieu de 5, toujours présentant la même moyenne.
b) Peu-t-on rejeter l'hypothèse que le ciment dans son ensemble a une résistance de 30 MPa sur le 5-échantillon avec $\alpha = 0.05$ en supposant que la variance de la résistance du ciment est inconnue : $X \sim \mathcal{N}(30, \sigma)$?

Test unilatéral à droite

33. Nous étudions le tableau donnant la répartition de 200 étudiants suivant le sexe et la couleur des cheveux, en supposant qu'ils ont été tirés au hasard dans l'ensemble des étudiants de l'université. Le tableau est le suivant :

	Cheveux blonds (j = 1)	Cheveux bruns (j = 2)	Autre couleur (j = 3)	Effectifs marginaux
Masculin (i = 1)	25	51	17	n1. = 93
Féminin (i = 2)	62	31	14	n2. = 107
Effectifs marginaux	n.1 = 87	n.2 =82	n.3 = 31	200

Peut-on considérer comme vraisemblable, avec un risque d'erreur de 5%, l'hypothèse selon laquelle le sexe et la couleur des cheveux sont indépendants ?

Tests non-paramétriques

Test de khi-deux d'ajustement ou d'adéquation

34. A un age donné, on a pu déterminer que : 50 % des bébés normaux marchent, 12 % ont une ébauche de marche, 38 % ne marchent pas.
Population étudiée : Les bébés prématurés.
Observations :
On a observe 80 prématurés à l'age donné : 35 de ces bébés marchent, 4 ont une ébauche de marche, 41 ne marchent pas
Les bébés prématurés développent-ils la marche de la même manière que les bébés normaux ?

Test de khi-deux d'ajustement de conformité

35. Équiprobabilité des sexes à la naissance
L'étude de 320 familles ayant 5 enfants s'est traduite par la distribution suivante :

Classe	A	B	C	D	E	F	Total
Nombre de garçons	5	4	3	2	1	0	
Nombre de filles	0	1	2	3	4	5	
Nombre de familles	18	56	110	88	40	8	320

On veut comparer cette distribution à la distribution théorique qui correspond à l'équiprobabilité de la naissance d'un garçon et de la naissance d'une fille.
a) Quelle est la loi de probabilité du nombre de garçons dans une famille de cinq enfants, dans l'hypothèse d'équiprobabilité des naissances des garçons et des filles.
b) La comparaison de la distribution observée à la distribution théorique s'effectue par un test Khi deux. Que peut-on en conclure ?

36. Influence de la place de départ dans une course
Au départ d'une course de chevaux, il y a habituellement huit positions de départ et la position numéro 1 est la plus proche de la palissade. On soupçonne qu'un cheval a plus de chances de gagner quand il porte un numéro faible, c'est-à-dire qu'il est plus proche de la palissade intérieure. Voici les données de 144 courses :

									Total
Numéro de départ	1	2	3	4	5	6	7	8	8
Nombre de victoires d'un cheval ayant ce numéro	29	19	18	25	17	10	15	11	144

a) Poser les hypothèses à tester (hypothèse nulle et hypothèse alternative).
b) Calculer le khi deux observé et la probabilité critique. Conclure.

Test de khi-deux d'indépendance

37. Pour l'étude de la relation entre le niveau d'étude et le fait de subir ou non un chômage de longue durée, on a fait des observations sur un échantillon de 100 individus. Pour chaque individu, on a relevé le niveau d'étude : « secondaire » ou « supérieur » et s'il a subi un chômage de longue durée : « oui » ou « non ». On observe que : 40 ont un niveau d'étude secondaire et ont subi un chômage long ; 26 ont un niveau d'étude secondaire et n'ont pas subi un chômage long ; 12 ont un niveau d'étude supérieur et ont subi un chômage long ; 22 ont un niveau d'étude supérieur et n'ont pas subi un chômage long.
a) Représenter la répartition des 100 sujets selon le niveau d'étude et le fait de subir ou non un chômage long par un tableau de contingence.
b) Peut-on en conclure qu'il existe un lien entre le niveau d'étude et le fait de subir ou non un chômage long ?

38. On désire tester l'effet d'une antibiothérapie systématique sur l'apparition d'une infection post-opératoire. Une expérience randomisée est conduite. Un premier groupe de patients reçoit une antibiothérapie. Un deuxième groupe reçoit un placebo. Les résultats sont les suivants :

	Sujets ayant reçu une antibiothérapie	Sujets ayant reçu un placebo
Infection postopératoire	10	29
Pas d'infection post-opératoire	75	27

L'antibiothérapie est-elle efficace dans la prévention des complications infectieuses ?

Tests de khi-deux d'homogénéité

39. Dans un échantillon de 400 femmes et un échantillon de 300 hommes, on observe que 25 femmes et 25 hommes développent une certaine forme de maladie mentale.
Peut-on dire que cette forme de maladie n'atteint pas les femmes et les hommes de la même façon ?

40. Les observations d'une variable qualitative sur k échantillons permettent-elles de conclure que les échantillons proviennent de la même population
Existe t-il un lien entre le nombre de grossesse et le décès des bébés ?
Fréquences observées :

Age du décès	Nombre de grossesses inférieur à 3	Nombre de grossesses supérieur à 3
Inférieur à 3 mois	18	6
Supérieur à 3 mois	17	19

Feuille 4 : Préparation pour les contrôles

Exercices

Contrôle 1.

41. [2] La SGM souhaite mieux connaitre la répartition des impayés dans son portefeuille de clients. Sur l'ensemble des 20000 dossiers traités annuellement au service contentieux, un échantillon aléatoire de 30 dossiers a été prélevé aux fins d'étude, qui a permis d'obtenir un montants moyen observé d'impayés de 2660,50 K€ et un écart-type observé des impayés de 279,66 K€.
a) Quelle serait la probabilité pour que, sur l'ensemble des dossiers, le montant moyen d'impayés soit inférieur à 2300 K€ ?
b) Quel serait l'intervalle de confiance à 95% de cette moyenne et quelle en serait l'interprétation ?
c) Quel serait l'intervalle de confiance à 95% de l'écart-type des impayés et quelle en serait l'interprétation ?
d) Quel est le risque d'erreur que l'on attribue à l'intervalle de confiance, bilatéral symétrique du montant moyen d'impayés : [2539,5 - 2781,497] obtenu àn pratir de cette série de 30 dossiers.
e) Quel serait l'intervalle de confiance à 95% de la moyenne de la population, obtenu à la base des observations d'un échantillon de 25 dossiers, dont la moyenne observée d'impayés est de 2600 K€ et l'écart-type observé est de 277 K€.
f) Quel serait l'intervalle de confiance à 99% de l'écart-type des dossiers impayés de la population, obtenu à la base des observations d'un échantillon de 200 dossiers, dont la moyenne observée d'impayés est de 2650 K€ et l'écart-type observé est de 280 K€.

42. [2] 96% des ménages français possèdent un réfrigérateur.
a) Quelle est la probabilité pour que, dans un échantillon de 1 200 ménages, la fréquence relative soit comprise entre 0,95 et 0,97. Que pourrait-on dire si la fréquence relative de l'échantillon était de 0,99 ?
b) Quelle doit être la taille de l'échantillon pour que la probabilité de trouver une fréquence relative de l'échantillon comprise entre 0,95 et 0,97 soit de 99%.

Contrôle 2.

43. Une société de gérance de projets a demandé à une firme d'expertises en contrôle de matériaux, d'évaluer la qualité d'un mélange bitumineux provenant de deux usines. Il a été convenu d'effectuer une vérification en évaluant la résistance à la compression, à l'âge de 3 jours, sur des cylindres de béton. La résistance à la compression est supposée normalement distribuée. Les résultats pour les deux usines se résument comme suit :

	Usine 1	Usine 2
Nombre de cylindres	$n_1 = 16$	$n_2 = 12$
Résistance moyenne (kg/cm^2)	$\bar{x}_1 = 90,6$	$\bar{x}_2 = 96,1$
Somme des Carrés des Écarts à la moyenne	SCE1 = 1200	SCE2 = 1068

a) Peut-on considérer comme vraisemblable, avec un risque d'erreur de 5%, l'hypothèse selon laquelle les variances des résistances à la compression des cylindres de ces deux usines sont identiques ?
b) Est-ce que la firme d'expertises peut affirmer, au risque de 5%, que le mélange bitumineux de l'usine 1 est moins résistant à la compression que celui de l'usine 2 ?

44. Euro-pratique : Avant le passage à l'euro, un sondage a montré que 50% des achats sont effectués avec une carte bancaire. Depuis le passage à l'euro, un récent sondage effectué sur un échantillon de 500 personnes choisies au hasard a révélé que 270 personnes utilisent leur carte bancaire.
a) Peut-on conclure, avec un risque d'erreur $\alpha = 5\%$, que la proportion d'utilisateurs de cartes bancaires est restée stable depuis le passage à l'euro ?

45. On a compté le nombre de fruits portés par des arbres choisis au hasard dans deux parcelles. On suppose que le nombre de fruits par arbre est une variable aléatoire approximativement normale. Les résultats pour les deux parcelles se résument comme suit.

	Parcelle I	Parcelle II
Nombre d'arbres	$n_1 = 12$	$n_2 = 16$
Récolte moyenne par arbre observée	$\bar{x}_1 = 109,5$	$\bar{x}_2 = 77$
Somme des Carrés des Écarts à la moyenne	SCE1 = 35721	SCE2 = 20979

a) Tester, avec un risque d'erreur de 5%, l'hypothèse selon laquelle les deux parcelles ont même variance ?
b) Tester, avec un risque d'erreur de 5%, l'hypothèse d'égalité des récoltes moyennes par arbre ?
c) Peut-on affirmer, avec un risque d'erreur de 5%, que la récolte sera plus importante dans la parcelle I que dans la parcelle II ?

Le manuel "Statistique inférentielle" fait connaissance aux notions et les méthodes, lies a la quatrième phase de la méthode statistique l'Interprétation. Le manuel offre le matériel théorique et pratique, nécessaire à apprendre d'après le programme en "Statistique appliquée» des spécialités Gestion et Economie de la filière de gestion à l'Université de Sofia « Sv. Kliment Ohridski ». On y considère les notions de base et les deux groupes de méthodes de l'Interprétation - l'estimation des paramètres de la population et des tests d'hypothèses. Après un court rappel du thème de la statistique descriptive - Echantillonnage, Distribution de la moyenne, de la dispersion et de la fréquence échantillonnalles, on présente les thèmes d'Estimation – estimation ponctuelles e t intervalles et de Tests d'hypothèses – test paramétriques et non-paramétriques.

Le manuel a pour but de donner des connaissances théoriques et de développer des compétences pratiques pour le choix de modèles convenables pour tester d'hypothèses et prendre de décisions.

eISSN 2367-8666