Chapitre 4

Les statistiques inférentielles

L'étude statistique portant sur tous les éléments d'une population étant, soit impossible à réaliser (à cause d'un trop grand nombre d'individus à étudier), soit trop onéreuse, il faut obtenir des résultats fiables sur les caractéristiques d'une population en se limitant à l'étude des éléments ou unités d'un échantillon. Mathématiquement, l'opération est inhabituelle. D'habitude, on travaille sur un ensemble de grande taille pour déduire des informations sur un ensemble plus petit inclus dans celui-ci. La démarche de l'inférence consiste à partir de l'ensemble plus petit pour déduire des informations sur l'ensemble de grande taille.

Les mots **inférence** et **inférentiel** sont utilisés pour décrire les situations dans lesquelles on cherche, grâce aux données, à tirer des conclusions précises concernant les paramètres inconnus de la population statistique étudiée. Un grand nombre d'applications de la statistique sont de ce type, souvent sous la forme de **tests de signification**.

La théorie de l'estimation est nécessaire lorsque l'information totale dans une population est impossible, comme l'ensemble des poissons contenus dans l'océan. Cette population appelée **population mère** doit être **échantillonnée**. Échantillonner consiste à prélever dans une population mère une partie de celle-ci au hasard avec une **taille** n **fixée**. Lorsque c'est possible, il est évidemment nécessaire de réaliser plusieurs échantillons. Chaque échantillon fournit alors un résultat. La variation des résultats s'appelle **fluctuation de l'échantillonnage**. Dit autrement, à partir d'un échantillon, on ne possède aucune certitude sur les paramètres statistiques ; il convient ainsi de les estimer selon le **principe de vraisemblance** [1].

Pour finir cette introduction, quelques termes doivent être définis. L'**enquête** est une opération consistant à questionner ou observer les individus d'un échantillon. Le **recensement** est une enquête exhaustive. Le **sondage** est une enquête sur un un échantillon représentatif de la population.

1. Les méthodes d'inférence par le principe de vraisemblance ne sont pas les seules envisageables. Il existe également l'**inférence bayésienne**.

4.1 L'échantillonnage

Un **échantillon** est un groupe restreint, c'est-à-dire un sous-ensemble, issu d'une variable aléatoire X de la population. Cette population d'où est tiré l'échantillon est appelée **population mère**. Il existe différentes façons de tirer un échantillon de la population mère. La plupart nécessitent de disposer une **base de sondage**. La plupart des cas montrent qu'il est impossible de constituer cette base de sondage, on passe alors par une autre méthode.

De fait, on utilise un **échantillon aléatoire** qui offre des résultats recueillis sur ce sous-ensemble qui doivent pouvoir être étendus, c'est-à-dire **inférés**, à la population mère. Pour définir un tel échantillon, une méthode consiste à prélever, au hasard, un sous-ensemble d'individus. On distingue deux types d'échantillons aléatoires : l'échantillon non biaisé et l'échantillon biaisé. Un **échantillon non biaisé** est tiré au hasard dans lequel tous les individus ont la même chance de se retrouver dans l'échantillon. Le tirage est équiprobable. À l'opposé, dans un **échantillon biaisé** (ou non équiprobables), les éléments n'ont pas été pris au hasard.

Un échantillon est utile lorsqu'il n'est pas possible de tenir compte de tous les avis de la population totale. Par exemple, il est impossible de connaître l'intention de vote de 40 millions de personnes. Il faut alors interroger 1 000 personnes représentatives (erreur de 5 %), ou mieux 10 000 personnes représentatives (erreur 0,3 %).

L'échantillon doit être **« représentatif »** de la population c'est-à-dire qu'il doit être pertinent et aléatoire pour l'étude menée. Pour obtenir une telle condition, un échantillonnage aléatoire est généralement la meilleure technique. Un échantillon est représentatif si l'on peut en **généraliser les paramètres** à toute la population mère. Cette généralisation doit prendre en compte l'**erreur d'échantillonnage** (la marge d'erreurs dues aux fluctuations d'échantillonnage) c'est-à-dire à la fois le risque d'avoir un échantillon non représentatif et un certain flou dans l'extension à toute une population.

Il existe deux types d'échantillons. 1. Les échantillons sont constitués par des individus différents. Ce sont des **échantillons indépendants**. 2. S'il s'agit des mêmes individus soumis, dans un ordre tiré au sort et avec un délai suffisant, les individus sont associés deux à deux. On dit que les **échantillons** sont **appariés**.

Remarque fondamentale. Un petit échantillon représentatif est préférable à un grand échantillon biaisé.

Remarque. En règle générale, la population étudiée est toujours considéré comme un échantillon.

N.B. 1. Les lettres minuscules sont réservées aux observations et aux estimations.

Paramètre	Population mère	Échantillon
Effectif	N	n
Variable	X	x
Moyenne	μ	m
Variance	σ^2	s^2
Écart type	σ	s
Fréquence	ϖ	f

TABLE 4.1 – Notations particulières matérialisant les relations entre la population mère et un de ses échantillons

N.B. 2. Les lettres majuscules sont réservées aux variables aléatoires.

N.B. 3. Les lettres grecques correspondent aux paramètres exacts d'une variable au sein d'une population.

4.2 Les méthodes d'échantillonnage

On distingue deux méthodes : 1. les **méthodes aléatoires** (c'est-à-dire qui fait appel à un tirage au sort) ; 2. les **méthodes non aléatoires**.

4.2.1 Méthodes aléatoires

Les **modèles de sondage aléatoire simple** (S.A.S.) et à équiprobabilité des individus procèdent par **tirage au sort**.

Tirer au sort des individus dans une population mère implique que l'on dispose d'une **base de sondage**, c'est-à-dire une liste et une localisation des individus numérotés de 1 à N. On effectue alors le tirage au sort de numéros de 1 à N avec des dés, une roue de loterie, une table de nombres au hasard, la fonction *random* d'une calculatrice ou d'un logiciel statistique. Les numéros tirés au hasard désignent les individus composant l'échantillon.

Il existe théoriquement deux façons de sélectionner des individus au hasard.

1. Le **tirage avec remise** implique que l'on tire un numéro, que l'on note et que l'on ne le raye pas de la liste. Ce modèle est irréaliste, mais il est le plus simple du point de vue des calculs ultérieurs : le taux de sondage $\frac{n}{N}$ n'y intervient pas. On parle aussi d'un **échantillonnage non exhaustif**.
2. Le **tirage sans remise** suppose le même déroulement des opérations mais un numéro déjà tiré est rayé de la liste et ne peut plus apparaître dans les tirages ultérieurs. C'est un modèle plus réaliste mais qui complique quelque peu les calculs : il faut tenir compte du taux de sondage $\frac{n}{N}$, c'est-à-dire de la

représentation de la fraction de la population constituant l'échantillon. S'il est fort, il rend les estimations plus précises. On parle aussi d'un **échantillonnage exhaustif**.

Chaque observation individuelle dans un échantillon aléatoire possède la même distribution de probabilité que sa population mère. À partir de la distribution, on calcule la moyenne μ et l'écart type de la population mère, lorsque la taille de celle-ci est connue. Dès lors, on peut étudier une caractéristique mesurable X, c'est-à-dire une variable aléatoire au sein d'une population de taille finie ou infinie, caractérisée par deux paramètres μ et σ. La composition de la population, vis-à-vis du caractère X, est entièrement définie par la connaissance des quantités $\varpi(x)$, qui est la **proportion des individus** tels que $X < x$, pour toutes les valeurs de $x \in \mathbb{R}$.

Soit E l'expérience consistant à choisir au hasard un élément de la population. Avant le tirage, on se propose de prévoir la valeur du caractère X que l'on obtiendra. Ce caractère est une variable aléatoire X telle que $Pr(X < x) = \varpi(x)$ pour toute valeur $x \in \mathbb{R}$. À l'expérience E, est associée une variable aléatoire X dont la fonction de répartition est $F(x)$.

Le premier échantillon tiré de taille n est noté, de manière indicielle, $x_1^1, x_1^2, \ldots, x_1^i, \ldots, x_1^n$. Pour bien comprendre cette notation, il est important de préciser que l'exposant est un indice renvoyant au numéro de l'échantillon, tandis que l'indice renvoie à la position de l'individu dans la liste. Lorsque le tirage est accompli, il est possible de calculer une moyenne de cet échantillon, notée soit $\hat{\mu}_1$, soit $\bar{x}_1$ – l'indice ici correspond au numéro de l'échantillon. Sa valeur est une **estimation ponctuelle** de la moyenne μ. De fait, si l'échantillon est de taille suffisante, alors $\hat{\mu}_1$ est proche de la valeur de la moyenne exacte μ dans la population. Ce type d'échantillon est appelé **échantillon aléatoire très simple** (E.A.T.S.).

Cela étant, il existe une infinité de tirages, donc d'échantillons aléatoires de taille n dans la population. Dit autrement, il est possible de réaliser n fois la même expérience E, dans des conditions indépendantes, par exemple en remettant dans la population l'élément tiré. À ces n expériences, on associe n variables aléatoires indépendantes X_i, suivant la même loi que la variable aléatoire X. On obtient une matrice regroupant les k tirages (Tab. 4.2). X_i correspond à la première colonne, donc au premier élément de chaque tirage. Par exemple, X_1 est la variable aléatoire du premier élément tiré dans l'ensemble des k échantillons. X_2 est la variable aléatoire du deuxième élément tiré dans l'ensemble des k échantillons. X_k est la variable aléatoire du n-ième élément tiré dans l'ensemble des k échantillons. Dit autrement, il est proposé ici n variables aléatoires indépendantes et identiquement distribuées de même loi X, appelée **variable aléatoire parente**. La liste $(X_1, X_2, \ldots, X_n)$ est un n**-échantillon**, qui est lui-même une variable aléatoire. La moyenne $\bar{X}_n$ du n-échantillon vaut :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4.1}$$

Position dans le tirage	**Premier élément**	**Deuxième élément**	**...**	**l-ième élément**	**...**	**n-ième élément**	**Moyenne de l'échantillon**
Échantillon n° 1	x_1^1	x_2^1	...	x_l^1	...	x_n^1	$\bar{x}_1$
Échantillon n° 2	x_1^2	x_2^2	...	x_l^2	...	x_n^2	$\bar{x}_2$
...	...	...	...	...	...	...	...
Échantillon n° j	x_1^i	x_2^i	...	x_l^j	...	x_n^j	$\bar{x}_j$
...	...	...	...	...	...	...	...
Échantillon n° k	x_1^k	x_2^k	...	x_l^k	...	x_n^k	$\bar{x}_n$
Variable aléatoire X_i	X_1	X_2	**...**	X_i	**...**	X_n	$\bar{X}_n$

TABLE 4.2 – Variable aléatoire des n-échantillons

Tout échantillon est une **estimation ponctuelle** de la variable X, tandis que le n-échantillon est un **estimateur**.

Loi de la moyenne. La moyenne de variables aléatoires M suit une loi normale, même si la distribution de départ n'est pas normale d'espérance $\mathbb{E}(M) = \mu$, de variance $\mathbb{V}(M) = \frac{\sigma^2}{n}$ et d'erreur $\mathrm{SE} = \sigma(M) = \frac{\sigma}{\sqrt{n}}$.

La méthode d'échantillonnage aléatoire est parfaite pour des études où la répétition des tirages n'engendre qu'un coût limité. Il convient par conséquent d'utiliser d'autres méthodes moins onéreuses, mais tout aussi fiable.

Remarque. Une réalisation de l'échantillon sera notée $(x_1, x_2, \ldots, x_n)$. Cette liste n'est alors plus une variable aléatoire.

Étude de la statistique de la moyenne du n-échantillon $\bar{X}$

On suppose que les moments d'ordre 1 et 2 de la variable aléatoire parente X existent. On pose $\mathbb{E}(X) = \mu$ et $\mathbb{V}(X) = \sigma^2$.

Dans le cas d'une **population infinie**, l'espérance et la variance de $\bar{X}$ valent respectivement :

$$\mathbb{E}\left(\bar{X}\right) = \mu \tag{4.2}$$

et

$$\mathbb{V}\left(\bar{X}\right) = \frac{\sigma^2}{n} \tag{4.3}$$

où n est la taille de l'échantillon.

Soit N la taille de la population mère. Le tirage s'effectue **sans remise**. Dans le cas d'une **population finie**, l'espérance et la variance de $\bar{X}$ valent respectivement :

$$\mathbb{E}\left(\bar{X}\right) = \mu \tag{4.4}$$

et

$$\mathbb{V}\left(\bar{X}\right) = \frac{N-n}{N-1}\frac{\sigma^2}{n} \approx \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n} \tag{4.5}$$

L'erreur sur la moyenne vaut alors :

$$\sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}} \tag{4.6}$$

Le comportement asymptotique de $\bar{X}$ Par la loi faible des grands nombres, $\bar{X}$ converge en probabilité vers μ lorsque n tend vers l'infini.

Par la loi forte des grands nombres, $\bar{X}$ converge presque sûrement en probabilité vers μ lorsque n tend vers l'infini, car la série $\sum_{i=1}^{n} \frac{\sigma^2}{i^2} = \sigma^2 \sum_{i=1}^{n} \frac{1}{i^2}$ est convergente. Dans ce cadre, si les moments d'ordre 1 et 2 existent, on peut appliquer le théorème central limite à la variable aléatoire Y_n :

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \tag{4.7}$$

La variable Y_n converge en loi vers une variable suivant la loi normale $N(0,1)$ lorsque n tend vers l'infini.

Étude de la statistique de la variance du n-échantillon $\hat{s}$

On suppose que les moments d'ordre 1 et 2 de la variable aléatoire parente X existent. On pose $\mathbb{E}(X) = \mu$ et $\mathbb{V}(X) = \sigma^2$.

La variance empirique $\hat{s}^2$ d'un échantillon de taille n est définie par :

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \tag{4.8}$$

Il est possible d'écrire la variance empirique en introduisant la moyenne de la population mère.

$$\hat{s}^2 = \frac{1}{n}\left[\sum_{i=1}^{n}\left(X_i - \mu\right)^2\right] - \left(\mu - \bar{X}\right)^2 \tag{4.9}$$

L'espérance de la variance empirique est :

$$\mathbb{E}\left(\hat{s}^2\right) = \frac{n-1}{n}\sigma^2 \tag{4.10}$$

Si la taille n de l'échantillon est grande, cette espérance a pour valeur limite :

$$\mathbb{E}\left(\hat{s}^2\right) = \frac{1}{n}\left(\mu_4 - \sigma^4\right) \tag{4.11}$$

μ_4 est le moment d'ordre 4 de la variable aléatoire X issue de la population mère.

La variance de la variance empirique est :

$$\mathbb{V}\left(\hat{s}^2\right) = \frac{n-1}{n^3}\left[(n-1)\,\mu_4 - (n-3)\,\sigma^4\right] \tag{4.12}$$

Si la taille n de l'échantillon est grande, la variance de $\hat{s}^2$ a pour valeur limite :

$$\mathbb{V}\left(\hat{s}^2\right) = \frac{\mu_4 - \sigma^4}{n} \tag{4.13}$$

Le comportement asymptotique de $\hat{s}^2$ La variable $\frac{\hat{s}^2 - \left(1-\frac{1}{n}\right)\sigma^2}{\sqrt{\mathbb{V}(\hat{s}^2)}}$ converge en loi vers une variable suivant la loi normale $N(0,1)$ lorsque n tend vers l'infini.

En prenant les limites de l'espérance et de la variance pour n grand, on obtient la variable aléatoire $\frac{\hat{s}^2-\sigma^2}{\sqrt{\frac{\mu_4-\sigma^4}{n}}}$ qui converge en loi vers la loi normale $N(0,1)$.

La corrélation entre $\bar{X}$ et $\hat{s}^2$ Pour définir la corrélation entre $\bar{X}$ et $\hat{s}^2$, on calcule la covariance entre ces deux variables aléatoires :

$$\operatorname{cov}\left(\bar{X}, \hat{s}^2\right) = \frac{n-1}{n^2}\mu_3 \tag{4.14}$$

μ_3 est le moment centré d'ordre 3 de la variable aléatoire X.

Trois cas sont possibles.

1. Si n tend vers l'infini, la covariance entre ces variables tend vers 0, les statistiques $\bar{X}$ et $\hat{s}^2$ sont asymptotiquement non corrélées.
2. Si la distribution de la variable X est symétrique, le moment centré μ_3 est égal à 0, les statistiques $\bar{X}$ et $\hat{s}^2$ sont non corrélées quelle que soit la valeur de n.
3. Si, de plus, X suit une loi normale, les statistiques $\bar{X}$ et $\hat{s}^2$ sont indépendantes quelle que soit la valeur de n.

4.2.2 Méthodes non aléatoires

Les méthodes de sondage non aléatoires tentent de fabriquer des « modèles réduits » d'une population mère en utilisant d'autres procédés que le tirage au sort.

L'échantillonnage systématique

L'échantillonnage systématique suppose l'existence d'une **base de sondage** dans laquelle les individus de la population mère sont numérotés de 1 à N.

La méthode consiste à :

1. déterminer la taille n de l'échantillon ;
2. fixer le pas de sondage $k = \frac{n}{N}$;
3. choisir un numéro de départ d entre 1 et k. L'échantillon sera constitué des individus numérotés : $d, d+k, d+2k, \ldots, d+(n-1)\,k$.

Cette méthode fournit des résultats au moins aussi précis que les sondages aléatoires simples. Elle ne peut pas donner lieu aux mêmes calculs, sauf si le pas de sondage correspond à une régularité dans la base de sondage.

La méthode des quotas

La **technique des quotas** permet d'obtenir des échantillons non biaisés. Elle correspond à un échantillon de la population qui respecte la proportion d'éléments distinctifs de sa population totale. Elle est fondée sur le respect, dans l'échantillon, des proportions connues dans la population mère pour des variables corrélées avec la variable à recueillir par sondage. Dit autrement, l'échantillon est un modèle réduit de la population mère. Si la méthode est conduite correctement, les résultats sont plus précis que ceux avec un sondage aléatoire simple.

4.2.3 Méthodes d'échantillonnage « Monte Carlo »

Elle propose une transformation des valeurs en moyenne obtenues par un échantillon. L'objectif est de déterminer une estimation fiable de μ pour cette variable aléatoire. La méthode confirme la règle d'approximation normale.

Dans des situations mathématiquement intraitables, elle fournit souvent le seul moyen pratique afin de déterminer les distributions d'échantillonnage.

D'un point de vue positionnement dans le cours, cette partie devrait figurer en fin de chapitre pour en comprendre tous les enjeux et tous les termes.

L'échantillonnage d'une petite population à l'aide de nombres pris au hasard

Exemple : la distribution de l'absentéisme des 100 employés d'une petite entreprise d'ameublement.

On pose la variable aléatoire X correspondant au nombre de jours d'absence au cours du premier trimestre de 1990, à l'exclusion des congés de longue maladie.

Distribution de la population			
Jours d'absences X	**Effectif**	**Fréquence** $\Pr(x)$	**Intervalles avec l'effectif relatif**
$x_1 = 0$	30	$\Pr(X = x_1) = 0,30$	1 - 30
$x_2 = 1$	26	$\Pr(X = x_2) = 0,26$	31 - 56
$x_3 = 2$	22	$\Pr(X = x_3) = 0,22$	57 - 78
$x_4 = 3$	12	$\Pr(X = x_4) = 0,12$	79 - 90
$x_5 = 4$	7	$\Pr(X = x_5) = 0,07$	91 - 97
$x_6 = 5$	2	$\Pr(X = x_6) = 0,02$	98 - 99
$x_7 = 6$	1	$\Pr(X = x_7) = 0,01$	100 - 100
	$N = 100$	$\sum_{i=1}^{n} p_i = 1,00$	

TABLE 4.3 – Distribution de la population de X

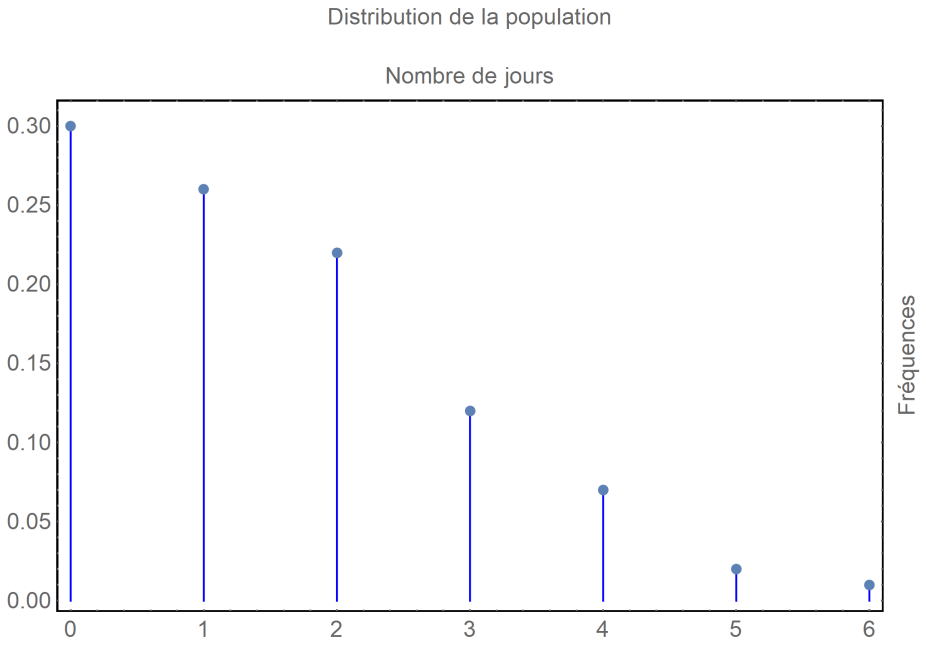


FIGURE 4.1 – Distribution du nombre de jours d'absence

La moyenne de l'échantillon vaut : $\bar{X} = 3$, c'est-à-dire le rang n° 3. Sa variance vaut : $s^2 \approx 2,1$.

L'espérance de l'échantillon se calcule avec les fréquences relatives. Elle vaut : $\mathbb{E}(X) = 1,5$. La variance des fréquences relatives vaut : $\mathbb{V}(X) \approx 1,4$.

Que se passerait-il si l'on traitait cette population de façon aveugle ? Que se passerait-il si l'on tentait d'estimer l'espérance en considérant un petit échantillon

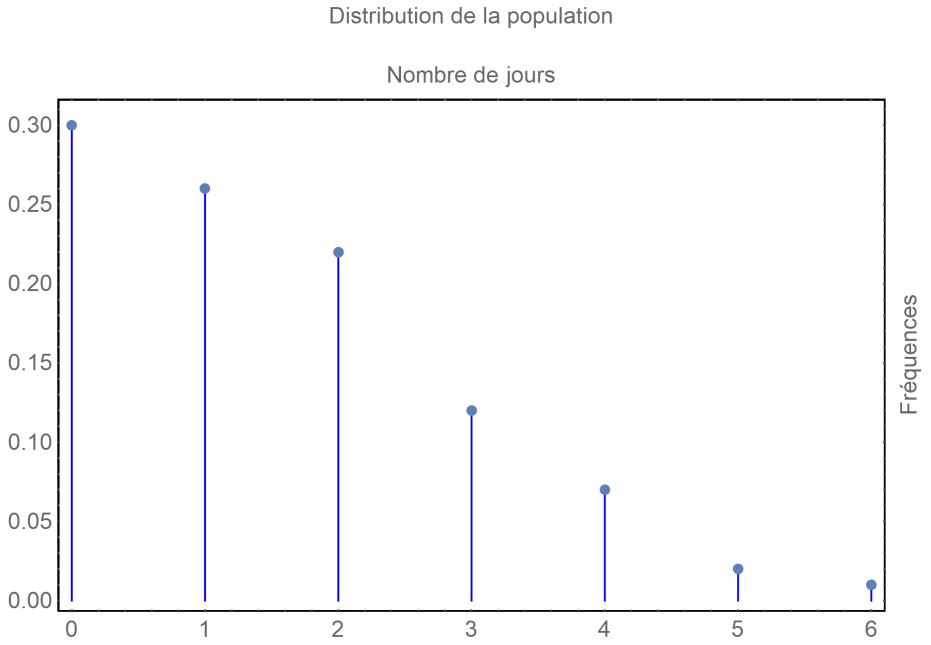


FIGURE 4.2 – Fonction de répartition du nombre de jours d'absence

aléatoire (E.A.T.S.) de $n = 5$ observations.

On crée cinq nombres au hasard en utilisant une fonction « Random [2] » par exemple. Un jet donne la liste suivante : $(37, 48, 79, 88, 74)$. Elle correspond à une chaîne dont les nombres possibles varient de 0 à N (c'est-à-dire 100). Chacun nombre correspond à un intervalle de l'effectif cumulée des jours d'absence (Tab. 4.3). Par exemple, 37 correspond à la variable aléatoire $X = x_1 = 1$ jour d'absences. On fait de même avec les quatre autres valeurs, et on obtient la liste suivante de jours d'absence : $(1, 1, 3, 3, 2)$. On calcule la moyenne des jours aléatoirement tirés de l'échantillon : $\bar{X} = 2$, ce qui est proche de l'espérance des valeurs observées dans l'échantillon de référence. On répéte l'opération à plusieurs reprises à partir d'un nombre m de tirages.

On construit la distribution d'échantillonnage montrant la manière dont $\bar{X}$ varie autour de son objectif : $\mathbb{E}(X) = 1,5$.

2. Comment créer un bon générateur ? Il existe trois méthodes :

   1. utiliser une table de nombre au hasard ;
   2. utiliser un nombre pseudo-aléatoire (avec des carrés progressifs) ;
   3. utiliser un générateur congruenciel. Par exemple, $x_{n+1} = 7^5 x_n$ modulo $2^{31} - 1$ avec n très grand.

L'échantillonnage d'une population de dimension quelconque

Dans l'échantillonnage avec remise, seule la fréquence relative importe. La taille de la population N n'est pas pertinente. C'est cette propriété qui est au cœur de la méthode de Monte-Carlo.

La méthode d'échantillonnage Monte-Carlo suppose que l'on dispose d'une source aléatoire suivant une certaine loi de probabilité désirée.

L'estimation d'une intégrale

On souhaite calculer l'intégrale $\int_0^1 f(x)\,\mathrm{d}x$.

Soit n variables aléatoires X_1, X_2, …, X_n uniformément réparties entre 0 et 1

$$\Pr(x \leq X_i < x + \mathrm{d}x) = \mathrm{d}x \tag{4.15}$$

alors $f(X_1), f(X_2), \ldots, f(X_n)$ sont des variables aléatoires.

Par définition

$$\mu = \mathbb{E}[f(X_i)] = \int_0^1 f(x)\,\mathrm{d}x \tag{4.16}$$

et

$$\sigma^2 = \int_0^1 [f(x) - \mu]^2\,\mathrm{d}x \tag{4.17}$$

donc la moyenne M des variables aléatoires vaut :

$$M = \frac{f(X_1) + f(X_2) + \ldots + f(X_n)}{n} \to \mu \tag{4.18}$$

avec un écart type caractérisé par $\frac{\sigma}{\sqrt{n}}$.

Soit n réalisations $(x_1, x_2, \ldots, x_n)$ d'une variable aléatoire uniforme entre 0 et 1. Si n est suffisamment grand, alors

$$m = \frac{f(x_1) + f(x_2) + \ldots + f(x_n)}{n} \approx \int_0^1 f(x)\,\mathrm{d}x \tag{4.19}$$

L'estimation de l'erreur

Obtenir une estimation n'est pas tout. En effectuant deux expériences similaires indépendantes l'une de l'autre, on devrait obtenir deux résultats différents puisque des nombres aléatoires différents ont été utilisés pour la génération de

variables. **Comment estimer l'erreur d'estimation commise ?** La réponse est simple : il suffit d'utiliser le théorème central limite.

L'erreur vaut la probabilité critique z_C d'une variable suivant une loi normale multipliée par la racine carrée du rapport entre la variance de l'échantillon et le nombre d'expériences indépendantes $z_C\sqrt{\frac{\mathbb{E}(X)}{n}}$.

Si on répète les tirages aléatoires un grand nombre de fois, la valeur de l'erreur finit par converger. C'est là qu'un ordinateur est un précieux outil pour opérer ces tirages et calculer moyennes et erreurs.

Appliquée aux intervalles de confiance, la méthode de Monte-Carlo permet de proposer la méthode dite *bootstrap* [3].

4.2.4 Intervalle de fluctuation

L'intervalle de fluctuation suppose que la vraie proportion théorique p soit connue. C'est un échantillonnage, et non une estimation. On estime un intervalle de fluctuation asymptotique au seuil de 95 % avec un effectif n.

L'objectif est de prendre une décision avec un risque d'erreur α % en fonction de l'appartenance ou non de la fréquence observée à l'intervalle de confiance asymptotique.

Si $n \geq 30$, $np \geq 5$ et $p\,(1-p) \geq 5$, la formule est :

$$\left[p - z_C\frac{\sqrt{p\,(1-p)}}{\sqrt{n}}, p + z_C\frac{\sqrt{p\,(1-p)}}{\sqrt{n}}\right] \tag{4.20}$$

avec z_C la probabilité critique asymptotique de la loi normale. Au seuil de 95 %, $z_C = 1,96$.

Exercice type. Une urne contient un grand nombre de boules blanches et noires. La proportion de boules blanches est $p = 0,3$/ On tire n boules de l'urne. On note X_n la variable aléatoire qui compte le nombre de boules blanches. Déterminer l'intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence $F_n = \frac{X_n}{n}$.

1. Pour $n = 50$

$$I_{50} = \left[0,3 - 1,96\frac{\sqrt{0,3\,(1-0,3)}}{\sqrt{50}}, 0,3 + 1,96\frac{\sqrt{0,3\,(1-0,3)}}{\sqrt{50}}\right] = [0,173; 0,427] \tag{4.21}$$

2. Pour $n = 500$

$$I_{500} = [0,36; 0,34] \tag{4.22}$$

3. méthode du va-et-vient

4.3 Les estimateurs et les estimations

La théorie de l'estimation permet d'**estimer les paramètres d'une loi de probabilité**. En effet, le travail du statisticien consiste, à partir d'observations, à reconstituer le modèle probabiliste d'une situation aléatoire. La première étape diagnostique le type de loi étudiée : la loi de Poisson, la loi normale, *etc*. Cette première étape se fait normalement sans trop de difficultés, car chaque loi a son champ d'application spécifique. La seconde étape consiste à estimer les divers paramètres attachés à cette loi.

Une population d'individus est caractérisée par une proportion du caractère ϖ, sa moyenne μ, son écart type σ, et plus généralement par un paramètre θ. Le problème est que la taille N de la population mère est souvent inconnue. De fait, on tente d'approximer ses valeurs par un échantillon de taille n, ayant $\hat{\varpi}$, sa moyenne $\hat{\mu}$, son écart type $\hat{\sigma}$, et plus généralement un paramètre $\hat{\theta}$. Ses valeurs de l'échantillon sont dites **estimées**. Elles permettent de proposer une estimation des valeurs inconnues de la population en étudiant les biais que ces paramètres pourraient avoir par rapport aux valeurs exactes. Pour ce, un intervalle de confiance peut être construit afin de lier $\hat{\theta}$ avec θ. Par exemple, pour la moyenne μ, l'intervalle de confiance vaut :

$$\hat{\mu} - t_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} < \mu < \hat{\mu} + t_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \tag{4.23}$$

où t est la variable de Student au risque α.

En général, du fait de la non connaissance de l'espérance et de la variance de la population mère, il est impossible de spécifier un **modèle statistique** de manière précise. De fait, le modèle statistique contient plusieurs fonctions de répartition de la variable observée, et non une seule. Modéliser des données consiste à spécifier les répartitions possibles de la variable observée. Une fois le modèle choisi, l'intérêt se tourne vers l'estimation des paramètres inconnus du modèle. Intuitivement, la notion d'estimation est claire. On observe les réalisations d'une variable aléatoire dont on connaît la distribution à l'exclusion de quelques paramètres. À l'aide des réalisations observées, on doit estimer les valeurs des paramètres inconnus.

4.3.1 Définition des notions d'estimateur ponctuel et d'estimation

La théorie de l'estimation fait intervenir des fonctions ou statistiques particulières, appelées estimateurs. L'**estimateur** concerne la variable aléatoire. Un estimateur est une fonction des données. Il est construit de telle façon que sa valeur soit proche de la vraie valeur du paramètre. Le but de la théorie de l'estimation est de choisir, parmi toutes les statistiques possibles, le meilleur estimateur,

c'est-à-dire celui qui donnera une **estimation ponctuelle** la plus proche possible du paramètre, et ceci quel que soit l'échantillon. Soit θ un paramètre, et soient $X_1, \ldots, X_n$, n variables aléatoires indépendantes suivant la loi modèle. On appelle **estimateur de** θ une variable aléatoire Y_n fonction des $X_1, \ldots, X_n$:

$$Y_n = y\left(X_1, \ldots, X_n\right) \tag{4.24}$$

Si on a observé expérimentalement les valeurs $x_1, \ldots, x_n$ l'estimateur Y_n fournira une estimation y_n de θ donnée par :

$$y_n = y\left(x_1, \ldots, x_n\right) \tag{4.25}$$

Bien entendu, la définition précédente ne garantit pas du tout que Y_n donne une estimation correcte de θ. On définit alors un **estimateur convergent** comme un « bon » estimateur de θ.

Le processus s'appelle l'**estimation**. Un aspect fondamental de l'inférence statistique consiste à obtenir des **estimations fiables** des caractéristiques d'une population à partir d'un échantillon extrait de cette population. C'est un problème de décision concernant des paramètres tels que :

— l'espérance mathématique notée μ (pour un caractère mesurable) ;

— la variance notée σ^2 ou l'écart-type noté σ ;

— la proportion ϖ (pour un caractère dénombrable).

Comme un échantillon ne peut donner qu'une information partielle sur la population, les estimations ainsi obtenues seront inévitablement entachées d'**erreurs** que l'on doit minimiser autant que possible.

Estimer un paramètre, c'est donner une valeur approchée de ce paramètre, à partir des résultats obtenus sur un échantillon aléatoire extrait de la population.

Exemple. L'estimation de la moyenne L'estimateur d'un échantillon de taille n est la variable aléatoire $\bar{X}$

La moyenne par $\hat{\mu}$ est **une** estimation de μ.

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.26}$$

L'estimateur $\bar{X}$ vaut :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{4.27}$$

L'estimateur $\bar{X}$ subit forcément des fluctuations d'échantillonnage, $\mathbb{E}\left(\bar{X}\right)$ et $\mathbb{V}\left(\bar{X}\right)$ sont-ils sans biais, sont-ils convergents ?

4.3.2 Notion de biais pour un estimateur ponctuel

Un **biais** correspond à la différence entre l'espérance de l'estimateur $\hat{\theta}$ et la valeur à estimer θ dans la population ; on l'appelle également **erreur d'estimation**. Il est dit **sans biais** si :

$$\mathbb{E}\left(\hat{\theta}\right) - \theta = 0 \tag{4.28}$$

Dans le cas contraire, on dira que l'estimateur est biaisé ; on parlera alors d'**erreur systématique**, car l'estimateur $\hat{\theta}$ varie autour de son espérance mathématique $\mathbb{E}\left(\hat{\theta}\right)$, et non autour de la valeur θ du paramètre. Par exemple, la moyenne est un estimateur sans biais :

$$\mathbb{E}\left(\hat{\mu}\right) - \mu = 0 \tag{4.29}$$

Un estimateur est dit **asymptotiquement sans biais** lorsque la limite de sa valeur lorsque n tend vers l'infini vaut :

$$\lim_{n \to +\infty} \mathbb{E}\left(\hat{\mu}\right) - \mu = 0 \tag{4.30}$$

Plus généralement, un **estimateur sans biais** $\hat{\theta}$ aura son espérance qui tendra vers θ lorsque n tend vers l'infini.

Un estimateur sans biais est tel que, sous l'influence du hasard, il donnera une fois des valeurs estimées trop grandes, une autre fois des valeurs estimées trop petites, mais, en moyenne, ces erreurs se balancent, et l'estimateur donne la valeur juste.

Un **estimateur de** θ est dit **efficace** s'il est sans biais et s'il est de variance minimale parmi tous les estimateurs sans biais de θ.

Remarque 1. Un estimateur biaisé donne des estimations qui peuvent s'écarter systématiquement de la valeur à estimer ; il est de fait moins satisfaisant qu'un estimateur sans biais.

Remarque 2. L'absence de biais n'est pas la garantie absolue d'un « bon estimateur ». Il faut aussi tenir compte de sa variance.

4.3.3 Précision d'un estimateur ponctuel

Le carré du biais d'un estimateur $\mathrm{CME}(\theta) = \left(\hat{\theta} - \theta\right)^2$ est appelé le **carré moyen de l'erreur** (C.M.E.). Le carré du biais d'un estimateur permet de calculer l'**erreur quadratique moyenne** (Er.Q.M.) qui correspond à sa **précision**.

$$\mathrm{ERQM}(\theta) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathbb{V}(\theta) + \left(\hat{\theta} - \theta\right)^2 \tag{4.31}$$

Le carré moyen de l'erreur $\mathrm{CME}(\theta)$ d'un estimateur $\hat{\theta}$ est égal à la somme de la variance de l'estimateur et du carré du biais de l'estimateur.

Pour rendre l'erreur quadratique moyenne la plus petite possible, il faut remplir deux conditions :

1. $\mathbb{E}\left(\hat{\theta}\right) = \theta$, c'est-à-dire choisir un estimateur sans biais ;
2. $\mathbb{V}\left(\hat{\theta}\right)$ doit être petite.

Parmi les estimateurs sans biais, on choisira ainsi celui qui a la variance la plus petite. Cette propriété traduit l'**efficacité** de l'estimateur. Dit autrement, si l'espérance mathématique μ est connue, on choisira comme estimateur de la variance σ^2, la statistique $\hat{\theta}$.

La variance étant un indicateur de dispersion, on préfère un estimateur sans biais et convergent, dont la variance est aussi faible que possible. De fait, on choisit l'estimateur minimisant l'E.R.Q.M. Pour deux estimations quelconques, avec ou sans biais, l'**efficacité de l'estimateur $\hat{\theta}_1$ comparée à celle de l'estimateur** $\hat{\theta}_2$, notée W, correspond au rapport de l'E.R.Q.M. de $\hat{\theta}_2$ et de l'E.R.Q.M. de $\hat{\theta}_2$.

$$W = \frac{\mathrm{ERQM}\left(\hat{\theta}_2\right)}{\mathrm{ERQM}\left(\hat{\theta}_1\right)} \tag{4.32}$$

4.3.4 Définition d'un estimateur consistant

Un **estimateur consistant** est un estimateur dont la distribution se concentre dans une zone de plus en plus étroite autour de son objectif au fur et à mesure que la taille de l'échantillon tend vers l'infini. Il existe deux conditions pour démontrer la consistance :

1. l'erreur quadratique moyenne de l'estimateur tend vers zéro lorsque n tend vers l'infini ;
2. le biais et la variance de l'estimateur tendent l'un et l'autre vers zéro.

4.3.5 Définition d'un estimateur convergent

$\hat{\theta}$ est un estimateur convergent s'il converge en probabilité vers θ, c'est-à-dire si :

$$\forall \varepsilon > 0, \Pr\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \xrightarrow{n \to +\infty} 0 \tag{4.33}$$

La notation se lit $\hat{\theta}$ converge en probabilité vers 0.
Dans la pratique, l'inégalité de Bienaymé-Tchebychev [4] fait qu'il suffit de vérifier que l'espérance mathématique $\mathbb{E}(X_n)$ tende vers θ et que $\mathbb{V}(X_n)$ tende vers 0.

Irénée-Jules Bienaymé (1796-1878)

$$\forall \varepsilon > 0, \lim_{n \to +\infty} \Pr\left(\left|\hat{\theta} - \mathbb{E}(X)\right| < \varepsilon\right) < \frac{\mathbb{V}(X)}{n\varepsilon^2} \tag{4.36}$$

Pafnouti Lvovitch Tchebychev (1821-1894)

Appliquer à un estimateur, cela devient :

$$\hat{\theta} \xrightarrow[n \to +\infty]{\Pr} \theta \tag{4.37}$$

Dit autrement, un estimateur $\hat{\theta}$ est convergent si sa distribution tend à se concentrer autour de la valeur inconnue du paramètre θ lorsque la taille n de l'échantillon tend vers l'infini.

Remarque importante. Pour un paramètre donné, on peut trouver différents estimateurs convergents. Toutefois, ils ne disposent pas des mêmes vitesses de convergence.

4.3.6 Propriétés des estimateurs

Il s'agit de vérifier le biais et la convergence de la moyenne et de la variance.

4. L'inégalité de Tchebychev donne un moyen d'évaluer la distance entre les valeurs prises par une variable aléatoire X et son espérance. Plus précisément, elle donne une majoration de la probabilité que l'écart soit grand.

Lemme. Soit X une variable aléatoire positive définie sur un espace probabilisé. On note $\mathbb{E}(X)$ l'espérance de X, alors :

$$\forall k > 0, \Pr(X \geq k) \leq \frac{\mathbb{E}(X)}{k} \tag{4.34}$$

Théorème de l'inégalité de Tchebychev. Soit X une variable aléatoire définie sur un espace probabilisé et qui admet une espérance $\mathbb{E}(X)$ et une variance $\mathbb{V}(X)$, alors :

$$\forall k > 0, \Pr(|X - \mathbb{E}(X)| \geq k) \leq \frac{\mathbb{V}(X)}{k^2} \tag{4.35}$$

La moyenne

L'estimateur est-il sans biais ? Par définition, l'espérance de l'estimateur est :

$$\mathbb{E}\left(\bar{X}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \tag{4.38}$$

En appliquant une des propriétés de l'espérance, on peut écrire :

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \tag{4.39}$$

L'espérance d'une somme de variables aléatoires est la somme de leurs espérances :

$$\frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i\right) \tag{4.40}$$

Les espérances $\mathbb{E}\left(X_i\right)$ étant indépendantes et identiquement distribuées de même loi que X, on peut écrire :

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X\right) = \frac{1}{n}n\mathbb{E}\left(X\right) = \mathbb{E}\left(X\right) = \mu \tag{4.41}$$

L'espérance de l'estimateur est par conséquent sans biais.

L'estimateur est-il convergent ? Par définition, la variance de l'estimateur $\bar{X}$ est :

$$\mathbb{V}\left(\bar{X}\right) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \tag{4.42}$$

En appliquant une des propriété de la variance, on peut écrire :

$$\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) \tag{4.43}$$

La variance d'une somme de variables aléatoires indépendantes est la somme de leur variance :

$$\frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left(X_i\right) \tag{4.44}$$

X_i est indépendante et identiquement distribuée de même loi que X :

$$\frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X) = \frac{1}{n^2}n\mathbb{V}(X) = \frac{1}{n}\mathbb{V}(X) = \frac{\sigma^2}{n} \tag{4.45}$$

En passant à la valeur limite, on peut conclure que l'estimateur est convergent puisque :

$$\lim_{n\to+\infty}\frac{\sigma^2}{n} = 0 \tag{4.46}$$

Si la variance vaut $\mathbb{V}(\mathbb{E}(X)) = \frac{\sigma^2}{n}$, alors l'écart type $\sigma(\mathbb{E}(X))$ (ou l'erreur quadratique à la moyenne) vaut :

$$\sigma(\mathbb{E}(X)) = \frac{\sigma}{\sqrt{n}} \tag{4.47}$$

Cette caractéristique de l'écart entre $\bar{X}$ et sa cible, la vraie valeur de paramètre, μ représente l'**erreur d'estimation** (ou standard erreur (S.E.)). L'écart type de l'échantillon (ou écart type de $\bar{X}$). Cette formule montre explicitement comment l'écart type de $\bar{X}$ diminue lorsque la taille de l'échantillon aléatoire augmente. Cela signifie que SE diminue avec l'accroissement de la taille de l'échantillon n. Cela précise l'idée simple selon laquelle plus l'échantillon est grand, plus $\bar{X}$ donne une estimation exacte de la moyenne de la population. Par ailleurs, SE sert dans l'établissement d'un intervalle de confiance.

La moyenne est un estimateur sans biais et convergent. Cela signifie que la probabilité qu'elle s'écarte de sa cible est quasiment nulle.

La variance

L'estimateur est-il sans biais ? Par définition, l'espérance de l'estimateur est :

$$\begin{array}{l}
\mathbb{E}(\hat{s}^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)\\
\mathbb{E}(\hat{s}^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left({X_i}^2 - \bar{X}^2 - 2X_i\bar{X}\right)\right)\\
\mathbb{E}(\hat{s}^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left({X_i}^2\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\bar{X}^2\right) - \frac{2}{n}\sum_{i=1}^{n}\left(X_i\bar{X}\right)\right)\\
\mathbb{E}(\hat{s}^2) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left({X_i}^2\right) + \frac{1}{n}n\left(\bar{X}^2 - 2\bar{X}\bar{X}\right)\right)\\
\mathbb{E}(\hat{s}^2) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}\left({X_i}^2\right)\right) + \mathbb{E}\left(\bar{X}^2\right)\\
\mathbb{E}(\hat{s}^2) = \frac{1}{n}n\mathbb{E}(X^2) + \mathbb{E}\left(\bar{X}^2\right)\\
\mathbb{E}(\hat{s}^2) = \mathbb{E}(X^2) + \mathbb{E}\left(\bar{X}^2\right)
\end{array} \tag{4.48}$$

En appliquant le théorème de König-Huygens, on peut écrire :

$$\mathbb{E}\left(\hat{s}^2\right) = \frac{1}{n}\left[n\left(\mu^2 + \sigma^2\right)\right] - \left(\mu^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \tag{4.49}$$

La variance empirique est par conséquent un estimateur biaisé. Le facteur $\frac{n}{n-1}$ permet de corriger le biais dans la formule de la **variance empirique corrigée du biais** :

$$\hat{s^*}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(X_i - \bar{X}\right)^2\right] \tag{4.50}$$

$$\Rightarrow \sigma^2 = \frac{n-1}{n}\hat{s}^2 = 1 + \frac{\hat{s}^2}{n} = 1 + \left(\frac{\hat{s}}{\sqrt{n}}\right)^2 \tag{4.51}$$

Le biais de $\hat{s^*}^2$ est nul :

$$\mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right) - \sigma^2 = 0 \tag{4.52}$$

La proportion p

Dans le cadre d'un tirage avec remise, La proportion p correspond généralement à une fréquence, mais ce peut être juste un pourcentage. Son estimateur $\hat{p}$ est sans biais et convergent.

$$\mathbb{E}\left(\hat{p}\right) = p \tag{4.53}$$

et

$$\mathbb{V}\left(\hat{p}\right) = \sqrt{\frac{p\left(1-p\right)}{n}} \tag{4.54}$$

La notion d'estimateur absolument correct

Un estimateur sans biais convergent est dit **absolument correct**, mais il n'est pas **nécessairement unique**.

4.3.7 Statistique exhaustive, information et estimateur

Soit X une variable aléatoire dont la fonction de répartition $F\left(x, \theta\right)$ et la densité $f\left(x, \theta\right)$ dépendant d'un paramètre θ.

Soit $D_\theta \subset \mathbb{R}$ l'ensemble de définition de θ.

Soit un échantillon de taille n de la variable aléatoire $X : (X_1, \ldots, X_n)$.

Soit une statistique T qui est une fonction mesurable T des variables aléatoires X_i notée $T(X_1, \ldots, X_n)$.

Définition d'une statistique exhaustive

Afin de faire de l'inférence statistique, le statisticien va devoir extraire de l'information de la suite de variables aléatoires $X_1, \ldots, X_n$ dont il dispose. Lorsque la taille de l'échantillon n est grande, il est naturel de tenter de réduire l'échantillon et de résumer l'information qui y est contenue. Lorsque cela est possible, $X_1, \ldots, X_n$ sera remplacé par une statistique $T(X_1, \ldots, X_n)$. Ainsi, une statistique T, dépendant d'un échantillon de la variable aléatoire X de taille n, apporte des informations sur un paramètre θ, si sa loi de probabilité dépend de ce paramètre. Une telle statistique contient l'ensemble de l'information sur le ou les paramètres de la loi de probabilité. Toutefois, comme savoir si la réduction des données opérée par la statistique T ne conduit pas à une perte d'information ? Cette interrogation est la question clé que cherche à résoudre la notion d'**exhaustivité**.

Une **statistique** t est dite **exhaustive** [5] pour le paramètre θ si et seulement si la probabilité conditionnelle d'observer X sachant $T(X) = t$ est indépendante de θ. De fait, cet échantillon ne peut plus donner d'informations sur θ. Dit autrement, la statistique t apporte ainsi toute l'information possible sur le paramètre.

Soit l'ensemble de définition E_θ tel que :

$$E_\theta = \{\forall x \in \mathbb{R}, \forall \theta \in D_\theta\} \tag{4.55}$$

Si E_θ ne dépend pas de θ, on le note E.

Comme les variables aléatoires X_i sont indépendantes, la densité de l'échantillon vaut :

$$\forall x \in \mathbb{R}^n, \forall \theta \in D_\theta, \mathcal{L}(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta) \tag{4.56}$$

où $x = (x_1, \ldots, x_n)$ est une réalisation de l'échantillon.

La densité $\mathcal{L}(x, \theta)$ est une fonction de θ appelée **vraisemblance de l'échantillon**. Elle peut se mettre sous la forme :

$$\mathcal{L}(x, \theta) = g(t, \theta)\, h(x, \theta \setminus T = t) \tag{4.57}$$

$g(t, \theta)$ est la densité de la statistique T. $h(x, \theta \setminus T = t)$ est la densité conditionnelle de l'échantillon sachant $T = t$.

5. ou résumé exhaustif

La statistique T est exhaustive si la densité conditionnelle de l'échantillon sachant T ne dépend pas de θ, c'est-à-dire si

$$\mathcal{L}(x,\theta) = g(t,\theta)\, h(x,\theta) \tag{4.58}$$

Lorsque la valeur t de la statistique est connue, l'échantillon n'apporte plus aucune information sur le paramètre θ.

Propriété. Soient T une statistique exhaustive pour le paramètre θ, et ψ une fonction strictement monotone de T, alors la statistique $S = \psi(T)$ est une statistique exhaustive pour le paramètre θ.

Théorème de Darmois. Soit une variable aléatoire X dont le domaine de définition ne dépend pas de θ. Le théorème de Darmois fournit les conditions d'existence d'une statistique exhaustive. S'il existe un entier $n > 1$ tel que l'échantillon admette une statistique exhaustive pour le paramètre θ, la fonction $f(x,\theta)$ est de la forme :

Georges Darmois (1888-1960)

$$\forall x \in E, \forall \theta \in D_\theta, f(x,\theta) = e^{a(x)\alpha(\theta)+b(x)+\beta(\theta)} \tag{4.59}$$

ou la forme équivalente :

$$\forall x \in E, \forall \theta \in D_\theta, \ln[f(x,\theta)] = a(x)\,\alpha(\theta) + b(x) + \beta(\theta) \tag{4.60}$$

Si la densité f est de la forme exponentielle précédente et si l'application $x_j \to t = \sum_{i=1}^{n} a(x_i)$ est bijective et continûment différentiable pour tout x_j, alors la statistique $T = \sum_{i=1}^{n} a(x_i)$ est une statistique exhaustive particulière pour le paramètre θ.

Remarque 1. Si l'ensemble E dépend de θ, la première partie du théorème de Darmois est vraie, mais la seconde est fausse.

Remarque 2. Il peut exister une statistique exhaustive que l'on trouve par d'autres méthodes.

Information de Fisher

Sous réserve de l'existence de l'intégrale, la **quantité d'information apportée par un échantillon** est :

$$I_n(\theta) = I_{x_1,x_2,\ldots,x_n}(\theta) = \mathbb{E}\left[\left(\frac{\mathrm{d}\ln L(x,\theta)}{\mathrm{d}\theta}\right)\right] = \int_{E_\theta}\left(\frac{\mathrm{d}\ln L(x,\theta)}{\mathrm{d}\theta}\right)^2 L(x,\theta)\,\mathrm{d}x \tag{4.61}$$

Si l'ensemble E_θ ne dépend pas de θ et si la vraisemblance $L(x,\theta)$ est dérivable au moins jusqu'à l'ordre deux, la quantité d'information de Fisher possède les propriétés suivantes :

Ronald Aylmer Fisher (1890-1962)

— $\frac{\mathrm{d}\ln L(x,\theta)}{\mathrm{d}\theta}$ est une variable aléatoire centrée ;

— $I_n(\theta) = \mathbb{V}\left(\frac{\mathrm{d}\ln L(x,\theta)}{\mathrm{d}\theta}\right)$;

— $I_n(\theta) = -\mathbb{E}\left(\frac{\mathrm{d}^2\ln L(x,\theta)}{(\mathrm{d}\theta)^2}\right)$;

— $I_n(\theta) = nI_1(\theta)$.

Exemple. Pour la loi normale, $I_n(\theta) = \frac{n}{\sigma^2} = nI_1(\theta)$

Au niveau de la **dégradation de l'information**, si l'ensemble E_θ ne dépend pas du paramètre θ, l'information apportée par un échantillon est supérieure ou égale à l'information apportée par une statistique. Il est à noter que l'égalité n'existe que si la statistique est exhaustive.

4.3.8 Recherche du meilleur estimateur

La recherche du meilleur estimateur d'un paramètre θ est un problème difficile à résoudre.

1. La précision d'un estimateur $\hat{\theta}$ dépend de sa variance, c'est-à-dire que la loi de $\hat{\theta}$ dépend elle-même de la loi de la variable aléatoire X. Il faut de fait connaître la forme de cette loi.
2. Une statistique est un résumé apporté par un échantillon, il est par conséquent très important de ne pas perdre l'information.

En tenant compte de ces deux impératifs, on peut aborder la recherche du meilleur estimateur suivant deux méthodes :

1. soit en recherchant des statistiques exhaustives qui conduisent à des estimateurs sans biais de variance minimale ;
2. soit en étudiant la quantité d'information de Fisher qui apporte des indications sur la précision d'un estimateur.

Estimateur sans biais de variance minimale

Les propriétés d'un estimateur sans biais de variance minimale peuvent se résumer par quatre termes :

1. l'**unicité**. S'il existe un estimateur sans biais de variance minimale, il est unique presque sûrement ;
2. le **théorème de Rao-Blackwell**. Soient $\hat{\theta}$ un estimateur sans biais du paramètre θ, et U une statistique exhaustive pour ce paramètre, alors $\hat{\theta}^* = \mathbb{E}\left(\hat{\theta} \setminus U\right)$ est un estimateur sans biais de θ au moins aussi bon que $\hat{\theta}$;

Calyampudi Radhakrishna Rao (1920-2023)

David Harold Blackwell (1919-2010)

3. la **statistique exhaustive** et l'**estimateur de variance minimale**. S'il existe une statistique exhaustive U, alors l'estimateur sans biais de variance minimale du paramètre θ ne dépend que de la statistique U. Il peut exister plusieurs estimateurs sans biais, fonction d'une statistique exhaustive U. Pour obtenir l'unicité, il faut introduire la notion de statistique complète. Une statistique U est dite **complète** pour une famille de lois $f(x, \theta)$ si :

$$\forall \theta, \mathbb{E}(h(U)) = 0 \tag{4.62}$$

entraîne $h = 0$ presque sûrement, h étant une fonction réelle ;

Erich Leo Lehmann (1917-2009)

Henry Scheffé (1907-1977)

4. le **théorème de Lehmann-Scheffé**. Soit $\hat{\theta}^*$ un estimateur sans biais du paramètre θ dépendant d'une statistique exhaustive complète U. $\hat{\theta}^*$ est l'unique estimateur sans biais de variance minimale. En particulier, si on connaît un estimateur $\hat{\theta}$ sans biais, $\hat{\theta}^*$ est donné par $\hat{\theta}^* = \mathbb{E}(\theta \setminus U)$. Le meilleur estimateur d'un paramètre est un estimateur sans biais dépendant d'une statistique exhaustive, par exemple, la famille des lois exponentielles.

On peut ainsi calculer la dégradation de l'information. On note :

— $I_n(\theta)$ l'information apportée par l'échantillon ;

— $I_T(\theta)$ l'information apportée par la statistique choisie ;

— $I_{n\setminus T}(\theta)$ l'information conditionnelle apportée par l'échantillon sachant la statistique T.

L'ensemble de définition E_θ ne dépend pas de θ.

Les quantités d'information vérifient les propriétés suivantes :

— $I_n(\theta) = I_T(\theta) + I_{n\setminus T}(\theta)$;

— $I_n(\theta) \geq I_T(\theta)$.

Si l'ensemble E_θ ne dépend pas de θ, l'information apportée par un échantillon est supérieure ou égale à l'information apportée par une statistique. L'égalité a lieu si et seulement si la statistique est exhaustive. Dit autrement, toute l'information apportée par un échantillon, concernant un paramètre est contenue dans une statistique exhaustive.

Précision intrinsèque d'un estimateur et inégalité de Cramer-Rao

Pour obtenir la précision intrinsèque d'un estimateur, les hypothèses suivantes doivent être vérifiées :

1. la densité $f(x, \theta)$ est telle que la quantité d'information de Fisher $I_n(\theta)$ existe et est finie, et que, en particulier, $\mathbb{E}\left(\hat{\theta}\right)$ et $\mathbb{V}\left(\hat{\theta}\right)$ existent ;
2. l'ensemble E_θ ne dépend pas de θ ;

3. les dérivées par rapport à θ de $L(x, \theta)$ existent et sont intégrables dans $\mathbb{R}^n$, l'inégalité de Cramér-Rao apparaît :

$$\mathbb{V}\left(\hat{\theta}\right) = \frac{1}{I_n(\theta)} \left[\frac{\mathrm{d}\mathbb{E}\left(\hat{\theta}\right)}{\mathrm{d}\theta}\right]^2 \tag{4.63}$$

Basée sur l'information de Fisher, la borne Cramér-Rao exprime une borne inférieure sur la variance d'un estimateur sans biais. Harald Cramér (1893-1985)

Il existe deux particuliers intéressants à retenir :

1. pour un estimateur sans biais $\hat{\theta}$ d'une fonction $h(\theta)$, c'est-à-dire $\mathbb{E}\left(\hat{\theta}\right) = h(\theta)$, l'inégalité de Cramér-Rao s'écrit :

$$\mathbb{V}\left(\hat{\theta}\right) \geq \frac{\left[h'(\theta)\right]^2}{I_n(\theta)} \tag{4.64}$$

2. pour un estimateur sans biais $\hat{\theta}$ du paramètre θ, c'est-à-dire pour $\mathbb{E}\left(\hat{\theta}\right) = \theta$, l'inégalité de Cramér-Rao s'écrit :

$$\mathbb{V}\left(\hat{\theta}\right) \geq \frac{1}{I_n(\theta)} \tag{4.65}$$

La variance d'un estimateur sans biais est minorée par une quantité indépendante de cet estimateur. Elle ne peut pas être inférieure à une certaine borne.

4.3.9 Estimateurs robustes

Un estimateur est **robuste** s'il est peu sensible aux données aberrantes. La notion de **point de rupture** permet de mesurer la **robustesse** d'un estimateur.

Les valeurs aberrantes

L'analyse statistique de données tient compte d'erreurs aléatoires, mais il faut être conscient du fait que pratiquement tout jeu de données contient également des erreurs d'un autre genre.

Des mesures sont parfois complètement fausses, des chiffres dans une base de données sont transmis ou copiés avec des fautes, une expérience est effectuée avec les fausses valeurs des facteurs, *etc*. Même des données de bonne qualité contiennent souvent 10 % à 20 % de telles **valeurs aberrantes**. Une bonne vieille

pratique dans l'analyse des données consiste à ne pas tenir compte des observations qui ne suivent pas la structure de la majorité. Cette pratique n'est pas à recommander, même si l'idée que l'estimation de paramètres doit résister à la présence de telles valeurs est valable et importante.

Par exemple, on prend la liste ordonnée $\{3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 60\}$. On peut dresser un tableau avec deux colonnes, l'une sans 60, l'ature avec 60, et comparer les paramètres mesurés (Tab. 4.4). S'il n'existe aucune différence notable entre les paramètres mesurés sans 60 et avec 60, le paramètre est dit robuste.

	Sans 60	**Avec 60**	**Remarque**
Échantillon n	10	11	
Somme des valeurs	50	110	
Moyenne m	5	10	Statistique non robuste
Somme de l'écart à la moyenne	20	2770	
Variance s^2	2	251,8	Statistique non robuste
Écart type s	1,41	15,87	Statistique non robuste
Médiane m_e	5	5	Statistique robuste
Quartile 1 Q_1	4	4	Statistique robuste
Quartile 3 Q_3	6	7	Statistique robuste
$Q_3 - Q_1$	2	3	Statistique robuste
Étendue w	4	57	Statistique non robuste

TABLE 4.4 – Extraction d'une valeur aberrante

Il est assez clair que la présence de valeurs aberrantes peut complètement invalider les résultats d'une méthode statistique. Pour cette raison, il est important de disposer d'outils qui aident à identifier ces observations. Une observation aberrante est souvent liée à un résidu exceptionnel – positif ou négatif. De fait, **la recherche des valeurs aberrantes dans une base de données s'effectue au moyen des résidus**.

Face à des valeurs aberrantes, plusieurs réactions sont possibles.

1. On peut essayer d'identifier ces valeurs et ne pas en tenir compte lors de l'estimation des paramètres.
2. Une meilleure alternative sont les estimateurs robustes qui bornent l'influence des valeurs aberrantes, et pour lesquels il n'est pas nécessaire de connaître par avance les valeurs aberrantes.

Remarque. La médiane et les quartiles ne permettent pas d'effectuer des calculs comme la moyenne ou l'écart type.

L'estimateur robuste et les moyennes tronquées

Un **estimateur** est dit **robuste** s'il est peu influencé par des valeurs aberrantes. L'idée fondamentale de la construction de tels estimateurs est la **troncature**, notion qui est à la base des moyennes tronquées.

Soit $y_{1\backslash n}, y_{2\backslash n}, \ldots, y_{n\backslash n}$ un échantillon trié. Pour calculer une moyenne tronquée, on enlève quelques unes des observations les plus petites et les plus grandes. Par la suite, la moyenne des observations restantes est calculée. Le pourcentage α des observations tronquées de chaque côté caractérise cet estimateur que l'on nommera θ_α. De fait, on obtient :

$$\begin{array}{ll} \alpha = 0 & \frac{\left(y_{1\backslash n},\ldots,y_{n\backslash n}\right)}{n} = \theta_0 \\ \alpha = \frac{1}{n} & \frac{\left(y_{2\backslash n},\ldots,y_{n-1\backslash n}\right)}{n-2} = \theta_{1\backslash n} \\ \alpha = \frac{2}{n} & \frac{\left(y_{3\backslash n},\ldots,y_{n-2\backslash n}\right)}{n-4} = \theta_{2\backslash n} \\ \ldots & \\ \alpha = 50\ \% & \theta_{50\ \%} \end{array} \tag{4.66}$$

Une valeur $\alpha \approx 25\ \%$ représente un bon compromis entre la perte d'efficacité et le gain de robustesse. Malheureusement, l'estimation de l'écart type des estimateurs θ_α est assez compliquée.

L'analyse par médianes d'un plan d'expériences à deux voies

La sensibilité aux valeurs aberrantes est une faiblesse générale de la méthode des moindres carrés. En pratique, il est rare que les expériences effectuées lors d'une série ne produisent aucune valeur aberrante.

Des méthodes robustes, plus résistantes à des valeurs aberrantes, existent. Ce paragraphe présente la méthode du **« *median polish* »**.

Estimation des effets par moindres carrés dans un plan à deux facteurs L'estimation par moindres carrés des effets d'un plan d'expériences croisées, c'est-à-dire d'un tableau de mesures sous l'influence de deux facteurs f_1 et f_2, peut s'effectuer par l'algorithme A suivant :

1. calculer les moyennes dans les lignes du tableau ;
2. soustraire de chaque ligne sa moyenne ;
3. calculer les moyennes dans les colonnes, y compris la colonne des moyennes calculées sous (1) ;
4. soustraire de chaque colonne sa moyenne.

On obtient ainsi un tableau de résidus, une colonne d'effets de f_1, une ligne d'effets de f_2, et la moyenne globale.

Analyse d'un tableau à deux voies par médianes La médiane est un estimateur plus résistant aux influences des valeurs aberrantes que la moyenne. Si un jeu de données $y_1, \ldots, y_n$ contient une grosse erreur – on dit une valeur exotique, la moyenne est grandement influencée par cette observation, tandis que la médiane ne l'est pas. Une analyse robuste d'un tableau peut être obtenue en remplaçant les moyennes par les médianes. On procède à l'algorithme B :

1. calculer les médianes dans les lignes du tableau ;
2. soustraire de chaque ligne sa médiane ;
3. calculer les médianes dans les colonnes, y compris la colonne des médianes calculées sous (1) ;
4. soustraire de chaque colonne sa médiane ;
5. recommencer à l'étape (1) et continuer jusqu'à la convergence.

M-estimateurs

Les M-estimateurs forment une grande famille d'estimateurs robustes. Ces techniques s'adaptent facilement à différents problèmes statistiques, et sont ainsi plus flexibles que les estimateurs tronqués.

On peut avoir les M-estimateurs comme estimateurs de moindres carrés avec une pondération artificielle qui associe un poids réduit à des valeurs aberrantes. Parce que la grandeur des résidus est un indicateur du degré d'aberrance d'une observation, on peut baser une telle pondération sur les résidus. Le poids est proche de 1 lorsque le résidu r est proche de zéro. Le poids diminue lorsque $|r|$ augmente. Une fonction de ce genre est le **poids de Huber** qui vaut 1 pour $|r|$ suffisamment petit et qui décroît inversement proportionnellement à $|r|$ sinon. On peut également utiliser le **bicarré**.

Michel Huber (1875-1947)

Un M-estimateur de régression est défini par une fonction de poids $w\,(u, k)$, symétrique en u. La constante $k > 0$ règle le degré de robustesse. On calcule l'estimateur de la manière suivante :

1. calculer un estimateur préliminaire des paramètres du modèle de régression, par exemple l'estimateur des moindres carrés ;
2. calculer les résidus préliminaires $r_1, \ldots, r_n$, ainsi que la **distance interquartiles** (D.I.Q.) de ces résidus ;
3. calculer le poids de la pondération robuste :

$$w_i = w\left(\frac{r_i}{\mathrm{DIQ}}, k\right) \text{ avec } i = \{1, \ldots, n\} \tag{4.67}$$

4. calculer l'estimateur des moindres carrés pondérés en utilisant les poids $w_1, \ldots, w_n$. Ainsi, un nouvel estimateur θ_k des paramètres de modèle de

régression (qui remplace l'estimateur préliminaire) est trouvé. Cela permet également de recalculer les résidus ;

5. estimer la matrice de covariance de l'estimateur θ_k par :

$$\hat{\mathbb{V}}\left\{\hat{\theta}\right\} = \hat{\tau}\left(X^T X\right)^{-1} \tag{4.68}$$

où

$$\hat{\tau}^2 = \frac{n^2 \sum_{i=1}^{n} {w_i}^2 {r_i}^2}{(n-q)\left(\sum_{i=1}^{n}\left(w_i\left(\frac{r_i}{\mathrm{DIQ}\left(w'\left(\frac{r_i}{\mathrm{DIQ}},k\right)\right)}\right)\right)\right)^2} \tag{4.69}$$

Dans cette formule, X est la matrice du modèle de régression, q le nombre de coefficients et $w'(u,k)$ la dérivée de $w(u,k)$ par rapport u.

Remarque. Après l'étape (4), on peut reprendre avec (2), c'est-à-dire recalculer la distance interquartile des nouveaux résidus, recalculer la pondération et finalement recalculer l'estimateur.

Estimateur de Huber Deux estimateurs correspondent à l'**estimateur de Huber** par la fonction :

$$w_{\text{Huber}}(u,k) = \begin{cases} \frac{k}{|u|} & \text{si } |u| \geq k \\ 1 & \text{si } -k < u < k \end{cases} \tag{4.70}$$

Pour cet estimateur, la somme dans le dénominateur de $\hat{\tau}^2$ (sous (5)) est égale au nombre d'observations ayant un poids w_i égal à un. Un bon choix pour la constante k se trouve entre 0,7 et 0,9.

Estimateur bicarré L'estimateur bicarré se base sur :

$$w_{\text{bicarré}}(u) = \begin{cases} \frac{k^2-u^2}{k^4} & \text{si } -k < u < k \\ 0 & \text{sinon} \end{cases} \tag{4.71}$$

Dans ce cas, la somme dans le dénominateur de $\hat{\tau}^2$ vaut :

$$\hat{\tau}^2 = \sum_{i=1}^{n} \sqrt{w_t}\left(1 - 5\left(\frac{r_i}{(k \times \mathrm{DIQ})}\right)^2\right) \tag{4.72}$$

Les valeurs de k entre 3 et 4 constituent un bon choix pour cet estimateur.

4.3.10 Estimation d'un paramètre par intervalle de confiance

L'estimation ponctuelle n'est pas l'objectif de la statistique, car on ne souhaite généralement pas seulement connaître une valeur estimée, mais aussi l'erreur probable de l'estimation. On désire savoir si $\hat{\theta}$ est éloigné ou proche de θ. Pour ce, on construit un **intervalle de confiance**, lequel doit avoir de « grandes chances » de contenir la vraie valeur du paramètre. Il est toujours associé à un risque d'erreur α.

Principe d'un intervalle de confiance

À quelle valeur de la moyenne μ de la population mère peut-on s'attendre dans un échantillon de taille n ?

Soit ϖ la fréquence d'apparition d'un caractère A dans une population.

Soit f la fréquence d'apparition du même caractère dans un échantillon de taille n.

On sait que f est une estimation ponctuelle non biaisée de ϖ. **Quelle confiance peut-on accorder à cette estimation ?** En effet, on ne peut pas affirmer que ϖ soit exactement égale à f observée. Dit autrement, on doit construire un intervalle d'estimation appelée **intervalle de confiance** de la forme :

$$\varpi = f \pm \text{marge d'échantillonnage} \tag{4.73}$$

La marge d'échantillonnage est souvent appelée **erreur d'échantillonnage**.

Pour définir cet intervalle de confiance, on choisit un nombre $\alpha \in]0, 1[$ et on détermine un intervalle $]a, b[$ tel que l'on ait la probabilité α de se tromper en affirmant que ϖ appartienne à cet intervalle.

On dira que l'on a obtenu l'intervalle de confiance ϖ au **coefficient de risque** α, ou au **coefficient de sécurité** $(1 - \alpha)$ (ou seuil de confiance ou quasi-certitude).

Pour construire un intervalle de confiance, il suffit d'introduire une variable aléatoire dont on connaît la distribution de probabilité.

Pour l'estimation d'une proportion, d'une moyenne ou d'une variance, on suppose implicitement que la population est un effectif infini. Certains résultats ne sont plus valables si l'on suppose que la population possède un effectif fini.

Estimation d'une moyenne dans le cas où l'écart type est connu – Distribution normale

La forme de la distribution d'échantillonnage de $\bar{X}$ change lorsqu'on augmente la taille de l'échantillon. Lorsque la population mère est normale, ou lorsque la taille de l'échantillon est grande, la distribution d'échantillonnage de $\bar{X}$ a une forme approximativement normale.

Les échantillons considérés sont issus d'une population suivant la loi normale $N(\mu, \sigma)$. Les propriétés démontrées ne sont valables que sous cette hypothèse.

Dans le cas d'un tirage sans remise, la distribution des valeurs dans une loi normale permet un ajustement tendanciel à partir de la probabilité que la population mère soit dans l'intervalle des écarts types. Le **risque d'erreur** (ou le seuil de confiance) α de généralisation est calculable. On déduit la formule générale de l'intervalle de confiance encadrant, moyenne d'échantillon, intervalle ayant $100 \times (1 - \alpha)$ % de chance de contenir la moyenne de la moyenne de la population mère. $(1 - \alpha)$ est appelé le **niveau de confiance**. Pour ce, on utilise l'erreur quadratique à la moyenne SE.

$$\Pr\left(\bar{X} - z_C \times b \times \mathrm{SE} \leq \mu \leq \bar{X} + z_C \times b \times \mathrm{SE}\right) = 1 - \alpha \tag{4.74}$$

avec α le risque d'erreur choisi par l'utilisateur. Son choix définit et fixe la valeur de z_C. $\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$ avec σ est l'écart type des valeurs de l'échantillon et n est la taille de l'échantillon. b est lié au taux de sondage $\frac{n}{N}$. $b = 1$ si le tirage s'effectue avec remise, tandis que $b = \sqrt{1 - \frac{n}{N}}$.

N.B. La connaissance de l'écart type σ de la population mère est déterminante pour construire un intervalle de confiance autour de la moyenne μ de la population mère.

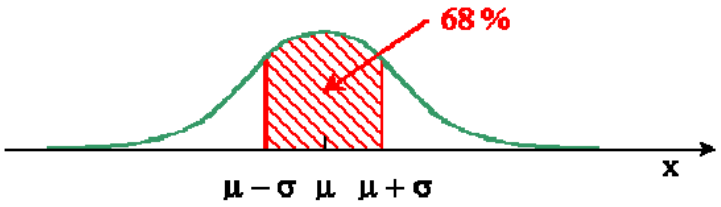
z_C est la probabilité critique (Tab. **??** ; Fig. 4.3).

Seuil de confiance	$\mathbf{z_C}$
$50,00$ %	$0,6745$
$68,27$ %	$1,00$
$80,00$ %	$1,28$
$90,00$ %	$1,645$
$95,00$ %	$1,96$
$95,45$ %	$2,00$
$96,00$ %	$2,05$
$98,00$ %	$2,33$
$99,00$ %	$2,58$
$99,73$ %	$3,00$

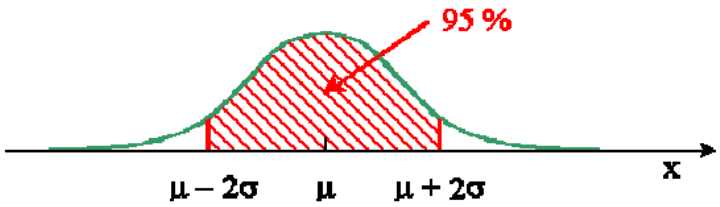
TABLE 4.5 – Probabilités critiques de la loi normale

Remarque. L'accroissement de la présence d'un estimateur correspond au gain résultant de l'augmentation de la taille n de l'échantillon.

- La moyenne (ici appelée μ) est égale au mode et à la médiane.
- On retrouve 68.26% de la population entre ± 1 écart-type (ici appelé s) autour de la moyenne



- On retrouve 95.44% de la population entre ± 2 écarts types autour de la moyenne.



- On retrouve 99.74% de la population entre ± 3 écarts types autour de la moyenne

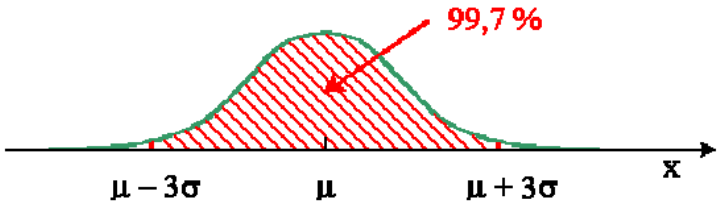


FIGURE 4.3 – Illustration des probabilités critiques de la loi normale

On pose $\varepsilon = z_C \frac{\sigma}{\sqrt{N}} = z_C \mathrm{SE}$. De cette relation, on peut déduire que $z_C = \varepsilon \frac{\sqrt{N}}{\sigma}$, et $N = \left(z_C \frac{\sigma}{\varepsilon}\right)^2$

Cela étant, l'écart type σ de la population mère est rarement connu. La loi normale ne pouvant être utilisée, est remplacée par la loi t de Student à $n-1$ degrés de liberté, puisque σ est remplacé par l'écart type estimé s.

Exemple. En économie, on choisit souvent un intervalle de confiance de 95 %. En d'autres termes, on utilisera une technique qui donnera, lorsqu'on tire un grand nombre d'échantillons, un intervalle correct 19 fois sur 20. Si on choisit d'estimer une moyenne μ par une distribution normale centrée et réduite, l'étendue la plus faible qui contient exactement une probabilité de 95,00 % est manifestement la partie centrale qui exclut une probabilité de 2,50 % à chaque queue de la distribution. On regarde à quoi correspond 2,50 % dans les tables statistiques de la loi normale ce qui correspond à $z_C = 1,96$ dans notre cas. Cela signifie que l'on doit considérer des valeurs de part et d'autre de la moyenne égales à 1,96 fois l'écart type SE de l'échantillon.

$$\Pr\left(\bar{X} - 1,96 \times \mathrm{SE} \leq \mu \leq \bar{X} + 1,96 \times \mathrm{SE}\right) = 0,95 \tag{4.75}$$

Cela signifie qu'il existe 95 % de chance que la variable aléatoire μ tombe entre $\mu - 1,96 \times \mathrm{SE}$ et $\mu + 1,96 \times \mathrm{SE}$.

Remarque importante. La manière de construire un tel intervalle est générique à toutes les lois statistiques. La loi normale sert d'exemple, car elle correspond à la loi la plus courante.

Dans le cas des échantillons de petite taille ($n < 30$) ou d'un tirage avec remise, SE vaut :

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{4.76}$$

avec $\sqrt{\frac{N-n}{N-1}}$ le **facteur d'exhaustivité**.

La variable $\bar{X}$ est une combinaison linéaire de variables aléatoires indépendantes suivant une loi normale, donc la variable $\bar{X}$ est également une variable suivant une loi normale $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Estimation d'une moyenne dans le cas où l'écart type est inconnu – Distribution normale

Cas d'un petit échantillon Pour un petit échantillon, la variable $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ suit une loi de Student à $\upsilon = n - 1$ degrés de liberté.

Pour un risque donné α, on lit le t_α dans la table et on peut affirmer au risque α que :

$$\bar{X} - t_\alpha \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_\alpha \frac{s}{\sqrt{n}} \tag{4.77}$$

Cas d'un grand échantillon Si $n \geq 30$, la variable aléatoire $U = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ suit approximativement la loi normale centrée réduite. L'intervalle de confiance de μ au risque α s'écrit :

$$\bar{X} - u_\alpha \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + u_\alpha \frac{s}{\sqrt{n}} \tag{4.78}$$

Cas d'un échantillon sur une population infinie Pour un échantillonnage sur une population infinie de taille N : $\bar{X} \pm z_C \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

Estimation de la variance dans le cas d'une moyenne connue – Distribution normale

La variance s'écrit $s^2 = \mathbb{V}\left(\bar{X}\right) = \frac{1}{n}\left[\sum_{i=1}^{n}\left(X_i - \mathbb{E}\left(X\right)\right)^2\right] - \left(\mathbb{E}\left(X\right) - \bar{X}\right)^2$. Il est possible de réarranger la décomposition en :

$$ns^2 = \left[\sum_{i=1}^{n}\left(X_i - \mathbb{E}\left(X\right)\right)^2\right] - n\left(\mathbb{E}\left(X\right) - \bar{X}\right)^2 \tag{4.79}$$

William Cochran (1909-1980)

Le théorème de Cochran concernant la décomposition d'une forme quadratique permet d'écrire :

$$\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \frac{ns}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}\right)^2 \tag{4.80}$$

Dit autrement, une loi du χ^2 est une somme de lois normales centrées réduites à n degrés de liberté.

ns^2 est la somme des n carrés de variables aléatoires centrées réduites indépendantes et suivant une loi normale est une variable $\chi^2\left(n\right)$.

$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$ est une somme de formes quadratiques. La première est de rang $(n-1)$, car les variables vérifient la relation $\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$. La deuxième est de rang 1.

L'intervalle de confiance bilatéral pour un risque α symétrique de la variable du χ_2 vaut :

$$\Pr\left(\chi^2{}_{\frac{\alpha}{2}}\left(n\right) < \chi^2\left(n\right) < \chi^2{}_{1-\frac{\alpha}{2}}\left(n\right)\right) = 1 - \alpha \tag{4.81}$$

On en déduit un intervalle de confiance bilatéral pour un risque α à risques symétriques (s est la valeur de la statistique S donnée par l'échantillon) :

$$\Pr\left(\frac{ns^2}{\chi^2{}_{1-\frac{\alpha}{2}}\left(n\right)} < \sigma^2 < \frac{ns^2}{\chi^2{}_{\frac{\alpha}{2}}\left(n\right)}\right) = 1 - \alpha \tag{4.82}$$

avec une loi du χ^2 à $n-1$ degrés de libertés.

Estimation de la variance dans le cas d'une moyenne inconnue – Distribution normale

Si la moyenne est inconnue, $\frac{ns^2}{\sigma^2}$ est une variable $\chi^2\left(n-1\right)$. De plus, $\bar{X}$ et s^2 sont deux variables indépendantes. La réciproque est vraie. Si $\bar{X}$ et s^2 sont indépendantes, la variable aléatoire X est une variable aléatoire suivant une loi normale.

$$\Pr\left(\chi^2{}_{\frac{\alpha}{2}(n-1)} < \chi^2\,(n-1) < \chi^2{}_{\frac{1-\alpha}{2}(n-1)}\right) = 1-\alpha \tag{4.83}$$

On en déduit un intervalle de confiance bilatéral pour un risque α à risques symétriques (s est la valeur de la statistique S donnée par l'échantillon) :

$$\Pr\left(\frac{ns^2}{\chi^2{}_{1-\frac{\alpha}{2}}\,(n-1)} < \sigma^2 < \frac{ns^2}{\chi^2{}_{\frac{\alpha}{2}}\,(n-1)}\right) = 1-\alpha \tag{4.84}$$

Cas d'un petit échantillon Pour un risque α donné, on détermine les nombres a et b tels que :

$$\begin{cases} \Pr\left(X \leq a\right) = \frac{\alpha}{2} \\ \Pr\left(X \geq b\right) = \frac{\alpha}{2} \end{cases} \tag{4.85}$$

Les nombres a et b se lisent dans la table du χ^2. La ligne correspond au degré de liberté, soit ici $n-1$. La colonne correspond à la surface à droite. Pour $\alpha = 0,05$, on lit b dans la colonne 0,025 et a dans la colonne 0,975, car la surface est égale à 1.

La seule valeur connue de s^2 étant s, on obtient comme intervalle de confiance de σ^2 au risque α :

$$\frac{(n-1)\,s^2}{b} \leq \sigma^2 \leq \frac{(n-1)\,s^2}{a} \tag{4.86}$$

Cas d'un grand échantillon Le théorème précédent est vrai quel que soit n, mais les tables χ^2 s'arrêtent généralement au degré de liberté $\upsilon = 30$. On ne peut pas les utiliser si $n > 31$. En l'absence d'ordinateur, on dispose du théorème d'approximation suivant.

Théorème d'approximation pour les grands échantillons. Si X est une variable aléatoire suivant une loi du χ^2 à υ degrés de liberté. Si $\upsilon > 30$, alors la variable aléatoire $U = \sqrt{2X} - \sqrt{2\upsilon - 1}$ suit à peu près la loi réduite $N\,(0,1)$.

Cas d'un échantillon sur une population infinie Les limites de confiance de l'écart type σ d'une population distribuée normalement, σ étant estimé par l'écart type s d'un échantillon, ont pour expression :

$$s \pm z_C \frac{\sigma}{\sqrt{2N}} \tag{4.87}$$

Conséquence des estimations de la moyenne et de la variance – Distribution normale

Comme

$$\begin{cases} \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} & \text{est une variable suivant la loi } N(0,1) \\ \frac{ns^2}{\sigma^2} & \text{est une variable } \chi^2(n) \end{cases} \tag{4.88}$$

on en déduit que $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n-1}}}$ est une variable suivant la loi de Student $T(n-1)$. Comme la variable aléatoire de Student ainsi définie ne dépend pas de σ, cette propriété sera utilisée dans la théorie de l'estimation lorsque l'écart type σ est inconnu.

Estimation d'une proportion – Distribution poissonnienne

Exemple. On réceptionne un lot d'articles dans lequel on prélève un échantillon de 100 articles. On compte le nombre $k = 3$ d'articles défectueux. Quelle est la proportion ϖ d'articles défectueux dans le lot ? 3 %.

Quelle est celle de la population ? Pour y répondre, il faut introduire un intervalle de confiance. La figure n° 4.4 permet de lire l'intervalle de confiance pour une proportion. Pour $varpi = 3$ %, l'intervalle de confiance de ϖ est compris entre 1 % et 8,6 %.

On suppose maintenant qu'il y a 6 % d'articles défectueux dans le loit. Comme la proportion d'articles défectueux est assez faible, et comme la taille de l'échantillon est assez grande, il est possible d'utiliser une approximation de Poisson avec $N = 100$ et $\lambda = N\varpi = 6$. Le problème posé est par conséquent celui de connaître la probabilité d'avoir un article défectueux d'une loi de Poisson de paramètre $\lambda = 6$? Pour un risque de 2,5 %, $k \approx 1$ sur 100 d'observer un article défectueux, tandis que, pour un risque de 97,5 %, $k \approx 11$ sur 100 d'observer un article défectueux.

Quelle est la probabilité d'avoir un article défectueux d'une loi de Poisson de paramètre $\lambda = 10$? Pour un risque de 2,5 %, $k \approx 4$ sur 100 d'observer un article défectueux, tandis que, pour un risque de 97,5 %, $k \approx 16$ sur 100 d'observer un article défectueux.

Estimation d'une proportion – Distribution normale

Les échantillons considérés sont issus d'une population suivant la loi normale $N(\mu, \sigma)$. Les propriétés démontrées ne sont valables que sous cette hypothèse.

Dans le cadre d'un tirage avec remise, l'erreur type de l'échantillon d'une fréquence vaut :

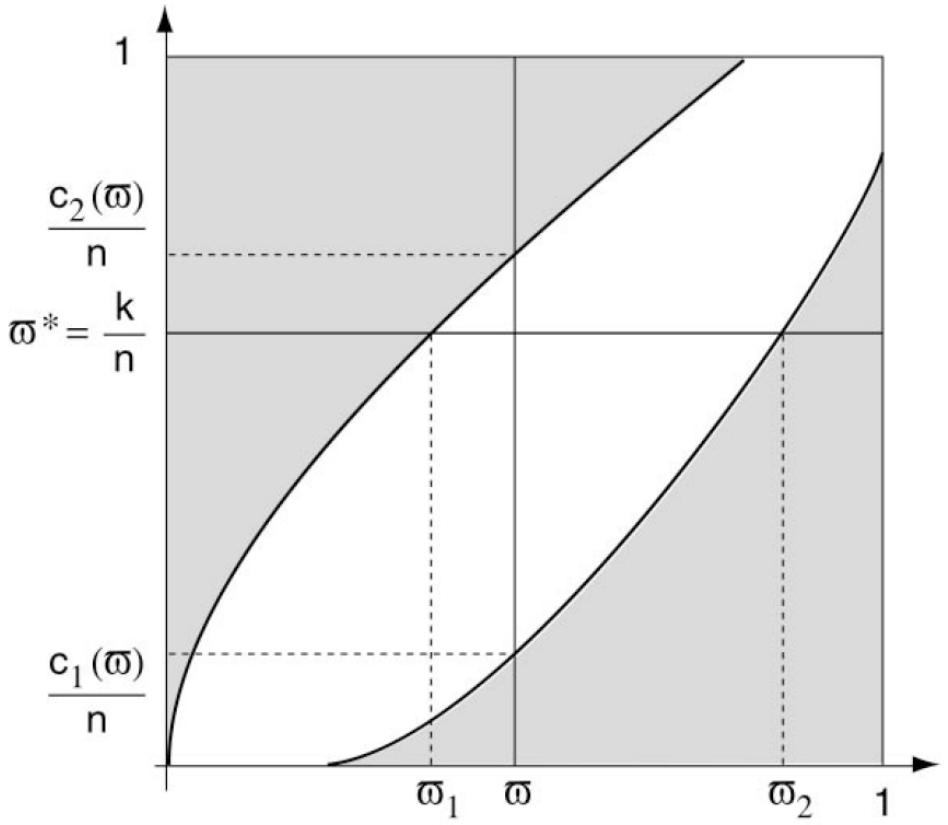


FIGURE 4.4 – Estimation d'une proportion
En abscisse : proportion ϖ d'articles défectueux contenus dans un lot ; en ordonné : nombre d'articles défectueux dans l'échantillon
En blanc : H_0 ; en gris : H_1

$$\mathrm{ET}\left(\varpi\right) = \sqrt{f \times \frac{1-f}{n}} \tag{4.89}$$

avec f la fréquence observée sur l'échantillon.

$$\Pr\left(f - z_C \times b \times \mathrm{ET} \leq \varpi \leq f + z_C \times b \times \mathrm{ET}\right) = 1 - \alpha \tag{4.90}$$

avec α le risque d'erreur choisi par l'utilisateur. Son choix définit et fixe la valeur de z_C. $\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$ avec σ est l'écart type des valeurs de l'échantillon et n est la taille de l'échantillon. b est lié au taux de sondage $\frac{n}{N}$. $b = 1$ si le tirage s'effectue avec remise, tandis que $b = \sqrt{1 - \frac{n}{N}}$.

Dans le cas des échantillons de petites tailles ($n < 30$) ou d'un tirage sans remise, ET vaut :

$$\mathrm{ET}\,(\varpi) = \sqrt{f \times \frac{1-f}{n}}\sqrt{\frac{N-n}{N-1}} \tag{4.91}$$

avec $\sqrt{\frac{N-n}{N-1}}$ le **facteur d'exhaustivité**.

Propriétés des intervalles de confiance

Propriété 1. Un intervalle de confiance est un **intervalle aléatoire**, car les bornes de cet intervalle correspondent à des variables aléatoires, elles-mêmes fonctions des observations.

Propriété 2. Le seuil α étant donné, il faut définir les nombres α_1 et α_2. Leur choix dépend des problèmes à traiter, des risques encourus à négliger les petites ou les grandes valeurs du paramètre. Si on choisit $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, on construit un **intervalle de confiance bilatéral à risques symétriques**. On peut construire des **intervalles de confiance unilatéraux**, soit avec $\alpha_1 = 0$ et $\alpha_2 = 0$.

Propriété 3. Le seuil α, les nombres α_1 et α_2 et la taille n de l'échantillon étant fixés, on peut construire un intervalle de confiance associé à chaque échantillon. Toutefois, parmi ces intervalles, une proportion égale à α % ne contiendra pas la valeur exacte du paramètre. Ce seuil α représente le risque que l'intervalle de confiance ne contienne pas la vraie valeur du paramètre. La situation la plus favorable correspond à choisir un risque α petit, associé à un intervalle de faible étendue.

Propriété 4. On peut diminuer la valeur du seuil α, et même à la limite, choisir $\alpha = 0$ pour avoir la certitude absolue. Dans ce cas, l'intervalle de confiance s'étend à tout le domaine de définition du paramètre $]-\infty, +\infty[$ pour l'espérance mathématique ou $[0, +\infty[$ pour l'écart type, par exemple. Dit autrement, diminuer la valeur de α revient à augmenter l'étendue de l'intervalle.

Propriété 5. Dans la pratique, on donne à α une valeur acceptable de l'ordre de 5 %, puis, lorsque cela est possible, on augmente la taille de l'échantillon.

Propriété 6. La probabilité $(1 - \alpha)$ représente le **niveau de confiance** de cet intervalle qui est associé à l'intervalle, et non à la valeur inconnue du paramètre.

Propriété 7. Pour définir un intervalle de confiance, il faut connaître un estimateur ponctuel du paramètre, ainsi que sa loi de distribution.

Exemple pour comprendre les intervalles de confiance

Un intervalle de confiance est une estimation. On opère un échantillon de taille n. On observe une fréquence f. Comment savoir si cette fréquence est proche de

la proportion théorique p inconnue que l'on observerait pour l'ensemble de la population ? Quel est son intervalle de confiance I_C ?

$$I_C = \left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}} \right] \tag{4.92}$$

I_C ne dépend que de la taille de l'échantillon n. Il correspond à 95 % des cas, c'est-à-dire :

$$\Pr\left(f - \frac{1}{\sqrt{n}} \leq p \leq f + \frac{1}{\sqrt{n}} \right) \geq 0,95 \tag{4.93}$$

Exemple pour comprendre. Un institut de sondage interroge 1 052 personnes sur leur intention de vote. 614 déclarent avoir l'intention de voter pour le candidat A. 438 déclarent avoir l'intention de voter pour le candidat B. Qui a le plus de chance d'être élu ?

1. On calcule les fréquences.
   - Pour A, $f(A) = \frac{614}{1052} = \frac{307}{526}$
   - Pour B, $f(B) = \frac{438}{1052} = \frac{219}{526}$
2. On calcule l'intervalle de confiance I_C de chaque fréquence.
   - Pour A, $I_C = \left[f(A) - \frac{1}{\sqrt{1052}}, f(A) + \frac{1}{\sqrt{1052}} \right] = \left[\frac{307}{526} - \frac{1}{\sqrt{1052}}, \frac{307}{526} + \frac{1}{\sqrt{1052}} \right] \approx [0,543; 0,624]$
   - Pour B, $I_C = \left[f(B) - \frac{1}{\sqrt{1052}}, f(B) + \frac{1}{\sqrt{1052}} \right] = \left[\frac{438}{526} - \frac{1}{\sqrt{1052}}, \frac{438}{526} + \frac{1}{\sqrt{1052}} \right] \approx [0,369; 0,464]$
3. Le candidat A a la plus grande probabilité d'être élu.

Pour $n = 1052$, $n_A = 570$ et $n_B = 482$, $f(A) = \frac{285}{526}$ et $f(B) = \frac{241}{526}$. On en déduit que $I_C(A) \approx [0,500; 0,584]$ et $I_C(B) \approx [0,413; 0,504]$. Le candidat B peut atteindre 0,504 dans 95 % des cas. Il n'est pas possible de déterminer qui gagnera, mais le candidat A dispose de plus de chances que le candidat B.

Pour $n = 1052$, $n_A = 550$ et $n_B = 502$, $f(A) = \frac{275}{526}$ et $f(B) = \frac{251}{526}$. On en déduit que $I_C(A) \approx [0,480; 0,565]$ et $I_C(B) \approx [0,433; 0,522]$. Il est impossible de déterminer qui gagnera. Si on augmente l'échantillonnage en multipliant les résultats du sondage par 100. L'opération permet de conserver les mêmes fréquences. On obtient $n = 105200$, $n_A = 55000$ et $n_B = 50200$, $f(A) = \frac{275}{526}$ et $f(B) = \frac{251}{526}$. On en déduit que $I_C(A) \approx [0,519; 0,527]$ et $I_C(B) \approx [0,473; 0,482]$. Dans ce cas, le candidat A a davantage de chances de gagner.

Comparaison de deux moyennes issues de deux échantillons distincts – Distribution normale

Différence et somme de deux moyennes dans le cas d'échantillons indépendants suivant des lois normales de variances inconnues Soient deux échantillons indépendants entre eux issus de la même population mère et représentés par deux variables aléatoires, X_1 et X_2, suivant la même loi de probabilité.

La moyenne de la différence ou de la somme des moyennes $\bar{X}_1$ et $\bar{X}_2$ vaut :

$$\bar{X}_{\bar{X}_1 \pm \bar{X}_2} = \bar{X}_1 \pm \bar{X}_2 \tag{4.94}$$

L'écart type de la différence ou de la somme des moyennes s_1 et s_2 vaut :

$$s_{\bar{X}_1 \pm \bar{X}_2} = \sqrt{{\mathrm{SE}_1}^2 \pm {\mathrm{SE}_2}^2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} \tag{4.95}$$

L'erreur quadratique sur la différence et la somme des moyennes SE de ces deux variables est :

$$\mathrm{SE} = \frac{\bar{X}_1 \pm \bar{X}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \tag{4.96}$$

où s_1 est l'écart type de l'échantillon 1, n_1 est l'effectif de l'échantillon 1, s_2 est l'écart type de l'échantillon 2, et n_2 est l'effectif de l'échantillon 2.

Soient deux échantillons issus de la même population mère représentés par deux variables aléatoires indépendantes, X_1 et X_2, suivant la même loi de probabilité. On suppose que les variables $(X_1 + X_2)$ et $(X_1 - X_2)$ sont indépendantes, alors les variables aléatoires X_1 et X_2 sont des variables aléatoires suivant une loi normale. La variable aléatoire

$$\frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}_1\right)^2 \sum_{i=2}^{n} \left(X_i - \bar{X}_2\right)^2}} \times \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.97}$$

suit une loi de Student à $(n_1 + n_2 - 2)$ degrés de liberté. Les sommations sont effectuées sur les échantillons 1 et 2. Ce résultat permet de construire un intervalle de confiance pour la différence des moyennes de deux échantillons indépendants suivant une loi normale.

Ainsi, afin de déterminer un intervalle de confiance pour la différence des moyennes de deux lois normales, ou pour tester l'égalité de ces deux moyennes, il suffit d'utiliser la table statistique de Student.

Dans le cas limite d'une population infinie, l'intervalle de confiance entre deux moyennes vaut :

$$\left(\bar{X}_1 \pm \bar{X}_2\right) \pm z_C \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}} \tag{4.98}$$

avec $\bar{X}_1, s_1, n_1, \bar{X}_2, s_2, n_2$ respectivement les moyennes, les écarts types et les tailles de deux échantillns extraits de la population mère. z_C est la probabilité associée au risque d'erreur.

Différence et somme de deux moyennes dans le cas d'échantillons indépendants suivant des lois normales pour des échantillons appariés L'intervalle de confiance à 95 % pour des échantillons appariés X_1 et X_2 correspond à :

$$\bar{X}_1 \pm \bar{X}_2 \pm t_{0,025} \times \frac{s_{\bar{X}_1 \pm \bar{X}_2}}{\sqrt{n}} \tag{4.99}$$

avec $t_{0,025}$ la variable de Student et $s_{\bar{X}_1 \pm \bar{X}_2}$ l'écart type de la différence ou de la somme.

Les échantillons appariés disposent de deux avantages.

1. L'appariement est un mariage qui maintient constantes beaucoup de variables exogènes.
2. L'appariement est manifestement une caractéristique qu'il est souhaitable d'introduire chaque fois que cela est possible.

   Remarque. Si l'on ne peut pas recourir à cette modalité par impossibilité d'utiliser deux fois le même individu, on doit rechercher d'autres moyens.

Comparaison de deux variances issues de deux échantillons – Distribution normale

Soient n_1 réalisations indépendantes d'une variable aléatoire X_1 suivant la loi normale $N(\mu_1, \sigma_1)$ et n_2 réalisations indépendantes d'une variable aléatoire X_2 suivant la loi normale $N(\mu_2, \sigma_2)$. Les variables X_1 et X_2 sont indépendantes.

De la propriété

$$\begin{cases} \frac{n_1 {s_1}^2}{{\sigma_1}^2} = \chi^2(n_1 - 1) \\ \frac{n_2 {s_2}^2}{{\sigma_2}^2} = \chi^2(n_2 - 1) \end{cases} \tag{4.100}$$

on en déduit le résultat suivant qui sera utilisé dans la théorie de l'estimation :

$$\frac{n_1 {s_1}^2}{{\sigma_1}^2 (n_1 - 1)} = \frac{{\sigma_2}^2 (n_2 - 1)}{n_2 {s_2}^2} = F(n_1 - 1, n_2 - 1) \tag{4.101}$$

avec F est la valeur de Fisher-Snedecor.

Ce résultat permet de déterminer un intervalle de confiance pour le rapport de deux variances ou pour tester l'égalité de deux variances dans le cas d'une population mère connue. On peut utiliser les méthodes connues. On obtient par exemple :

$$\Pr\left(\frac{{s_1^*}^2}{{s_2^*}^2}\frac{1}{F_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)} \leq \frac{{\sigma_1}^2}{{\sigma_2}^2}\frac{{s_1^*}^2}{{s_2^*}^2} \leq \frac{{s_1^*}^2}{{s_2^*}^2}F_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)\right) = 1-\alpha \tag{4.102}$$

avec s^* désigne la valeur de la statistique S^* donnée par l'échantillon considéré. Les valeurs de $F_{\frac{1-\alpha}{2}}\left(n_1-1,n_2-1\right)$ et de $F_{\frac{\alpha}{2}}\left(n_1-1,n_2-1\right)$ sont lues sur les tables. On en déduit l'égalité, ou non, des variances.

Comparaison de deux proportions issues de deux échantillons – Distribution normale

Si les populations sont infinies, les limites de confiance de la différence ou la somme de deux fréquences estimées valent :

$$f_1 \pm f_2 \pm z_C\sqrt{\frac{f_1\left(1-f_1\right)}{n_1}+\frac{f_2\left(1-f_2\right)}{n_2}} \tag{4.103}$$

avec f_1 et f_2 les deux fréquences des échantillons de taille n_1 et n_2.

Généralisation

On généralise toutes ces formules en les appliquant aux distributions de probabilité :

— la loi normale ;
— la loi de Bernoulli ;
— la loi exponentielle ;
— la loi uniforme.

Pour ce, on utilise des estimateurs permettront d'opérer des **tests statistiques** :

— le test de Student ;
— le test du χ^2
— *etc*.

Cela dépend des hypothèses statistiques et du paramètre à estimer (moyenne, variance, *etc*.). On obtient un **intervalle de fluctuation** si on connaît la proportion p.

Les tests permettent de définir la probabilité que p théorique soit dans un intervalle de confiance supérieur à 0,99.

4.3.11 Estimation d'un paramètre par intervalle de pari

À partir des paramètres d'une population mère X, peut-on estimer un **intervalle de pari** estimant l'encadrement des paramètres de la population de l'échantillon $\bar{X}$ de taille n ?

Par exemple, on suppose que la moyenne m de l'échantillon et la moyenne μ de la population mère soient égales. On peut alors calculer l'intervalle de pari de la moyenne de l'échantillon. On note σ l'écart type de la population mère.

$$\mu - z_C \frac{\sigma}{\sqrt{n}} < m < \mu + z_C \frac{\sigma}{\sqrt{n}} \tag{4.104}$$

avec $z_C \sim N(0,1)$, $X \sim N(\mu, \sigma)$ et $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

L'intervalle de pari est plus précis que l'intervalle de confiance.

4.3.12 Estimation d'un paramètre par la méthode du *bootstrap*

Pour expliquer la méthode du *bootstrap*, on propose de calculer un intervalle de confiance pour une médiane ?

Soit un échantillon $(x_1, x_2, \ldots, x_n)$ ayant une médiane m_e. On considère un échantillon de l'échantillon qui a la taille n de l'échantillon. On tire au sort avec remise dans l'échantillon de départ n fois l'une des valeurs. On répète le processus p fois tel que p soit très grand (autour de dix millions de tirages). Chaque p-échantillon possède sa propre médiane m_{e_p}. En appliquant la loi de la moyenne, on peut proposer une valeur médiane moyenne ${m_e}^* = \frac{1}{p}\sum_{i=1}^{p} m_{e_i}$, une variance estimée $s^2 = \frac{1}{p-1}\sum_{i=1}^{p}\left(m_{e_i} - {m_e}^*\right)^2$, donc une erreur quadratique $\mathrm{SE} = \sqrt{\frac{s^2}{p}}$. Il est alors possible d'appliquer un intervalle de confiance :

$${m_e}^* - t_{\frac{\alpha}{2}} \times \mathrm{SE} \leq m_e \leq {m_e}^* + t_{\frac{\alpha}{2}} \times \mathrm{SE} \tag{4.105}$$

En général, la méthode du *bootstrap* consiste à rééchantillonner plusieurs fois l'échantillon de départ dans le but d'obtenir un intervalle de confiance pour des paramètres calculables (moyenne, écart type, *etc*.) ou non calculables (médiane, quartiles, *etc*.). C'est dans ce dernier cas que la méthode est particulièrement intéressante. Le *bootstrap* est un exemple particulier utilisant les méthodes de Monte-Carlo.

4.4 Les méthodes d'estimation d'un paramètre

Toutes les définitions sur les estimateurs font partie de l'estimation ponctuelle d'un paramètre. L'objectif de cette partie est proposée d'autres méthodes d'esti-

mation d'un paramètre.

4.4.1 Estimation d'un paramètre par la méthode des moindres carrés

La puissance de cette estimation est maximale lorsque plusieurs variables aléatoires sont étudiées. De fait, ce paragraphe n'est que l'introduction d'un chapitre ultérieur beaucoup plus complet.

La méthode des moindres carrés est un principe utilisable lorsque les quantités à estimer sont des espérances.

À la base de la méthode des moindres carrés se trouve un modèle de la forme générale :

$$y_i = \mu_i + \varepsilon_i \tag{4.106}$$

avec $i = \{1, 2, \ldots, n\}$ et les ε_i sont les erreurs qui vérifient :

$$\begin{cases} \mathbb{E}(\varepsilon_i) = 0 \\ \mathbb{V}(\varepsilon_i) = 0 \\ \text{cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ lorsque } i \neq j \end{cases} \tag{4.107}$$

L'**estimation d'un paramètre par la méthode des moindres carrés** se base sur une décomposition analogue :

$$y_i = \hat{y}_i + r_i \tag{4.108}$$

avec $i = \{1, 2, \ldots, n\}$, $\hat{y}_i$ les valeurs ajustées, et les résidus r_i. Les paramètres qui interviennent dans μ_i sont estimés de telle façon que la somme $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ soit minimale et que :

$$\sum_{i=1}^{n} {y_i}^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} {r_i}^2 \tag{4.109}$$

Pour d'autres modèles, il est nécessaire de disposer d'une méthode d'estimation plus générale. C'est la méthode de la vraisemblance.

4.4.2 Estimation d'un paramètre par la méthode du maximum de vraisemblance (M.V.)

La **vraisemblance** offre une approche générale de l'estimation de paramètres inconnus à l'aide de données. La vraisemblance est une fonction des observations et du paramètre dont l'objectif est de trier les différentes valeurs du paramètre selon leur *likelihood* (probabilité en anglais).

La fonction de vraisemblance

Soient $y_1, \ldots, y_n$ des observations et soit $F(t_1, \ldots, t_n \setminus \theta)$ un modèle statistique sous la forme d'une famille de fonctions de répartition conjointes. La fonction de vraisemblance V vaut alors :

$$V(\theta) = f(y_1, \ldots, y_n \setminus \theta) \tag{4.110}$$

la densité conjointe évaluée pour les observations. De fait, cette méthode revient à supposer que l'événement qui s'est produit, était le plus probable.

Si le modèle stipule des observations indépendantes et identiquement distribuées, il suffit de spécifier la loi marginale $F(t \setminus \theta)$ et la **fonction du maximum de vraisemblance** V vaut alors :

$$V(\theta) = f(y_1 \setminus \theta) \times \ldots \times f(y_n \setminus \theta) \tag{4.111}$$

L'estimateur du maximum de vraisemblance

L'**estimateur du maximum de vraisemblance** $\hat{\theta}$ est le point qui maximise la vraisemblance $V(\theta)$. Dit autrement, $V\left(\hat{\theta}\right) \geq V(\theta)$ pour chaque θ.

Dans la pratique, la fonction f étant en général strictement positive, on peut maximiser $\ln f$, ce qui est équivalent à maximiser f. On prend comme estimation du maximum de vraisemblance, la solution de l'**équation de la vraisemblance** :

$$\frac{\mathrm{d} \ln f(y_1, \ldots, y_n \setminus \theta)}{\mathrm{d}\theta} = 0 \tag{4.112}$$

S'il existe une statistique exhaustive T, l'estimateur du maximum de vraisemblance en dépend. Par contre, s'il n'existe pas de statistique exhaustive, il existe une suite θ_n de racines de l'équation de la vraisemblance qui converge presque sûrement vers θ lorsque n tend vers l'infini. Il est à noter que la variable aléatoire $\frac{\theta_n - \theta}{\frac{1}{\sqrt{I_n(\theta)}}}$ converge en loi vers la loi normale centrée réduite $N(0,1)$ lorsque n tend vers l'infini.

La méthode du maximum de vraisemblance consiste à choisir comme estimation de θ la valeur θ_0 qui rend f maximale. Si f est supposée deux fois dérivables, alors la valeur θ_0 vérifie :

$$\begin{cases} \frac{\mathrm{d}f(\theta_0)}{\mathrm{d}\theta} = 0 \\ \frac{\mathrm{d}^2 f(\theta_0)}{(\mathrm{d}\theta)^2} < 0 \end{cases} \tag{4.113}$$

Pour illustrer le propos, on peut proposer le tableau suivant (Tab. 4.6) :

Distribution	Paramètre à estimer	Estimateur
Loi uniforme sur $[0, a]$	a	$\sup x_i$
Loi binomiale k nombre de succès en n épreuves	p	$\hat{p} = \frac{k}{n}$
Loi de Poisson	λ	$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Loi normale	μ	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
	σ	$\hat{\sigma^*}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$

TABLE 4.6 – Estimateurs obtenus par la méthode du maximum de vraisemblance

Les conditions d'usage

En situation expérimentale, cette méthode nécessite l'usage d'un ordinateur.

Pour certains paramètres, l'estimation par le maximum de vraisemblance peut conduire à des résultats différents que l'estimation non biaisée.

La méthode peut se généraliser à l'estimation simultanée de plusieurs paramètres.

4.5 La réduction d'un échantillon

« Réduire » un échantillon consiste à le résumer par une valeur unique T la plus proche possible de l'ensemble des valeurs x_i de l'échantillon, revient à chercher un point $T(t_1, \ldots, t_n)$ dans $\mathbb{R}^n$ le plus proche de l'échantillon.

La démarche s'opère en deux temps :

1. choisir une distance d dans $\mathbb{R}^n$;
2. chercher le point T qui minimise $d(T, X)$. T est la caractéristique de la tendance centrale associée à la distance d ; $d(T, X)$ caractérise la dispersion associée à cette distance.

La distance retenue est calculée de manière générale par la **distance de Minkowski**.

Hermann Minkowski (1864-1909)

4.5.1 La distance de Minkowski

Pour $r \geq 1$, la distance de Minkowski vaut :

$$d(x, y) = \left[\sum_{i=1}^{n} p_i |x_i - y_i|^r\right]^{\frac{1}{r}} \tag{4.114}$$

4.5.2 La distance de Manhattan d_1

$$d_1(T, X) = \sum_{i=1}^{n} p_i |t - x_i| \tag{4.115}$$

La distance d_1 est définie et continue sur $\mathbb{R}$. Elle atteint un minimum en un point au moins qui est, par définition, la médiane, ce qui permet d'en proposer une définition rigoureuse. On appelle **médiane d'un échantillon** m_e toute valeur qui rend minimale la quantité : $d_1(T, X)$.

Avec $T = m_e$, $d_1(T, X) = \mathrm{EC}(X)$, c'est-à-dire l'écart moyen.

La médiane caractérise la valeur centrale, tandis que l'écart moyen caractérise la dispersion associée à la distance d_1.

4.5.3 La distance euclidienne d_2

$$d_1{}^2(T, X) = \sum_{i=1}^{n} p_i (t - x_i)^2 \tag{4.116}$$

La distance $d_1{}^2(T, X)$ est minimale pour $t = \bar{X}$ et $d_1{}^2(T, X) = s^2$. La moyenne caractérise la valeur centrale, tandis que l'écart type caractérise la dispersion associée à la distance euclidienne.

4.5.4 La distance d_∞

$$d_\infty(T, X) = \max(|t - x_i|) \tag{4.117}$$

avec $i = \{1, 2, \ldots, n\}$.

La distance $d_\infty(T, X)$ est minimale avec $T = \mathrm{ME}(X)$, c'est-à-dire la moyenne des valeurs extrêmes de l'échantillon.

$$\mathrm{ME}(X) = \frac{1}{2}[\min(x_1, \ldots, x_n) + \max(x_1, \ldots, x_n)] \tag{4.118}$$

c'est-à-dire la moitié de l'étendue de l'échantillon.

La moyenne des valeurs extrêmes est la caractéristique de la valeur centrale, tandis que la moitié de l'étendue est la dispersion associée à la distance d_∞.

4.6 La théorie de la décision : les tests statistiques

Un test d'hypothèse permet un **jugement sur échantillon**.

Pourquoi avoir besoin de comparer des moyennes, des variances, des proportions et des distributions ? Tout simplement pour établir si un **événement**, un phénomène ou une action a eu un impact sur les observations considérées. Existe-t-il une relation de cause à effet entre un premier échantillon avant la réalisation d'un événement et un second échantillon après la réalisation de l'événement ?

Pour comparer les deux échantillons (sans et avec l'événement réalisé), on utilise des tests d'hypothèse. Il en existe deux grandes catégories : les tests paramétriques (moyenne, écart type, type de distribution) et les tests non paramétriques (effectif, médiane, *etc*.). Parmi les tests paramétriques, le test t de Student permet d'évaluer l'impact d'une variable qualitative sur une variable quantitative. Parmi les tests non paramétriques, le test de Mann-Whitney évalue l'impact d'une variable qualitative sur une variable qualitative. Le test du χ^2 évalue les relations entre deux variables qualitatives. Il faut noter que les tests non paramétriques sont moins puissants que les tests paramétriques, mais ils sont presque toujours possibles.

Avec les comparaisons, la statistique inférentielle introduit la notion de **chemin causal**. Elle permet d'opérer des enquêtes sur les bases de données et de les profiler. Elle permet de prédire, déduire et modéliser, mais surtout de faire parler les données.

Le test est la base de toute science expérimentale. La **théorie des tests** consiste à formuler des hypothèses particulières sur les paramètres ou sur les lois qui interviennent dans les problèmes étudiés, puis apporter un jugement sur ces hypothèses. Ce jugement est basé d'une part, sur les résultats obtenus sur un ou plusieurs échantillons extraits de la population concernée, et, d'autre part, sur l'acceptation d'un risque dans la prise de **décision statistique** qui correspond à un choix opéré sur une population à partir de l'information que donne un échantillon.

Un test statistique (ou test d'hypothèse) est une méthode de calcul permettant de décider si une série statistique d'observations est compatible avec une loi de probabilité entièrement spécifique. Dit autrement, comment savoir si un résultat observé est en accord avec une distribution théorique ? Cela permet de valider ou d'infirmer des prédictions. De fait, on a besoin de tester pour répondre de manière affirmative ou négative à cette question. Il s'agit ainsi d'une démarche conduisant à élaborer une règle de décision permettant de faire un choix entre deux hypothèses statistiques.

Une **hypothèse statistique** est une assertion vraie ou fausse au sujet d'une population que l'on peut mettre à l'épreuve en tirant un échantillon au hasard. C'est un énoncé quantitatif sur les caractéristiques d'une population. Les hypothèses s'appellent l'**hypothèse nulle** H_0 et l'**hypothèse alternative** H_1. Avant

toute démarche statistique, il faut définir à quelle condition l'une ou l'autre des hypothèses sera considérée comme vraisemblable. Les deux hypothèses ne jouent pas le même rôle. En effet, c'est l'hypothèse nulle H_0 qui est soumise au test, et toute démarche statistique consiste à la considérer comme vraie. Si le test conduit à la rejeter, c'est l'hypothèse alternative H_1 qui sera considérée comme vraie. Dès lors, les tests statistiques conduisent à une conclusion dichotomique. On fait une hypothèse *a priori* H_0, qui sera acceptée ou refusée par le test. On peut en proposer quelques exemples.

Comme l'hypothèse nulle H_0, on peut tester :

— une valeur particulière d'un paramètre : $\hat{\theta} = \theta$;

— l'égalité des valeurs d'un paramètre défini sur deux populations différentes ;

— l'ajustement d'une distribution théorique à une distribution expérimentale.

L'hypothèse alternative H_1 peut-être :

— $\hat{\theta} = \theta_1$;

— $\hat{\theta} > \theta_0$ et $\hat{\theta} \neq \theta_0$;

— $\hat{\theta} < \theta_1$.

La notion de « test statistique » est à la fois simple et compliquée. Elle est simple, car il s'agit de se poser une question dichotomique. Elle est compliquée, car il existe plusieurs tests statistiques :

— comparer un échantillon à une référence théorique. Il compare la distribution théorique et la distribution expérimentale. L'hypothèse H_0 consiste à supposer que les différences observées sont suffisamment faibles pour être explicables par les hasards du tirage au sort. Ce **test** est dit **de conformité** ;

— comparer plusieurs échantillons (comparaison des moyennes, comparaison des variances, *etc.*). L'hypothèse H_0 consiste à supposer qu'ils proviennent d'une même population, c'est-à-dire que les différences observées sont explicables par les fluctuations de l'échantillonnage. Il s'agit de comparer entre elles deux ou plusieurs distributions statistiques observées. Ce **test** est dit **d'homogénéité** ;

— comparer les distributions de deux ou plusieurs séries appariées d'un même échantillon (test sur séries appariées) ;

— montrer qu'une distribution étudiée suit vraisemblablement une loi de probabilité donnée. Il s'agit de vérifier si la distribution de l'échantillon est compatible avec celle de la population mère. Ce **test** est dit **d'adéquation à une loi de probabilité** ;

— comparer deux caractères. L'hypothèse H_0 consiste à supposer que les deux caractères quantitatifs ou qualitatifs sont indépendants. Ce **test** est dit **d'indépendance de deux caractères**.

Dans ce cadre, la plus grande difficulté est de reconnaître le ou les tests à utiliser dans le contexte d'une analyse statistique spécifique.

Remarque importante. L'outil de travail d'un test statistique est la **méthode d'estimation par intervalle de confiance**. On définit une variable caractéristique de la loi de probabilité à tester, dont la loi de probabilité est elle-même établie sous l'hypothèse H_0 à tester.

4.6.1 Définitions

La notion de probabilité critique

La probabilité critique (z_C) résume très clairement le degré de concordance entre les données et H_0. Elle permet de mesurer la crédibilité de H_0. En général, tout test d'hypothèses définit la probabilité critique comme la probabilité que la valeur de l'échantillon soit égale à la valeur réellement observée sous l'hypothèse H_0. La probabilité critique est un excellent moyen de résumer ce qu'affirment les données par rapport à la crédibilité de H_0.

Les tests de signification ou les tests d'hypothèse

Les **tests de signification** sont basés sur un score de test intuitif et le calcul de la p-valeur.

Les **tests d'hypothèse** sont caractérisés par des taux d'erreurs.

Toutefois, 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.

Les tests paramétriques ou les tests non paramétriques

Pour les **tests paramétriques**, la forme des distributions testées est supposée connue *a priori* et le test porte sur la valeur d'un ou plusieurs paramètres d'une loi spécifique. Par exemple, pour savoir si oui ou non, une espérance est plus petite qu'une certaine constante. Ces réponses doivent être apportées sur la base des observations. On distingue :

1. les hypothèses simples du type $\theta = \theta_0$;
2. les hypothèses composites du type $\theta \in \delta_\theta$ où δ_θ est un intervalle de $\mathbb{R}$; elles ramènent, en général soit à $\theta > \theta_0$, soit à $\theta < \theta_0$ ou encore à $\theta \neq \theta_0$. Ces tests supposent, en général, l'existence d'une variable aléatoire X suivant une loi normale. Si les résultats obtenus sont valables même si la variable

aléatoire X n'est pas une variable normale, on dit que le **test** est **robuste**, les résultats restent valables après quelques modifications des données.

Pour les **tests non paramétriques**, la forme des distributions n'est pas prise en considération (en tant qu'hypothèse *a priori*). D'ailleurs, leur détermination peut être le but du test. Ils s'appliquent tout aussi bien aux variables quantitatives (test de Mann-Whitney, test de Wilcoxon, test t du coefficient de corrélation de Spearman, test sur les signes en utilisant la loi binomiale, *etc*.) qu'aux variables qualitatives (test du χ^2, test de Fisher exact, *etc*.).

Les tests robustes

Les **tests robustes** se décomposent en plusieurs catégories. Parmi elles, les **tests libres** sont les plus intéressants, car ils sont valables quelle que soit la forme de la loi de la variable aléatoire X. On peut par conséquent les utiliser lorsque l'on ne connaît pas la loi de la variable aléatoire X.

1. Les tests de moyenne de non-corrélation sont des tests robustes.
2. Les tests robustes sont souvent paramétriques.

Les risques et les probabilités d'erreur

L'information étant incomplète, toute décision est associée à un risque. Il en existe deux.

1. α est l'erreur de première espèce : c'est la probabilité de rejeter H_0 alors qu'elle est vraie.
2. Il existe également β, l'erreur de deuxième espèce qui est la probabilité d'accepter H_0 alors qu'elle est fausse. On réalise alors une hypothèse alternative H_1 qui est avèrèe. Dans ce cas, la décision correcte consisterait à rejeter l'hypothèse H_0 qui est fausse. Dit autrement, une erreur serait commise si la valeur mesurée tombée dans la région de H_0. Par exemple, dans un procès pour meurtre, H_0 : l'accusé est innocent ; H_1 : l'accusé est coupable.

Réduire α revient à augmenter β. *A contrario*, réduire β revient à augmenter α. Dit autrement, la seule manière de réduire une erreur sans augmenter l'autre est de rassembler de meilleures preuves. α est appelé le **risque du client**. β est appelé le **risque du fournisseur**. Plusieurs cas sont possibles.

1. La taille n de l'échantillon et le risque α sont fixés. Dans ces conditions, le risque β diminue si la différence entre les deux valeurs proposées, m_0 et m_1, augmente.

État	**Décisions**	
	H_0 acceptable	H_0 rejetée
Si H_0 est vraie	Décision correcte avec le seuil de confiance $p = 1 - \alpha$	Erreur de la première espèce avec le seuil de signification (ou le seuil du test) $p = \alpha$
Si H_0 est fausse	Erreur de la seconde espèce $p = \beta$	Décision correcte avec la puissance du test $p = 1 - \beta$

TABLE 4.7 – Exemple des quatre résultats possibles d'un test d'hypothèses

2. Si le risque α diminue, la zone de non-rejet de l'hypothèse H_0, on finit par la garder trop souvent. De plus, dans ces conditions, le risque β augmente, donc la région de refus de l'hypothèse H_1 augmente. Les deux risques de première et deuxième espèces sont antagonistes.
3. Si l'on fixe le risque α et si la taille n de l'échantillon augmente, la zone de non-rejet de l'hypothèse H_0 devient plus petite, d'où une diminution du risque β ; le test est ainsi plus puissant.

On peut préférer un test d'hypothèse classique au calcul de la probabilité critique si le seuil α du test peut être déterminé rationnellement, et s'il existe beaucoup d'échantillons à classer.

Les règles d'un test statistique

1. Se poser une question.
2. Poser une hypothèse de travail.
3. Formuler de façon précise une hypothèse nulle H_0 et hypothèse alternative H_1.
4. En déduire la loi de distribution d'un paramètre, sous cette hypothèse H_0.
5. Choisir un seuil d'erreur de probabilité α (5 % ou 1 %) avec l'expérience pour accepter H_0.
6. Préciser les conditions d'application du test :
   a. forme de la loi de probabilité de la population étudiée ;
   b. taille de l'échantillon ;
   c. variance connue ou inconnue.
7. Collecter les données d'un ou plusieurs échantillons que l'on suppose tiré au hasard dans la population mère, c'est-à-dire représentatif.

8. Choisir la statistique la mieux adaptée en fonction des caractéristiques de la population étudiée et donner sa loi de probabilité sous les deux hypothèses, ces lois doivent être différentes. La **statistique de test** est toujours $t = \frac{\hat{\theta}-\theta}{\sqrt{\sigma^2}}$ avec $\hat{\theta}$ l'estimation ponctuelle, θ le paramètre exact de la population mère et σ^2 la variance de la population mère.
9. Déterminer la région critique (ou région de rejet) de l'hypothèse nulle H_0 au profit de l'hypothèse alternative H_1 et en déduire la règle de décision :
   a. W est la région critique conduisant au rejet de H_0 : $H_0 : \Pr(W \setminus H_0) = \alpha$. Respectivement dans laquelle, si H_0 est vraie, le paramètre n'a pratiquement pas de chance de s'y trouver. Dit autrement, on rejette H_0 si $z_C \leq \alpha$ %.
   b. $\bar{W}$ est la région de non-rejet (ou région d'acceptation) de H_0 (et le refus par complémentarité) dans lesquelles doit normalement se trouver le paramètre si H_0 est vraie. $\Pr\left(\bar{W} \setminus H_0\right) = 1 - \alpha$. On en déduit la valeur du risque de deuxième espèce β : $\Pr(W \setminus H_1) = 1 - \beta$.
10. Calculer effectivement la valeur numérique t de la variable de décision en utilisant les résultats apportés par l'échantillon.
11. Donner les conclusions du test sur H_0 en fonction du résultat de la comparaison de la valeur de la probabilité p_{value} au risque α :
    a. si $t \in W$ ou si $p_{value} \leq \alpha$, on rejette l'hypothèse H_0 au risque α au profit de l'hypothèse H_1 sans conclure que l'hypothèse H_0 est fausse, mais elle a une forte probabilité de l'être, le test est significatif ;
    b. si $t \in \bar{W}$ ou si $p_{value} > \alpha$, on ne peut pas rejeter l'hypothèse H_0 au risque α, donc on garde cette hypothèse, le test n'est pas significatif. La différence observée est imputable aux fluctuations de l'échantillonnage. La différence n'est pas significative au seuil α. H_1 est rejetée.

Remarque La p_{value} peut s'interpréter comme **le plus petit seuil de signification** pour lequel l'hypothèse nulle H_0 ne sera pas rejetée. Sa valeur exacte permet de préciser le risque potentiel d'erreur qui accompagne le **rejet** de l'hypothèse nulle H_0. Par exemple, si $p_{value} = 0,02$, on dira que la différence est significative de 2 %.

Le test unilatéral ou le test bilatéral

Le **test unilatéral** est approprié lorsqu'il faut poser une affirmation unilatérale telle que : « plus que », « moins que », « mieux que », « pire que », « au moins », *etc*. Il en existe deux types :

1. le test unilatéral à droite – cela signifie que le rejet de H_0 est à droite ;

2. le test unilatéral à gauche – cela signifie que le rejet de H_0 est à gauche.

Toutefois, il existe des cas pour lesquels il est plus approprié d'employer un **test bilatéral**. On peut souvent reconnaître ces cas par des affirmations symétriques comme « différent de », « échangé en mieux ou en pire », « non égal à », *etc*. En général, chaque fois que l'hypothèse alternative est bilatérale, il est bien venu de calculer la probabilité critique bilatérale pour H_0. Chaque fois que la distribution de l'échantillonnage est symétrique, la probabilité critique bilatérale est exactement le double de la probabilité critique unilatérale.

Il existe deux situations d'application des tests d'hypothèse : soit il s'agit d'une situation d'expérience, soit il s'agit d'une situation d'observations. On considère que les conclusions en situation d'observations sont moins fortes que celle en situation d'expérience.

4.6.2 Le test de signification

Principe

Définition. Soit s_{obs} la valeur observée du score lors d'une expérience et soit S_{rep} le score du test si on répète l'expérience indépendamment et sous H_0. La probabilité $\Pr\left(S_{\text{rep plus extrême que } s_{\text{obs}}}\right)$ est dite p-valeur du test. Ici $\Pr\left((.) \setminus H_0\right)$ est la probabilité calculée en supposant la véracité de H_0.

Le score est utilisé pour quantifier l'écart entre la distribution des observations et l'hypothèse nulle. La phrase « plus extrême que » doit être interprétée dans le sens d'un « témoinage plus fort contre hypothèse que ».

Remarques.

Remarque 1. La formule $\Pr\left((.)\right) \setminus H_0$ est **conditionnelle** sur la valeur observée.

Remarque 2. La p-valeur d'un test élevée si s_{obs} se trouve près du centre de la distribution nulle.

Remarque 3. La p-valeur d'un test élevée si s_{obs} se trouve près des extrêmes de la distribution nulle.

Remarque 4. Si l'hypothèse H_0 est vraie, réaliser s_{obs} dans les queues est improbable.

La procédure générale comprend cinq étapes.

1. On fixe un test de signification à un niveau de α %.
2. On formule l'hypothèse nulle H_0.

3. On choisit un score de test S et on calcule la valeur de ce score s_{obs} pour les observations.
4. On calcule la probabilité :

$$\Pr\left(S_{\text{rep plus extrême que } s_{\text{obs}}}\right) = p\text{-valeur} \tag{4.119}$$

5. On rejette l'hypothèse si la p-valeur est suffisamment petite, en général, si p-valeur $\leq \alpha$ %.

Pour effecteur les calculs nécessaires, la distribution de S si l'hypothèse H_0 était juste doit être trouvée.

Exemple de test de signification

Objectif. Tester l'équilibre de la pièce de monnaie.
On effectue huit jets indépendants. On obtient 1 pile contre 7 faces.

$$\{1, 0, 0, 0, 0, 0, 0, 0\} \tag{4.120}$$

Dans ce cas, on soupçonne que $p = \Pr\{\text{PILE}\}$ est plus petite que $\frac{1}{2}$, la valeur souhaitée pour une pièce équilibrée.

Sous un modèle statistique simple, ces huit résultats sont des réalisations indépendantes d'une variable de Bernoulli $\beta(1, p)$. L'hypothèse nulle à tester est :

$$H_0 : p = \frac{1}{2} \tag{4.121}$$

Le score naturel est :

$$S = \text{nombre de pile} \tag{4.122}$$

ici

$$s_{\text{obs}} = 1 \tag{4.123}$$

Pour calculer la p-valeur, il faut d'abord trouver la distribution de S en supposant que H_0 est vraie. On dit qu'il faut trouver la loi de S sous H_0. Ici, il s'agit d'une loi binomiale $\beta\left(8, \frac{1}{2}\right)$.

Il faut écrire la **distribution nulle** pour cette loi, c'est-à-dire étudier les valeurs de S_{rep} qui donnent le plus fort témoignage contre H_0, soit :

$$\begin{cases} S = 0 \\ S = 8 \end{cases} \begin{cases} S = 1 \\ S = 7 \end{cases} \begin{cases} S = 2 \\ S = 6 \end{cases} \begin{cases} S = 3 \\ S = 5 \end{cases} \tag{4.124}$$

Sous l'hypothèse nulle et une répétition de l'expérience, la probabilité d'observer un score plus extrême que celui observé s_{obs} est :

$$\begin{aligned} &p_{\text{valeur}} = \Pr\left(S_{\text{rep}} = 0, S_{\text{rep}} = 1, S_{\text{rep}} = 7, S_{\text{rep}} = 8 \setminus H_0\right) \\ &\Pr\left(S_{\text{rep}} = 0, S_{\text{rep}} = 1, S_{\text{rep}} = 7, S_{\text{rep}} = 8 \setminus H_0\right) = p_1 + p_2 + p_3 + p_4 \end{aligned} \tag{4.125}$$

Il suffit d'appliquer la formule d'une variable aléatoire binomiale pour obtenir les valeurs de p_1, p_2, p_3 et p_4.

$$\begin{aligned} p_1 &= \binom{8}{0} p^0 (1-p)^8 = (1-p)^8 \\ p_2 &= \binom{8}{1} p^1 (1-p)^7 = 8p(1-p)^7 \\ p_3 &= \binom{8}{7} p^7 (1-p) = 8p^7(1-p) \\ p_4 &= \binom{8}{8} p^8 (1-p)^0 = p^8 \end{aligned} \tag{4.126}$$

p correspond à l'espérance que la pièce soit équilibrée (H_0), c'est-à-dire le cas pour lequel $p = \frac{1}{2}$

$$\begin{aligned} p_1 &= \tfrac{1}{256} \\ p_2 &= \tfrac{1}{32} \\ p_3 &= \tfrac{1}{32} \\ p_4 &= \tfrac{1}{256} \end{aligned} \tag{4.127}$$

$$p_{\text{valeur}} = p_1 + p_2 + p_3 + p_4 \approx 0,07 \text{ soit } 7\ \% \tag{4.128}$$

Conclusion. Dans 7 % des cas, on observe un score plus extrême qu'un pile lors de huit essais avec une pièce équilibrée.

4.6.3 Le test d'hypothèse

L'objectif d'un test d'hypothèse est de vérifier si les données de l'échantillon recueilli sont compatibles, ou non, avec une hypothèse effectuée sur la population. Après examen des résultats de l'échantillon, on pourra rejeter, ou non, l'hypothèse étudiée avec une faible marge d'erreur lorsque celle-ci est rejetée.

Principe

On souhaite étudier une caractéristique θ de la population. L'hypothèse nulle testée H_0 est de la forme $H_0 : \theta = \theta_0$ pour laquelle θ_0 est une **valeur de référence** (ou valeur standard). L'hypothèse alternative H_1 correspond à l'ensemble Θ_1 des valeurs possibles de θ lorsque l'hypothèse H_0 est rejetée. La méthodologie du test d'hypothèse permet de choisir entre H_0 et H_1 à partir d'un échantillon recueilli $x_1, \ldots, x_n$, en contrôlant la probabilité de rejeter H_0 alors qu'elle est vraie. Cette méthodologie comporte différentes étapes.

Étape 1. Poser le test d'hypothèse

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases} \tag{4.129}$$

Étape 2. Définir la statistique utilisée. Pour choisir entre les hypothèses H_0 et H_1, on utilise une statistique $T = f(x_1, \ldots, x_n)$ mesurant un écart normalisé entre l'estimation $\hat{\theta}$ et la valeur testée θ_0.

Étape 3. Définir la loi de probabilité de la statistique T lorsque l'hypothèse H_0 est vraie. Il est nécessaire de connaître, même approximativement, la loi de probabilité de la statistique T lorsque l'hypothèse H_0 est vraie. Dans la plupart des cas, la loi de probabilité de T, sous l'hypothèse H_0 vraie, est une des lois de probabilité connues.

Étape 4. Définir la règle de décision. La règle de décision pour choisir entre les hypothèses H_0 et H_1 utilise la statistique T et est adaptée au problème étudié. Dans certains cas, on rejette H_0 lorsque T est trop grand $(T \geq c)$, dans d'autres cas, lorsque T est trop petit $(T \leq c)$ et dans d'autres cas encore lorsque T et trop petit ou trop grand $(T \leq c_1$ ou $T \geq c_2)$. La suite de la méthodologie du test d'hypothèse va être présentée avec la règle de décision consistant à rejeter H_0 lorsque T est supérieur ou égal à un seuil critique c. L'adaptation aux autres cas est immédiate. La règle de décision utilisée peut conduire à deux types d'erreur :

1. l'**erreur de première espèce** qui rejette H_0 alors que H_0 est vraie ;
2. l'**erreur de deuxième espèce** qui accepte H_0 alors que H_0 est fausse.

Le risque de première espèce α est défini comme la probabilité que la règle de décision conduise à rejeter H_0 alors que H_0 est vraie :

$$\alpha = \Pr\left(T \geq c \setminus H_0 \text{ vraie}\right) \tag{4.130}$$

Le risque de deuxième espèce β est défini pour chaque valeur θ_1 de Θ_1 comme la probabilité d'accepter H_0 alors que le paramètre θ étudié vaut θ_1 :

$$\beta\theta_1 = \Pr\left(T < c \setminus H_1 : \theta = \theta_1\right) \tag{4.131}$$

Il est habituel de calculer le seuil critique c en fixant le risque de première espèce α. Le seuil critique c est le fractile d'ordre $1 - \alpha$ de la loi de probabilité de T lorsque H_0 est vraie.

Étape 5. Mesurer le niveau de signification de la statistique T. On note t la valeur de la statistique T calculée sur l'échantillon observé $x_1, x_2, \ldots, x_n$. Le niveau de signification de la statistique observée t est la plus petite valeur de P de α conduisant au rejet de l'hypothèse H_0 :

$$P = \Pr\left(T \geq t \setminus H_0 \text{ vraie}\right) \tag{4.132}$$

Ainsi, on rejette $H_0 : \theta = \theta_0$ au risque α lorsque $P \leq \alpha$.

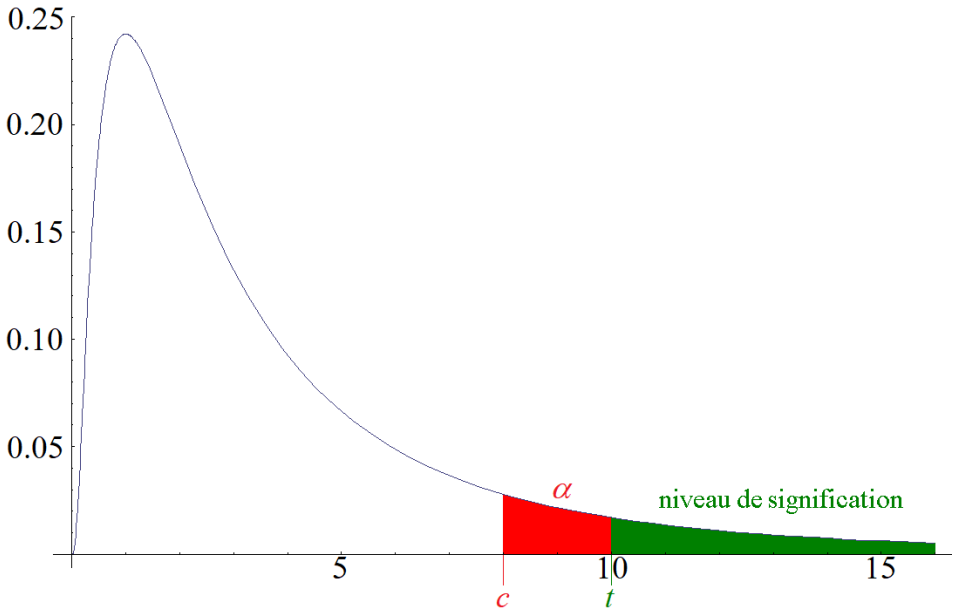


FIGURE 4.5 – Loi statistique T lorsque H_0 est vraie avec le risque α et le niveau de signification P.
H_0 est rejetée pour $c = 8$ et $t = 10$.

Test entre deux hypothèses simples et méthode de Neyman et Pearson

Jerzy Neyman (1894-1981)

Egon Pearson (1895-1980)

Le test de Neyman-Pearson est à la base de la théorisation des tests d'hypothèse.

X est une variable aléatoire dont la densité $f(x, \theta)$ dépend du paramètre réel θ. On extrait un échantillon aléatoire $X^1 = X_1^1, X_2^1, \ldots, X_n^1$ de taille n de cette variable.

On veut tester :

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases} \tag{4.133}$$

Soit $L(x, \theta)$ la densité de probabilité de l'échantillon (ou la vraisemblance).

La région critique W, pour un risque de première espèce valant α, est l'ensemble des points de $\mathbb{R}^n$ défini par :

$$\Pr(W \setminus H_0) = \alpha = \int_W L(x, \theta_0)\, \mathrm{d}x \tag{4.134}$$

La méthode de Neyman et Pearson permet de construire cette région critique. Elle repose sur le théorème qui porte leur nom. Il existe un test T, de puissance maximale, défini par la région critique W au seuil de signification α :

$$W = \{x \in \mathbb{R}^n \setminus L(x, \theta_1) > k_\alpha L(x, \theta_0) \text{ avec } k_\alpha \geq 0\} \tag{4.135}$$

La méthode de Neyman et Pearson consiste à rendre maximale la puissance du test, c'est-à-dire la quantité :

$$\Pr(W \setminus H_1) = 1 - \beta = \int_W L(x, \theta_1)\, \mathrm{d}x \tag{4.136}$$

La démonstration s'effectue en deux étapes :

1. s'il existe une constante k_α telle que l'ensemble W défini par :

$$W = \{x \in \mathbb{R}^n \setminus L(x, \theta_1) > k_\alpha L(x, \theta_0) \text{ avec } k_\alpha \geq 0\} \tag{4.137}$$

   soit de probabilité α sous H_0, alors cet ensemble W réalise le maximum de $(1 - \beta)$.

2. pour démontrer l'existence de la constante k_α, on définit une région $A(K)$ de $\mathbb{R}^n$ telle que :

$$A(K) = \{x \in \mathbb{R}^n \setminus L(x, \theta_1) > K L(x, \theta_0)\} \tag{4.138}$$

   K étant une constante positive donnée.

La probabilité $\Pr\left(A\left(K\right) \setminus H_0\right)$ est une fonction de K, continue, monotone si la variable aléatoire X est à densité continue.

1. Si $K = 0$, $L\left(x, \theta_1\right)$ étant positive, alors $\Pr\left(A\left(0\right) \setminus H_0\right) = 1$.
2. Si $K \to +\infty$, $L\left(x, \theta_1\right)$ étant une densité est bornée, alors $\Pr\left(A\left(h_\alpha\right) \setminus H_0\right) \to 0$.
3. Il existe ainsi une valeur intermédiaire k_α telle que $\Pr\left(A\left(h_\alpha\right) \setminus H_0\right) = \alpha$.

Le théorème de Neyman et Pearson est démontré.

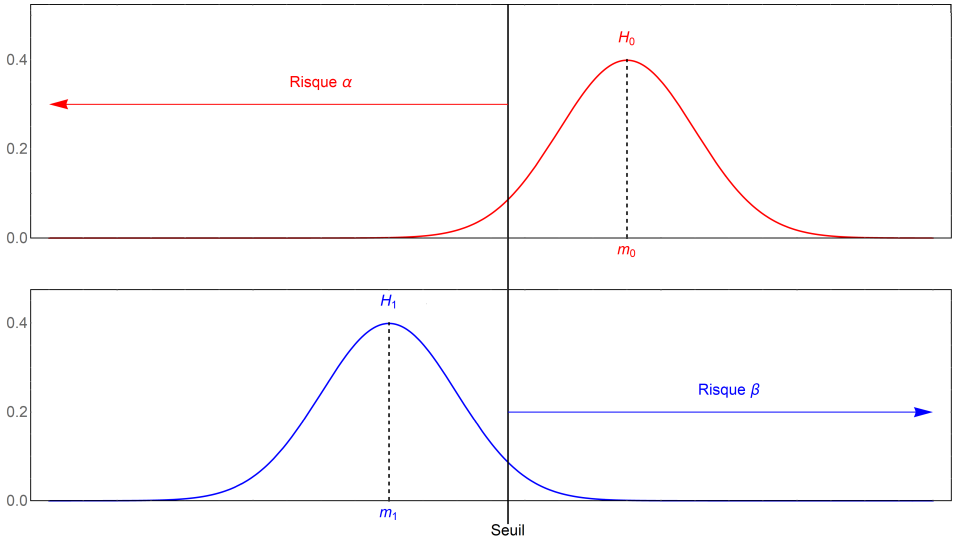


FIGURE 4.6 – Illustration des risques α et β à un seuil défini

L'étude de la puissance $(1 - \beta)$ du test repose sur le fait que $(1 - \beta) > \alpha$. Ce test est sans biais. Par définition,

$$1 - \beta = \int_W L\left(x, \theta_1\right) \mathrm{d}x \tag{4.139}$$

et

$$\alpha = \int_W L\left(x, \theta_0\right) \mathrm{d}x \tag{4.140}$$

Dans la région critique W, on observe $L\left(x, \theta_1\right) > k_\alpha L\left(x, \theta_0\right)$, donc $(1 - \beta) > \alpha k_\alpha$.

— Si $k_\alpha \geq 1$, le résultat est trivial : $(1 - \beta) > \alpha$.

— Si $k_\alpha \leq 1$ et $1-\alpha > \beta$, pour ce, il suffit d'intégrer dans la région $\bar{W}$. Dans la région $\bar{W}$, $L(x,\theta_1) \leq k_\alpha L(x,\theta_0)$, d'où :

$$\int_{\bar{W}} L(x,\theta_1)\,\mathrm{d}x < k_\alpha \int_{\bar{W}} L(x,\theta_0)\,\mathrm{d}x \tag{4.141}$$

avec $\int_{\bar{W}} L(x,\theta_0)\,\mathrm{d}x = 1-\alpha$ et $\beta = \int_{\bar{W}} L(x,\theta_1)\,\mathrm{d}x$.

Le test converge. Lorsque $n \to +\infty$, $1-\beta \to 1$.

Soit $T = \varphi(X_i)$ une statistique exhaustive, la densité de l'échantillon se met alors sous la forme :

$$L(x,\theta) = g(t,\theta)\,h(x) \tag{4.142}$$

avec $t = \varphi(x_i)$. La région critique, selon la méthode de Neyman et Pearson, est alors définie par :

$$g(t,\theta_1) > k_\alpha g(t,\theta_0) \tag{4.143}$$

Elle dépend de fait exclusivement de la statistique exhaustive.

Test entre deux hypothèses composites

Test d'une hypothèse simple contre une hypothèse composite Différents cas peuvent être envisagés, par exemple :

$$\begin{cases} H_0 : \theta = \theta_0 \text{ et } H_1 : \theta > \theta_0 \\ H_0 : \theta = \theta_0 \text{ et } H_1 : \theta \neq \theta_0 \end{cases} \tag{4.144}$$

L'hypothèse H_1 est composée d'un ensemble de valeurs. Le risque de deuxième espèce β doit être calculé pour chaque valeur du paramètre définie par cette hypothèse. On obtient ainsi une fonction $\beta(\theta)$, son graphe est la **courbe d'efficacité du test**, la fonction $1-\beta(\theta)$ est la **puissance du test**. Son graphe la **courbe de puissance du test**.

Un **test** est appelé **uniformément le plus puissant** (U.P.P.) si, quelle que soit la valeur du paramètre θ appartenant H_1, sa puissance est supérieure à la puissance de tout autre test. Il en est ainsi si, par exemple, la région critique ne dépend pas de la valeur θ du paramètre.

Test entre deux hypothèses composites L'hypothèse H_0 est aussi une hypothèse composite et le risque α de première espèce dépend de la valeur du paramètre. On imposera la condition $\alpha(\theta) \leq \alpha$ valeur donnée.

On dèmontre l'existence de tests U.P.P. dans certains cas, par exemple :

$$\begin{cases} H_0 : \theta < \theta_0 \\ H_1 : \theta \geq \theta_0 \end{cases} \tag{4.145}$$

4.6.4 Les tests paramétriques classiques

Les tests paramétriques classiques ont, en partie, été vus lors de la partie sur les méthodes d'estimation. Il s'agit ici de reformuler ce qui est déjà connu sous la forme de tests statistiques.

Tests sur la moyenne et l'écart type

Test sur la moyenne μ L'écart type σ est connu. L'objectif est de déterminer si la moyenne de m_0 de l'échantillon de taille n coïncide avec la moyenne de la population mère μ. Pour ce, on établit deux hypothèses :

$$\begin{cases} H_0 : \mu = m_0 \\ H_1 : m_1 > m_0 \end{cases} \tag{4.146}$$

La variable de décision est la statistique $\bar{X}$ dont la loi est facile à établir en fonction de H_0 et de H_1 :

$$\begin{cases} H_0 : \bar{X} \sim N\left(m_0, \frac{\sigma}{\sqrt{n}}\right) \\ H_1 : \bar{X} \sim N\left(m_1, \frac{\sigma}{\sqrt{n}}\right) \end{cases} \tag{4.147}$$

En désignant par α le risque de première espèce, la région critique est définie par :

$$\alpha = \Pr\left(\bar{X} > k \setminus H_0\right) = \Pr\left(U > \frac{k - m_0}{\frac{\sigma}{\sqrt{n}}}\right) \tag{4.148}$$

La valeur $u = \frac{k-m_0}{\frac{\sigma}{\sqrt{n}}}$ est lue dans la table de la loi normale centrée réduite. On en déduit la valeur de k, donc la région critique.

La forme de l'hypothèse H_1 conduit à rejeter les valeurs trop grandes de $\bar{X}$. Le risque de deuxième espèce β est défini par :

$$\beta = \Pr\left(\bar{X} < k \setminus H_1\right) = \Pr\left(U < \frac{k - m_1}{\frac{\sigma}{\sqrt{n}}}\right) = \Pr\left(U < U_\beta\right) \tag{4.149}$$

avec $U_\beta = \frac{k-m_1}{\frac{\sigma}{\sqrt{n}}}$. Dans la table de la loi normale centrée réduite, on lit la valeur de U_β. Ainsi, la valeur de β vaut :

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m = m_1 < m_0 \end{cases} \tag{4.150}$$

La variable de décision est encore la statistique $\bar{X}$. La région critique est définie par :

$$\alpha = \Pr\left(\bar{X} < k \setminus H_0\right) = \Pr\left(U < \frac{k - m_0}{\frac{\sigma}{\sqrt{n}}}\right) = \Pr\left(U < U_\alpha\right) \tag{4.151}$$

De façon analogue, on peut déduire la valeur de β :

$$\beta = \Pr\left(\bar{X} < k \setminus H_1\right) = \Pr\left(U > \frac{k - m_1}{\frac{\sigma}{\sqrt{n}}}\right) = \Pr\left(U < U_\beta\right) \tag{4.152}$$

Dans ce cas, les deux hypothèses deviennent :

$$\begin{cases} H_0 : m = m_0 \\ H_1 : m = m_1 \neq m_0 \end{cases} \tag{4.153}$$

L'hypothèse H_1 implique $m_1 < m_0$ ou $m_1 > m_0$. La région critique est alors déterminée par la même variable de décision $\bar{X}$ par :

$$\alpha = \Pr\left(\left|\bar{X}\right| < k \setminus H_0\right) = \Pr\left(|U| > \frac{k - m_0}{\frac{\sigma}{\sqrt{n}}}\right) = \Pr\left(|U| > u\right) \tag{4.154}$$

La valeur u est lue dans la table de la loi normale centrée réduite. Une fois estimée, on calcule la valeur de β comme dans les cas précédents.

Remarque 1. Si l'**écart type** σ n'est pas connu mais **estimé**, la variable aléatoire $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n-1}}}$ suit une loi t de Student à $(n-1)$ degrés de liberté. On procède par la suite de façon analogue en utilisant la table de la loi t de Student.

Remarque 2. Si la loi suivie par la variable aléatoire X n'est pas une loi normale, mais, si la taille de l'échantillon est supérieure à 30, on peut admettre que la loi limite est la loi normale. Si l'écart type n'est pas connu, on utilisera de nouveau la loi t de Student.

Test sur l'écart type σ Lorsque la moyenne μ est connue, les hypothèses à tester sont, par exemple :

$$\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma = \sigma_1 > \sigma_0 \end{cases} \tag{4.155}$$

La variable aléatoire $\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$ suit la loi du χ^2 à n degrés de liberté.

La variable de décision est la statistique D, un estimateur sans biais de la variance :

$$D = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 \tag{4.156}$$

La région critique (ou de rejet de H_0) est définie par :

$$\alpha = \Pr\left(D > k \setminus H_0\right) = \Pr\left(\chi^2\left(n\right) > \frac{nk}{{\sigma_0}^2}\right) \tag{4.157}$$

En utilisant la table du χ^2, on peut déterminer k, puis le risque β :

$$\beta = \Pr\left(D < k \setminus H_1\right) = \Pr\left(\chi^2\left(n\right) < \frac{nk}{{\sigma_1}^2}\right) \tag{4.158}$$

Les hypothèses se reformulent alors de manière suivante :

$$\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma = \sigma_1 < \sigma_0 \end{cases} \tag{4.159}$$

La région critique est définie par :

$$\alpha = \Pr\left(D < k \setminus H_0\right) = \Pr\left(\chi^2\left(n\right) < \frac{nk}{{\sigma_0}^2}\right) \tag{4.160}$$

En utilisant la table du χ^2, on détermine k, puis le risque β :

$$\beta = \Pr\left(D > k \setminus H_1\right) = \Pr\left(\chi^2\left(n\right) > \frac{nk}{{\sigma_1}^2}\right) \tag{4.161}$$

Remarque 1. Si la **moyenne** μ n'est pas connue mais **estimée**. Dans ce cas, la variable de décision est la statistique :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \tag{4.162}$$

La variable aléatoire $\frac{nS^2}{\sigma^2}$ suit la loi du χ^2 à $(n-1)$ degrés de liberté. Les différents tests s'étudient comme dans le cas où la moyenne est connue.

Remarque 2. Tous ces résultats ne sont valables que dans le cas où la variable aléatoire X suit une loi normale.

Tests sur une proportion

Le but est de tester si la proportion p d'individus d'une population, présentant un certain caractère qualitatif peut être considérée comme égale ou non à une valeur p_0.

Un estimateur sans biais de p est la proportion F d'individus présentant u caractères dans un échantillon aléatoire de taille n. La variable aléatoire du nombre d'individus présentant ce caractère $K = nF$ suit la loi binomiale $B(n, p)$. Si np et $n(1-p)$ sont supérieurs à S, on peut remplacer la loi binomiale par une loi normale. On en déduit que F suit approximativement la loi normale $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

Test d'hypothèse par intervalle de confiance

Il s'agit de considérer un intervalle de confiance simplement comme l'ensemble des hypothèses acceptables. Une fois calculé, un intervalle de confiance peut être employé immédiatement pour tester n'importe quelle hypothèse. « Statistiquement significatif au seuil de signification de 5 % » est la phrase traditionnelle que l'on rencontre typiquement dans la littérature scientifique. Elle signifie exactement la même chose que « statistiquement discernable au seuil d'erreur de 5 % ». L'hypothèse H_0 consiste à établir si oui ou non elle tombe dans l'intervalle de confiance.

Tests bilatéraux et intervalles de confiance Il est facile de transformer un test unilatéral en test classique bilatéral. En effet, il suffit de rejeter H_0 si la probabilité critique bilatérale tombe en-dessous du seuil fixé α.

Dans une situation répétitive, pour laquelle on doit tester beaucoup d'échantillons, on peut, au lieu de calculer les nombreuses probabilités critiques, calculer une fois pour toute une région de rejet classique. Chaque fois qu'un test bilatéral est approprié un intervalle de confiance bilatéral ordinaire l'est également. Si l'hypothèse nulle H_0 se situe en dehors de l'intervalle de confiance, on peut rejeter H_0. La raison pour laquelle l'intervalle de confiance et le test sont équivalents est claire : dans les deux cas, on vérifie simplement si la différence entre la valeur observée et la valeur de l'hypothèse nulle dépasse $1,96 \times \mathrm{SE}$ pour un seuil de 5 %. Ces deux approches ne diffèrent que parce que le test classique emploie l'hypothèse nulle comme point de référence, alors que l'intervalle de confiance utilise la valeur observée comme référence.

Tests unilatéraux et intervalles de confiance De la manière qu'un intervalle de confiance bilatéral est équivalent à un test bilatéral, on peut développer un intervalle de confiance unilatéral équivalent à un test unilatéral.

L'intervalle de confiance unilatéral situe la totalité de la marge d'erreur de 5 % dans une seule queue de distribution. Cela signifie qu'il n'existe qu'un point limite que d'un côté de la distribution, et pas de limite de l'autre côté.

Remarque importante. N'importe quel intervalle de confiance bilatéral peut être rendu unilatéral. Par exemple,

$$(\mu_1 - \mu_2) = (X_1 - X_2) + t_{0,05} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{4.163}$$

avec μ_1 la valeur de l'hypothèse H_0 à droite, μ_2 la valeur de l'hypothèse H_0 à gauche, X_1 la valeur observée à droite, X_2 la valeur observée à gauche, $t_{0,05}$ la variable t de Student à un seuil d'erreur de 5 %, ${s_p}^2$ la variance commune définie telle que ${s_p}^2 = \frac{\sum_{i=1}^{n}\left(X_1 - \bar{X_1}\right)^2 + \sum_{i=1}^{n}\left(X_2 - \bar{X_2}\right)^2}{(n_1 - 1) + (n_2 - 1)}$, n_1 l'effectif de l'échantillon n° 1 et n_2 l'effectif de l'échantillon n° 2.

L'intervalle de confiance bilatéral est la forme type qui est la plus facile à comprendre. Il fournit habituellement la valeur ponctuelle estimée. C'est la forme la plus naturellement liée à certaines techniques plus avancées que l'on étudiera plus tard.

4.6.5 Les tests d'ajustement

Principes

Les tests d'ajustement permettent de **juger l'adéquation entre une situation réelle et un modèle théorique**.

Deux problèmes différents peuvent se rencontrer en statistique :

1. soit ajuster une loi de probabilité à un échantillon. La loi est inconnue. Sa forme et les valeurs des paramètres sont obtenues à partir des caractéristiques de l'échantillon ;
2. soit ajuster un échantillon à une loi de probabilité donnée, la loi est connue, c'est-à-dire que la fonction de répartition ou la densité sont entièrement spécifiées. On doit vérifier l'adéquation entre la loi théorique et l'échantillon.

Le choix d'une loi est lié :

1. à la nature du phénomène étudié afin de choisir entre loi discrète et loi continue ;

2. à la forme de la distribution par un histogramme ;
3. à la connaissance et à l'interprétation des principales caractéristiques de l'ensemble des données, espérance, médiane, variance, ou écart type, coefficients d'asymétrie et d'aplatissement... ;
4. au nombre de paramètres des lois, une loi dépendant de plusieurs paramètres peut s'adapter plus facilement à une distribution donnée.

Une loi étant proposée, différents tests peuvent utilisés pour juger de la concordance entre une distribution théorique et une distribution réelle :

1. le test le plus utilisé est le **test de Pearson** (dit χ^2). Il peut aussi être utilisé pour tester l'égalité de k proportions, l'indépendance de deux variables aléatoires étudiées suivant différentes modalités par le tableau de contingence ;
2. le **test de Kolmogorov-Smirnov** ;
3. le **test de Cramér-Von-Mises**.

Un **test d'ajustement** permet de juger si une hypothèse concernant une loi de probabilité, c'est-à-dire une loi théorique, est compatible avec la réalisation d'un échantillon de taille n d'une variable aléatoire X.

1. Mise en œuvre d'un test d'ajustement :
   a. Prélever un échantillon suffisamment important de la population étudiée ;
   b. Classer les observations par ordre croissant dans le cas d'une variable aléatoire continue, d'égale amplitude ou d'égale probabilité ;
   c. Définir une variable de décision D permettant de mesurer les écarts entre la distribution théorique F et la distribution empirique F^* de l'échantillon.
2. Vérification de la concordance des deux distributions :
   a. Définir les hypothèses H_0 et H_1.
      — H_0 : les observations suivent une distribution théorique spécifiée : $F = F_0$.
      — H_1 : les observations ne suivent pas la distribution théorique spécifiée : $F \neq F_0$.
   b. Accepter un risque de première espèce α de refuser l'hypothèse H_0 alors qu'elle est vraie.
   c. Calculer la valeur d de la variable de décision D à partir des valeurs données par l'échantillon.
   d. Énoncer une règle de décision.

— On rejette l'hypothèse H_0 si la valeur calculée d est supérieure à une valeur d_0 n'ayant qu'une probabilité α d'être dépassée par la variable D.

— Sinon, on garde l'hypothèse H_0. On considère que la distribution théorique spécifiée peut décrire le phénomène étudié, c'est-à-dire $F = F_0$.

Tests d'ajustement – Méthodes empiriques

La forme de l'histogramme La forme de l'histogramme permet de privilégier certains modèles si des conditions de symétrie sont respectées, ou au contraire d'éliminer des modèles.

1. Une distribution symétrique peut suggérer une loi normale, une loi de Cauchy ou une loi de Student.
2. Une distribution dissymétrique fait penser à une loi log-normale, à une loi gamma, à une loi de Weibull ou à une loi bêta de type II.

Toutefois, puisque l'on étudie des phénomènes réels, certains modèles devront être privilégiés alors que d'autres devront être systématiquement écartés. Par exemple, les lois utilisées en fiabilité sont surtout les lois exponentielles ou de Weibull.

La vérification de certaines propriétés mathématiques L'échantillon permet de calculer $\bar{x}$ et s^2, c'est-à-dire des estimations de l'espérance mathématique μ et de la variance σ^2.

L'ajustement graphique Soit L une loi de probabilité de la fonction de répartition F. Cette fonction varie de 0 à 1 et est représentée dans un plan par une courbe Γ.

On considère une série classée, par ordre croissant, de n observations réparties en k classes d'effectifs n_i. La fonction de répartition empirique F^* de l'échantillon doit être peu différentes de la fonction de répartition théorique F.

Soit z_i le centre de la classe $[x_i - 1, x_i]$ et h_i l'étendue de cette classe ; le point P_i d'abscisse :

$$x_{i-1} + \frac{k_i}{2} \tag{4.164}$$

et d'ordonnée :

$$\frac{1}{n} \sum_{j=1}^{n} n_j \tag{4.165}$$

est un point de la fonction empirique.

Si les points P_i ne sont pas trop éloignés de la courbe Γ, on peut admettre que la loi suivie par les observations est voisine de la loi L.

Pour estimer la distance des points P_i à la courbe théorique Γ, on cherche une transformation mathématique plus simple, permettant de représenter la fonction de répartition par une droite. Cette transformation, ou anamorphose, existe pour la plupart des lois de probabilité.

Loi normale et droite de Henry Soit U la variable centrée réduite associée à la variable normale X. Les valeurs de U sont comprises entre $-\infty$ et $+\infty$.

Par ailleurs, il existe une relation linéaire simple entre les variables X et U :

$$U = \frac{X - \mu}{\sigma} \tag{4.166}$$

La transformée de la fonction de répartition dans le plan (U, X) est une droite de pente $\frac{1}{\sigma}$ est appelée **droite de Henry** (1894). On utilise un papier spécial, que l'on trouve dans le commerce, dit gausso-arithmétique. Il suffit de graduer l'axe des ordonnées selon les valeurs de F, mais proportionnellement aux valeurs de U (Tab. 4.8).

U	$F(U)$
-2	0,0228
-1	0,1583
0	0,5000
1	0,8417
2	0,9772

TABLE 4.8 – Exemple de valeurs normales pour une loi centrée réduite

On répète ce procédé pour toutes les valeurs de la variable U. On peut, de la même façon, choisir les valeurs de F et en déduire les valeurs de U.

Pour vérifier si un échantillon est extrait d'une population normale, on porte :

1. en abscisse, les valeurs des observations, c'est-à-dire les limites supérieures des classes ;
2. en ordonnée, les frèquences cumulées correspondantes.

Si les points sont sensiblement alignès, on peut accepter comme distribution thèorique une loi normale.

L'intersection avec la droite $U = 0$ et $F(U) = 0,50$ donne la valeur de l'espèrance mathèmatique $\mathbb{E}(X) = \mu$.

La valeur σ peut être obtenue de deux façons.

1. Si $U = 1$, alors $F(U) = 0,8415$ et $x_i - \mu = \sigma$.
2. Si $U = -1$, alors $F(U) = 0,1585$ et $x_i - \mu = -\sigma$.

Loi exponentielle On suppose que la durée de vie d'un composant suive une loi exponentielle de fonction de répartition F :

$$\Pr(X > x) = e^{-\lambda x} = 1 - F(x) \tag{4.167}$$

d'où $\ln(1 - F(x)) = -\lambda x$.

Si on dispose d'un échantillon de taille n, on porte :

1. en abscisse, les temps x_i de fonctionnement ;
2. en ordonnée, les pourcentages de « survivants » au temps x_i, en utilisant une échelle logarithmique.

En pratique, on ordonne les temps x_i par valeurs croissantes, et on prend pour ordonnées correspondantes, les valeurs $\ln\left(1 - \frac{1-i}{n}\right)$ avec $1 < i < n$.

Si l'échantillon est représentatif de la population, les points sont pratiquement alignés. Pour estimer la valeur de λ, c'est-à-dire la pente de la droite, on remarque que si $x = 1$, $\ln(1 - F(x)) = -\lambda$. L'intersection avec la droite $x = 1$ donne alors une estimation graphique du paramètre.

Loi de Weibull Une variable aléatoire réelle T suit une loi de Weibull si sa fonction de répartition F est :

$$\begin{cases} \forall t < \gamma, F(t) = 0 \\ \forall t \geq \gamma, F(t) = 1 - e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}} \end{cases} \tag{4.168}$$

et sa densité f est :

$$\begin{cases} \forall t < \gamma, f(t) = 0 \\ \forall t \geq \gamma, f(t) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}} \end{cases} \tag{4.169}$$

Le paramètre de position γ peut être pris égal à 0 (qui est une simple translation sur t), d'où la fonction de répartition simplifiée :

$$F(t) = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{4.170}$$

Les constantes β et η peuvent être estimées par la méthode du maximum de vraisemblance. Une estimation graphique de ces constantes est obtenue en utilisant le papier de Weibull ou d'Allen Plait en logarithme du logarithme en abscisse et en logarithme en ordonnée.

Allen Plait (-)

La transformation mathématique qui donne pour représentation de la fonction de répartition une droite, est la suivante :

$$\begin{cases} X = \ln t \\ Y = \ln\left(\ln\left(\frac{1}{1-F(t)}\right)\right) = \beta\left(X - \ln\eta\right) \end{cases} \tag{4.171}$$

La pente de la droite donne la valeur de β. L'intersection de la droite empirique avec $Y = 0$ donne la valeur de η. La valeur $Y = 0$ correspond à $F\left(t\right) = 0,632$.

$$\begin{cases} X = \ln t_i \\ Y = \ln\left(-\ln\left(1 - \frac{i-1}{n}\right)\right) \end{cases} \tag{4.172}$$

Un rapport permet de lire la valeur β, elle est comprise entre 0 et 4. Une échelle verticale donne, pour les valeurs de β, les valeurs de $\Gamma\left(1 + \frac{1}{\beta}\right)$; elle permet de calculer une estimation de l'espérance mathématique qui est égale à :

$$\mathbb{E}\left(T\right) = \eta\Gamma\left(1 + \frac{1}{\beta}\right) \tag{4.173}$$

Dans chaque classe, y compris les classes extrêmes, on doit avoir au moins cinq observations. Si cette condition n'est pas remplie, on regroupe certaines classes par le processus de discrétisation.

En fait, ce procédé n'est pas un test, mais une méthode rapide et simple pour voir si une distribution observée et compatible avec une loi que l'on s'est fixée à l'avance. Elle permet aussi de comparer les valeurs lues sur le graphique pour les paramètres aux estimations calculées sur l'échantillon.

Tests d'ajustement – Le test du χ^2

Le test du χ^2 utilise des propriétés de la loi multinomiale.

1. La fonction de répartition F est entièrement spécifiée, en particulier, les paramètres sont connus.
2. On connaît seulement la forme de la loi de distribution. Les paramètres de la fonction de répartition F sont estimés à partir d'un échantillon.

Soit X la variable aléatoire parente de fonction de répartition F. On considère une partition du domaine de définition en r intervalles $I_1, \ldots, I_r$ d'égale étendue ou non.

Pour chaque intervalle I_i, on considère l'ensemble E_i tel que :

$$\begin{cases} E_i = \{\omega; X(\omega) \in I_i\} \\ p_i = \Pr(E_i) \end{cases} \tag{4.174}$$

np_i est égal à la fréquence (absolue) théorique espérée dans la classe I_i que l'on compare à la fréquence observée N_i dans cette même classe I_i.

La variable de décision suit la statistique :

$$D^2 = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i} \tag{4.175}$$

Si l'hypothèse H_0 est vraie, c'est-à-dire qu'il existe une concordance acceptable entre la distribution théorique et les valeurs observées, cette quantité ne peut être trop grande. En fait, K. Pearson a montré que la statistique D^2 suit une loi du χ^2 à υ degrés de liberté quelle que soit la loi considérée, lorsque le nombre n d'observations tend vers l'infini. Le nombre υ degrés de liberté est égal à :

— $(r-1)$ si la distribution théorique est entièrement déterminée, aucun paramètre n'ayant été estimé ;

— $(r-1-k)$ si k paramètres ont été estimés à partir des observations pour définir complètement la distribution.

La règle de décision est la suivante : on rejette l'hypothèse H_0 si la valeur de la statistique D^2 obtenue à partir de l'échantillon est supérieure à une valeur n'ayant qu'une probabilité α d'être dépassée par la variable χ^2 considérée. Sinon, on garde l'hypothèse H_0 et on considère que la distribution théorique spécifiée est acceptable, c'est-à-dire $F = F_0$.

Remarque 1. La distribution limite de la statistique D^2 est indépendante de la loi F, ce test peut alors être utilisé dans de nombreuses situations.

Remarque 2. Les effectifs de chaque classe doivent être supérieurs à cinq. Si cette condition n'est pas vérifiée, on regroupe les classes d'effectifs trop faibles.

Exemple Un générateur a produit 1 000 nombres compris entre 0 et 1. Leur répartition est la suivante (Tab. 4.9).

On a obtenu 113 nombres entre 0 et 0,09. Si la répartition était uniforme, on aurait dû obtenir 100 nombres dans chaque classe. Un test du χ^2 permet de garder ou de rejeter cette hypothèse.

$$\begin{array}{l} D^2 = \frac{(113-100)^2}{100} + \frac{(73-100)^2}{100} + \frac{(125-100)^2}{100} + \frac{(115-100)^2}{100} + \frac{(90-100)^2}{100} + \\ + \frac{(101-100)^2}{100} + \frac{(95-100)^2}{100} + \frac{(93-100)^2}{100} + \frac{(110-100)^2}{100} + \frac{(85-100)^2}{100} \\ D^2 = 22,48 \end{array} \tag{4.176}$$

x	n
0,00 à 0,09	113
0,10 à 0,19	73
0,20 à 0,29	125
0,30 à 0,39	115
0,40 à 0,49	90
0,50 à 0,59	101
0,60 à 0,69	95
0,70 à 0,79	93
0,80 à 0,89	110
0,90 à 0,99	85
Total	**1 000**

TABLE 4.9 – Exemple de répartition de 1 000 nombres aléatoires en 10 classes.

Sous l'hypothèse de loi uniforme, la variable D^2 suit une loi du χ^2 à $(10-1) = 9$ degrés de liberté, aucun des paramètres n'ayant été estimé :

$$\begin{array}{l} \Pr\left(\chi^2\left(9\right) > 16,9\right) = 0,05 \\ \Pr\left(\chi^2\left(9\right) > 19\right) = 0,025 \end{array} \tag{4.177}$$

On doit de fait rejeter l'hypothèse de loi uniforme, car la valeur 22,48 a une probabilité inférieure à 0,025 de se réaliser.

Remarque. Le générateur de nombre n'est pas à l'abri de toute critique.

Tests d'ajustement – Test de Kolmogorov (1933)-Smirnov (1939)

On suppose que la fonction de répartition F de la variable aléatoire X est continue et strictement croissante. Soit F^* la fonction de répartition empirique d'un échantillon de talle n issu de cette population.

La variable de décision est la variable aléatoire D_n définie par :

$$D_n = \sup_{x \in \mathbb{R}} \left|F^*\left(x\right) - F\left(x\right)\right| \tag{4.178}$$

V. I. Glivenko et A. N. Kolmogorov ont démontré que la fonction $K_n\left(y\right)$ définie par :

Valery Ivanovitch Glivenko (1896-1940)

$$K_n\left(y\right) = \Pr\left(\sqrt{n}D_n < y\right) \tag{4.179}$$

converge, lorsque n tend vers l'infini, vers une fonction $K_n\left(y\right)$:

$$\begin{cases} K(y) = 0 \\ K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2y^2} \end{cases} \tag{4.180}$$

Des tables fournissent la valeur de la fonction K.

La règle de la décision est la suivante. On rejette l'hypothèse H_0 si la valeur statistique D_n, obtenue à partir de l'échantillon, est supérieure à une valeur $d(n)$ n'ayant qu'une probabilité α d'être dépassée. Sinon, on conserve l'hypothèse H_0 et on considère que la distribution théorique spécifiée est acceptable, c'est-à-dire $F = F_0$.

Remarque. Le test de Kolmogorov-Smirnov est préférable au test du χ^2 pour des variables continues. En effet, la variable aléatoire de décision D_n utilise l'échantillon tel qu'il présente, en revanche, le test du χ^2 appauvrit l'information en regroupant les données par classes et en assimilant les données d'une classe à la valeur centrale.

Tests d'ajustement – Test de Cramér-Von Mises

On considère la statistique $n\omega^2$ définie par :

$$n\omega^2 = \int_{-\infty}^{+\infty} F(x) - F^*(x)\, \mathrm{d}F(x) \tag{4.181}$$

Il existe des tables pour la loi de cette variable aléatoire, loi indépendante de F. Elle est utilisée pour évaluer l'écart entre une distribution empirique et une distribution théorique. Si les valeurs x_i de l'échantillon sont ordonnées par valeurs croissantes, on démontre que :

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F(x_i) \right]^2 \tag{4.182}$$

On rejette l'hypothèse H_0 si la valeur de la statistique $n\omega^2$ donnée par l'échantillon est supérieure à une valeur n'ayant qu'une probabilité α d'être dépassée. Pour $\alpha = 0,05$, on rejette l'hypothèse H_0 si $n\omega^2 > 0,46136$.

Tests d'ajustement – Test de normalité (ou de conformité)

La moyenne de l'échantillon est-elle conforme à la valeur attendue dans la population de référence ? La différence entre les deux moyennes est-elle significative ?

Soit un échantillon de taille n, on pose l'hypothèse nulle H_0 :

$$\begin{array}{l} X \sim N(\mu, \sigma) \\ \mu \text{ est estimée par } \bar{x} \\ \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \\ \sigma \text{ est estimé par } \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{array} \tag{4.183}$$

La règle de décision est la suivante. On rejette $H_0 : \bar{x} = \mu$ si $t = \frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$:

— au seuil $\alpha = 0,05$, soit si $\left(\left(\sqrt{n} + \frac{0,85}{\sqrt{n}}\right) - 0,01\right) D_n > 0,895$, soit si $\left(1 + \frac{0,50}{n}\right) n\omega^2 > 0,126$;

— au seuil $\alpha = 0,01$, soit si $\left(\left(\sqrt{n} + \frac{0,85}{\sqrt{n}}\right) - 0,01\right) D_n > 1,035$, soit si $\left(1 + \frac{0,50}{n}\right) n\omega^2 > 0,178$.

avec $D_n = \sup_{x\in\mathbb{R}} |F^*(x) - F(x)|$, $F^*(x)$ la fonction de répartition empirique et $F(x)$ la fonction de répartition théorique.

Tests d'ajustement – Test de normalité de Shapiro-Wilk (1965)

Samuel Sanford Shapiro (1930-2023)
Martin Bradley Wilk (1922-2013)

L'hypothèse nulle H_0 pose que la variable suit une loi normale dans la population d'origine. L'objectif de ce test consiste à ne pas rejeter H_0. Le test est très performant, notamment pour les petits effectifs $n \leq 50$. En pratique, on cherche à limiter le risque de se tromper en obtenant une p_{value} supérieure à 0,2. Ce test permet de justifier le test de Student, car, si H_0 est rejetée, il est impossible.

On classe la série x_i de données par ordre croissant $x_{(i)}$.

Pour ce, on utilise une statistique W :

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i \left(x_{(n-i+1)} - x_{(i)}\right)\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.184}$$

avec a_i les constantes générées à partir de la moyenne et la matrice de covariance des quantiles d'un échantillon de taille n suivant la loi normale – ces constantes sont fournies dans des tables spécifiques ;

$$[a_1, \ldots, a_n] = \frac{{}^t m V^{-1}}{\left({}^t m V^{-1} V^{-1} m\right)^2} \tag{4.185}$$

avec

$$m = {}^t[m_1, \ldots, m_n] \tag{4.186}$$

m les espérances des statistiques d'ordre d'une échantillon de variables indépendantes et identiquement distribuées suivant une loi normale, et V la matrice de covariance de ces statistiques d'ordre ; $x_{(i)}$; $x_{(i)}$ la série des données triées ; $\bar{x}$ la moyenne de la série ; $\left[\frac{n}{2}\right]$ la partie entière du rapport $\frac{n}{2}$.

N.B. Il existe une table pour les valeurs de a_i en fonction de l'effectif.

W peut être interprété comme le coefficient de détermination entre la série de quantiles générées à partir de la loi normale et les quantiles empiriques obtenues à partir des données.

Plus W est élevé, plus la compatibilité avec la loi normale est crédible. La région critique de rejet de la normalité s'écrit :

$$W < W_C \tag{4.187}$$

Les valeurs W_C pour différents risques α et les effectifs n sont lues dans la table de Shapiro-Wilk.

Il existe d'autres tests de normalité :

— le test de Kolmogorov-Smirnov ;

— le test du χ^2 ;

— le test de Lilliefors.

Tests d'ajustement – Test d'exponentiabilité

On pose l'hypothèse nulle H_0 la densité de la loi d'exponentialité $f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}$. Le paramètre θ est estimée par la moyenne de l'échantillon $\bar{x}$.

La règle de décision est la suivante. On rejette H_0 :

— au seuil $\alpha = 0,05$, soit si $\left(\left(\sqrt{n} + \frac{0,85}{\sqrt{n}}\right) - 0,26\right)\left(D_n - \frac{0,2}{n}\right) > 1,094$, soit si $\left(1 + \frac{0,16}{n}\right) n\omega^2 > 0,224$;

— au seuil $\alpha = 0,01$, soit si soit si $\left(\left(\sqrt{n} + \frac{0,50}{\sqrt{n}}\right) - 0,26\right)\left(D_n - \frac{0,2}{n}\right) > 1,308$, soit si $\left(1 + \frac{0,16}{n}\right) n\omega^2 > 0,337$.

avec $D_n = \sup_{x\in\mathbb{R}} |F^*(x) - F(x)|$, $F^*(x)$ la fonction de répartition empirique et $F(x)$ la fonction de répartition théorique.

4.6.6 Les tests de comparaison

Test de comparaison d'échantillons

Les tests de comparaison d'échantillons sont utilisés pour comparer deux ou plusieurs échantillons.

Pour les tests paramétriques, on peut comparer :

— deux moyennes de deux échantillons non normaux ;

— des moyennes de deux échantillons normaux indépendants de variances inconnues mais supposées identiques. On peut utiliser :

  — le test de Fisher-Snedecor ;

  — le test de Student.

— des moyennes de deux échantillons normaux avec les variances connues.

Pour les tests non paramétriques, on peut utiliser :

— le test de Smirnov ;

— le test de Wilcoxon.

Tests de comparaison – Tests paramétriques de comparaison des moyennes de deux échantillons

On considère deux échantillons aléatoires de tailles n_1 et n_2, prélevés indépendamment l'un de l'autre. On se pose la question : « Sont-ils, ou non, issus de la même population ? ». Soient X_1 la variable aléatoire parente et F_1 la fonction de répartition de la population dont est issu le premier échantillon, X_2 et F_2 les mêmes caractéristiques pour le second. Le test correct est le suivant :

$$\begin{cases} H_0 : F_1(x) = F_2(x) \\ H_1 : F_1(x) \neq F_2(x) \end{cases} \tag{4.188}$$

mais, il est beaucoup trop vague. Dans la pratique, on traite le problème plus général suivant : la comparaison des moyennes m_1 et m_2 de deux populations connaissant les estimations données par deux échantillons indépendants de tailles n_1 et n_2.

Pour caractériser la variable de décision $\bar{D} = \bar{X}_1 - \bar{X}_2$, il faut connaître la forme de la loi suivie par cette variable, son espérance mathématique et sa variance.

Selon les données, on distingue trois situations différentes :

1. la comparaison des moyennes de deux échantillons normaux indépendants, les variances étant connues ;
2. la comparaison des moyennes de deux échantillons normaux indépendants, les variances n'étant pas connues mais supposées égales ;
3. la comparaison des moyennes de deux échantillons non normaux indépendants.

Comparaison des moyennes de deux échantillons normaux indépendants, les variances étant connues (ou test d'homogénéité) La différence entre deux moyennes est-elle significative ?

Soient $N(\mu_i, \sigma_i)$, les lois suivies par les deux populations ($i = 1$ ou $i = 2$). La variable aléatoire $\bar{D}$ suit alors une loi normale :

$$\begin{cases} N(\mu_1 - \mu_2, \sigma_{\bar{D}}) \\ \sigma_{\bar{D}} = \sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}} \end{cases} \tag{4.189}$$

On considère la variable aléatoire centrée réduite $U = \frac{\bar{D} - (\mu_1 - \mu_2)}{\sigma_{\bar{D}}}$ avec $\mu_1 - \mu_2 = 0$:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \tag{4.190}$$

Sous H_0 et compte tenu de H_1, la région critique est de la forme $|U| > k$.

Comparaison des moyennes de deux échantillons normaux indépendants, les variances n'étant pas connues mais supposées égales Les hypothèses sont :

$$\begin{cases} H_0 : \mu_1 = \mu_2 \quad \sigma_1 = \sigma_2 \\ H_1 : \mu_1 \neq \mu_2 \quad \sigma_1 \neq \sigma_2 \end{cases} \tag{4.191}$$

À partir des estimations données par les échantillons, on doit vérifier dans l'ordre suivant :

1. l'égalité des variances ;
2. l'égalité des moyennes ou test de Fisher-Snedecor.

Test de l'égalité des variances ou test de Fisher-Snedecor Les hypothèses sont :

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 \\ H_1 : \sigma_1 \neq \sigma_2 \end{cases} \tag{4.192}$$

Comme les populations sont normaux, on sait que :

$$\frac{n_1 {s_1}^2}{(n_1 - 1)\,{\sigma_1}^2} \times \frac{(n_2 - 1)\,{\sigma_2}^2}{n_2 {s_2}^2} = F(n_1 - 1, n_2 - 1) \tag{4.193}$$

Sous H_0, on obtient :

$$\frac{n_1 {s_1}^2}{(n_1 - 1)} \times \frac{(n_2 - 1)}{n_2 {s_2}^2} = F(n_1 - 1, n_2 - 1) \tag{4.194}$$

que l'on peut écrire, en introduisant les estimateurs sans biais des deux variances :

$$\begin{array}{c} \frac{n_i {s_i}^2}{(n_i-1)} = {\hat{\sigma}_i}^2 \text{ avec } i = \{1,2\} \\ \frac{{\hat{\sigma}_1}^2}{{\hat{\sigma}_2}^2} = F\left(n_1-1, n_2-1\right) \end{array} \tag{4.195}$$

donc, sous H_0, le rapport des estimateurs sans biais des deux variances est une variable aléatoire de Fisher.

Les conclusions du test sont obtenues en calculant le rapport :

$$F = \frac{n_1 {s_1}^2}{(n_1-1)} \times \frac{(n_2-1)}{n_2 {s_2}^2} \tag{4.196}$$

pour les valeurs données par les échantillons.

L'hypothèse alternative étant $H_1 : {\sigma_1}^2 \neq {\sigma_2}^2$, la règle de décision consiste à rejeter H_0 si $F < F_{1-\frac{\alpha}{2}}\left(n_1-1, n_2-1\right)$ ou $F > F_{\frac{\alpha}{2}}\left(n_1-1, n_2-1\right)$ avec le seuil critique α.

Remarque. Pour déterminer la région critique, on doit toujours avoir une valeur du rapport F supérieure à 1.

Test de comparaison de deux moyennes ou test des espérances de Student
Si le test de Fisher-Snedecor a permis de conclure à l'égalité des variances des deux populations, la variable de décision $\bar{D} = \bar{X_1} + \bar{X_2}$ suit la loi normale de paramètres :

$$\mathbb{E}\left(\bar{D}\right) = \mu_1 - \mu_2 \tag{4.197}$$

et

$$\mathbb{V}\left(\bar{D}\right) = \sigma_{\bar{D}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \tag{4.198}$$

La variable aléatoire $\frac{\left(\bar{X_1}-\bar{X_2}\right)-(\mu_1-\mu_2)}{\sqrt{\sum_{i=1}^{n}\left(X_i^1-\bar{X}\right)^2+\sum_{i=1}^{n}\left(X_i^2-\bar{X}\right)^2}} \times \frac{\sqrt{n_1+n_2-2}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, avec X_i^1 et X_i^2 les variables composant les deux échantillons, suit une loi de Student à (n_1+n_2-2) degrés de liberté.

Sous H_1, $\mu_1 \neq \mu_2$, la région critique est de la forme $|T| > k$.

Remarque. Il est indispensable de tester avant tout l'égalité des variances pour appliquer le test de Student.

Comparaison des moyennes de deux échantillons non normaux indépendants Si les populations ne sont pas normaux, on ne peut pas appliquer le test des variances de Fisher. Toutefois, si les effectifs des échantillons sont assez grands, supérieurs à 30, on peut tester l'égalité des moyennes, que les variances soient égales ou non, avec la formule de Student. Le test de Student est un test robuste ; il est insensible à une modification des hypothéses de base.

Tests de comparaison – Tests non paramétriques de comparaison

Le problème consiste à décider si deux échantillons de tailles n_1 et n_2 sont issus ou non d'une même population de fonction de répartition F. Différents tests sont proposés. Ils interviennent si les conditions de Student ne sont pas vérifiées, à savoir que les distributions des variables sont normales et que les variances sont homogènes dans les échantillons qui sont comparés. Cette dernière condition est la propriété d'**homoscédasticité**[6].

Nikolai Smirnov (1900-1966)

Test de Smirnov Soient F^* et G^* les fonctions de répartition empiriques des deux échantillons.

Les hypothèses sont :

$$\begin{cases} H_0 : F(x) = G(x) \\ H_1 : F(x) \neq G(x) \end{cases} \tag{4.199}$$

On montre que :

$$\Pr\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup |F^*(x) - G^*(x)| < y\right) \to K(y) \tag{4.200}$$

On rejette H_0 si la valeur de la statistique $\sqrt{\frac{n_1 n_2}{n_1+n_2}} \sup |F^*(x) - G^*(x)|$ calculée à partir des échantillons est supérieure à une valeur qui a une faible probabilité d'être dépassée.

Remarque. La fonction $K(y)$ est celle du test de Kolmogorov-Smirnov.

Frank Wilcoxon (1892-1965)

Test de Wilcoxon (1945) Ce test ressemble beaucoup au test de Mann-Whitney. Ce sont des tests qui fonctionnent aussi bien sur des variables paramétriques que sur des variables non paramétriques. L'hypothèse nulle H_0 à vérifier est que la distribution est la même dans les deux groupes à comparer. Ces deux tests consistent à observer les rangs des valeurs prises par la variable d'étude de deux échantillons.

6. Le terme s'oppose à « hétéroscédascité ».

Si deux échantillons (x_i) avec $i \in [1, n]$ et (y_j) avec $j \in [1, m]$ sont issus de la même population, on doit obtenir, en mélangeant les observations et en les classant par valeurs croissantes, une population homogène.

Après avoir ordonné les suites, on désigne par U le nombre obtenu en comptant le nombre de couples (x_i, y_i) tel que :

1. $x_i > y_i$ si les variables sont quantitatives ;
2. le rang x_i est supérieur au rang de y_j si les variables sont qualitatives.

Si tous les x_i sont inférieurs à tous les y_j, le nombre U varie de 0 à nm dans le cas contraire. Si les deux échantillons sont issus de la même population :

$$\mathbb{E}(U) = \frac{nm}{2} \tag{4.201}$$

et

$$\mathbb{V}(U) = \frac{nm(n+m+1)}{12} \tag{4.202}$$

Si les effectifs n et m des échantillons sont supérieurs à huit, la loi de U tend asymptotiquement vers une loi normale ayant pour paramètres les valeurs $\mathbb{E}(U)$ et $\mathbb{V}(U)$ définies précédemment.

On rejettera H_0 « échantillons issus d'une même population » si la valeur observée de U est trop grande.

On propose un calcul plus rapide. Après avoir classé les observations, on calcule la somme des rangs des individus d'un des groupes, le groupe X par exemple. Soit $W_x = U + \frac{n(n-1)}{2}$ cette somme a pour espérance et variance :

$$\mathbb{E}(W_x) = \frac{n(n+m+1)}{2} \tag{4.203}$$

et

$$\mathbb{V}(W_x) = \frac{nm(n+m+1)}{12} \tag{4.204}$$

La somme des rangs des éléments de l'une ou l'autre des deux distributions comparées suit, sous H_0, une distribution tabulée pour des petits échantillons et une loi normale pour des échantillons de taille supérieure à 20. Il en est de même pour le test de Mann-Whitney.

Si les effectifs n et m des échantillons sont supérieurs à huit, la loi W_x tend asymptotiquement vers une loi normale dont les paramètres sont $\mathbb{E}(W_x)$ et $\mathbb{V}(W_x)$.

On rejette l'hypothèse H_0 « échantillons issus d'une même population », si la valeur calculée de W_x est trop grande pour le seuil de confiance choisi.

Remarque. On aurait pu appliquer le test de Fisher d'égalité des variances, puis le test de Student de comparaison des moyennes, si les conditions d'application de ces tests étaient vérifiées, à savoir une population normale (ou d'échantillons de tailles suffisantes).

Tests de comparaison – Test de comparaison de deux échantillons appariés

On considère un échantillon d'individus soumis à deux mesures successives d'une même variable. On se pose la question : « Les deux séries de valeurs sont-elles semblables ? ».

Soient X et Y les variables parentes associées à chaque série, ces variables sont indépendantes et suivent des lois normales.

On teste seulement l'égalité des moyennes $\mu_x = \mu_y$ avec la variable aléatoire $X - Y$ qui suit une loi normale d'espérance $\mu_x - \mu_y$.

On calcule $\bar{d}$ en calculant les écarts entre les deux variables X et Y.

$$x_1 - y_1 = d_1 x_2 - y_2 = d_2 \dots x_n - y_n = d_n \tag{4.205}$$

d'où $\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_n$ qui est la moyenne des différences.

On pose les hypothèses suivantes :

$$\begin{cases} H_0 : \mu_x = \mu_y \\ H_1 : \mu_x \neq \mu_y \end{cases} \tag{4.206}$$

Comme on ne connaît pas, en général, la variance estimée σ^2, on fait un test de Student sur la moyenne des différences :

$$T\,(n-1) = \frac{\bar{d}}{\frac{s_{\bar{D}}}{\sqrt{n-1}}} \tag{4.207}$$

On rejette H_0 si $|T| > k$, la valeur critique k dépend du seuil α choisi.

Tests de comparaison – Test de comparaison de plusieurs échantillons : test d'homogénéité du χ^2

On dispose de k échantillons, décrits par une variable aléatoire qualitative prenant r modalités.

Un tableau des observations est dressé (Tab. 4.10) :
avec n_{ij} l'effectif de l'échantillon i prenant la modalité j, n_{ij} l'effectif total de l'échantillon i avec $n_{i.} = \sum_{j=1}^{r} n_{ij}$, n_{ij} le nombre total des individus possédant le caractère j avec $n_{.j} = \sum_{i=1}^{k} n_{ij}$, et N le nombre total d'observations avec $N = \sum_{i=1}^{k} \sum_{j=1}^{r} n_{ij} = \sum_{j=1}^{r} n_{.j} = \sum_{i=1}^{k} n_{i.}$.

On pose les hypothèses suivantes :

	Modalité 1	...	Modalité j	...	Modalité r	**Total**
Échantillon 1	n_{11}	...	n_{1j}	...	n_{1r}	$n_{1.}$
...	...	...	...	...	...	...
Échantillon i	n_{i1}	...	n_{ij}	...	n_{rr}	$n_{i.}$
...	...	...	...	...	...	...
Échantillon k	n_{k1}	...	n_{kj}	...	n_{kr}	$n_{k.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.r}$	N

TABLE 4.10 – Tableaux des observations

$$\begin{cases} H_0 : \text{Les échantillons sont issus de la même population.} \\ H_1 : \text{Les échantillons sont issus de populations différentes.} \end{cases} \tag{4.208}$$

Sous H_0, on désigne par p_j, la probabilité théorique, mais inconnue, de posséder la modalité j. Si cette probabilité était connue, il serait possible de comparer les effectifs observés n_{ij} aux effectifs espérés $p_j n_{i.}$ pour toutes les valeurs des indices i et j. La statistique d^2

$$d^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - p_j n_{i.})^2}{p_j n_{i.}} \tag{4.209}$$

est une mesure de la distance entre une distribution théorique et la distribution observée. Sous H_0, cette variable est la réalisation d'une variable aléatoire D^2 suivant une loi du χ^2 à υ degrés de liberté. Le tableau des observations contient kr termes, liés par k relations. La variable aléatoire D^2 est alors une variable χ^2 à $(kr - k)$ degrés de liberté.

Toutefois, les probabilités p_j ne sont pas connues, mais estimées, par les quantités :

$$\hat{p}_j = \frac{n_{.j}}{N} \tag{4.210}$$

Il en résulte une nouvelle expression pour la statistique d^2 :

$$d^2 = N \left(\sum_{i=1}^{k} \sum_{j=1}^{r} \frac{{n_{ij}}^2}{n_{i.} n_{.j}} - 1 \right) \tag{4.211}$$

obtenue en remplaçant les probabilités p_j par leurs estimations.

Avec les fréquences relatives au lieu des fréquences absolues, on obtient :

$$d^2 = N\left(\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{{f_{ij}}^2}{f_{i.}f_{.j}} - 1\right) \tag{4.212}$$

Les r estimations des probabilités p_j sont liées par une relation – leur somme est égale à 1. En fait, on a estimé $(r-1)$ paramètres indépendants. La statistique D^2 est alors une variable aléatoire χ^2 à $(kr-k-r+1) = (k-1)(r-1)$ degrés de liberté.

On rejette H_0 si la valeur observée d^2 est trop grande pour un seuil α donné.

Tests de comparaison – Test d'homogénéité de deux proportions observées

On considère des échantillons de grandes tailles. Soient n_1 et n_2 les tailles de ces échantillons, f_1 et f_2 les pourcentages des individus présentants un certain caractère dans chaque échantillon, et soient p_1 et p_2 les probabilités correspondantes, on veut savoir si les probabilités p_1 et p_2 diffèrent significativement ou non à partir des pourcentages observés. Dans le cas contraire, les différences s'expliqueraient par le hasard des fluctuations d'échantillonnage. On teste ici l'éventuelle liaison entre deux variables qualitatives $VQ1$ et $VQ2$ chacune à deux modalités $VQ1 =$ valeur 1 et $VQ2 =$ valeur 2. Chaque proportion est rattachée à une valeur de leur population mère respective ϖ_1 et ϖ_2, mais elle est inconnue.

On formule les hypothèses suivantes :

$$\begin{cases} H_0 : \varpi_1 = \varpi_2 = \varpi \\ H_1 : \varpi_1 \neq \varpi_2 \end{cases} \tag{4.213}$$

N.B. H_1 est un test bilatéral, mais il est possible de poser $H_1 : \varpi_1 > \varpi_2$, c'est-à-dire un est unilatéral à droite, ou encore $H_1 : \varpi_1 < \varpi_2$, c'est-à-dire un est unilatéral à gauche.

Les échantillons étant de grande taille, les pourcentages observés f_1 et f_2 peuvent être considérés comme des réalisations de variables aléatoires F_1 et F_2 suivant des lois normales :

$$\begin{cases} F_1 \sim N\left(p, \sqrt{\frac{p(1-p)}{n_1}}\right) \\ F_2 \sim N\left(p, \sqrt{\frac{p(1-p)}{n_2}}\right) \end{cases} \tag{4.214}$$

Leur différence $F_1 - F_2$ suit alors la loi normale centrée réduite :

$$F_1 - F_2 \sim N\left(0, \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right) \tag{4.215}$$

On rejette H_0 si, au risque $\alpha = 0,05$ par exemple, on observe :

$$|f_1 - f_2| > 1,96\sqrt{p\,(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{4.216}$$

avec $t_{0,95} = 1,96$.

Remarque 1. La probabilité p étant inconnue, on prend comme valeur son estimation calculée à partir des deux échantillons :

$$\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} \tag{4.217}$$

Remarque 2. On aurait pu utiliser le test du χ^2 dont le calcul est simple dans ce cas particulier. En notant a, b, c, d, les effectifs observés dans chaque échantillon et pour chaque modalité. On obtient alors le tableau 4.11.

	Modalité 1	Modalité 2	**Total**
Échantillon 1	a	b	$a+b$
Échantillon 2	c	d	$c+d$
Total	$a+c$	$b+d$	$N = a+b+c+d$

TABLE 4.11 – Tableau des observations des deux échantillons

Pour mettre en œuvre le test du χ^2, on calcule la statistique D^2 qui est égale à :

$$\begin{aligned} D^2 &= N\left[\tfrac{a^2}{(a+b)(a+c)} + \tfrac{b^2}{(b+a)(b+d)} + \tfrac{c^2}{(c+a)(c+d)} + \tfrac{d^2}{(d+b)(d+c)} - 1\right] \\ D^2 &= N\tfrac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)} \end{aligned} \tag{4.218}$$

La statistique D^2 suit une loi du χ^2 à un degré de liberté.

Exemple On teste des micro-pipettes ayant besoin d'une révision au bout de six mois avec deux populations suivant des procédures de fabrication différentes. Soient $VQ1$ = fabrication 1 et $VQ2$ = fabrication 2, on obtient respectivement les paramètres suivants :

$$\begin{aligned} p_1 &= 0,45 \\ n_1 &= 52 \end{aligned} \tag{4.219}$$

et

$$\begin{aligned} p_2 &= 0,38 \\ n_2 &= 67 \end{aligned} \tag{4.220}$$

L'état des micro-pipettes au bout de six mois dépendt-t-il de leur mode de fabrication ?

Sous H_0, $\varpi_0 = \varpi_1 = \varpi_C$ avec ϖ_C la probabilité commune, dont on peut obtenir une estimation ponctuelle p_C telle que $p_C = \hat{\varpi}_C = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ d'estimateur $\frac{n_1 P_{n_1} + n_2 P_{n_2}}{n_1 + n_2}$ avec P_{n_1} et P_{n_2} des variables d'échantillonnage.

La statistique de test est basée sur l'étude des fluctuations d'échantillonage de différence $P_{n_1} - P_{n_2}$ sous H_0.

$$\begin{array}{l} P_{n_1} \sim \beta\left(n_1, \hat{\varpi}_C\right) \approx N\left(\hat{\varpi}_C, \sqrt{\frac{\hat{\varpi}_C(1-\hat{\varpi}_C)}{n_1}}\right) \\ P_{n_2} \sim \beta\left(n_2, \hat{\varpi}_C\right) \approx N\left(\hat{\varpi}_C, \sqrt{\frac{\hat{\varpi}_C(1-\hat{\varpi}_C)}{n_2}}\right) \end{array} \tag{4.221}$$

La variable $P_{n_1} - P_{n_2}$ subit des fluctuations approchables par une loi normale d'espérance nulle.

$$\begin{array}{l} \mathbb{E}\left(P_{n_1} - P_{n_2}\right) = \mathbb{E}\left(P_{n_1}\right) - \mathbb{E}\left(P_{n_2}\right) = \hat{\varpi}_C - \hat{\varpi}_C = 0 \\ \mathbb{V}\left(P_{n_1} - P_{n_2}\right) = \mathbb{V}\left(P_{n_1}\right) - \mathbb{V}\left(P_{n_2}\right) = \hat{\varpi}_C\left(1 - \hat{\varpi}_C\right)\left[\frac{1}{n_1} + \frac{1}{n_2}\right] \text{ (Les variables étant indépendante} \end{array} \tag{4.222}$$

On arrive à considérer la statistique de test normalisée :

$$Z = \frac{\left(P_{n_1} - P_{n_2}\right) - 0}{\sqrt{\hat{\varpi}_C\left(1 - \hat{\varpi}_C\right)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \tag{4.223}$$

La valeur observée de la statistique de test est :

$$z_{obs} = \frac{\left(p_1 - p_2\right) - 0}{\sqrt{p_C\left(1 - p_C\right)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \tag{4.224}$$

Les conditions d'application du test sont respectées :

$$\begin{array}{l} n_1 \geq 30 \\ n_2 \geq 30 \\ \min\left(n_1, n_2\right) \times \min\left(p_C, 1 - p_C\right) > 5 \end{array} \tag{4.225}$$

Sous H_1, on a soit $\varpi_2 > \varpi_1$ ou $\varpi_2 < \varpi_1$

$$\begin{array}{l} p_C = \frac{52 \times 0{,}45 + 67 \times 0{,}38}{52 + 67} \approx 0,411 \\ z_{obs} \approx 0,77 \end{array} \tag{4.226}$$

donc avec le risque $\alpha = 5\ \%$, $z_\alpha = 1,65$. Comme $z_{obs} < 1,65$, $p_{value} > 5\ \%$.

En conclusion, H_0 ne peut être rejetée. Les échantillons n'ont pas permis de mettre en évidence une liaison statistique entre la nécessité d'une révision des micro-pipettes au bout de six mois et leur mode de fabrication.

Analyse de la variance

L'analyse de la variance[7] permet de généraliser le test de comparaison de plusieurs échantillons au problème suivant : **la comparaison des moyennes de plusieurs échantillons indépendants**. Ainsi, comme son nom ne l'indique pas, l'analyse de la variance permet de **comparer des moyennes** – c'est la méthode à utiliser lorsqu'il faut comparer plusieurs moyennes (à partir de trois). Il faut ajouter que le procédé qui consiste à tester l'égalité des moyennes de chaque couple n'est pas satisfaisant. Aussi, la nécessité d'une procédure permettant de **tester globalement** l'ensemble de tous les échantillons est fournie par la **théorie de l'analyse de la variance**. Le but de cette théorie est d'étudier la variabilité d'un objet en fonction d'un ensemble de facteurs que l'on peut contrôler systématiquement, et que l'on souhaite dissocier la part revenant à chaque facteur.

L'An.O.Va. permet d'étudier la dépendance d'une variable quantitative à une ou deux variables qualitatives. Plus généralement, les variables qualitatives sont appelées **facteurs**. Le facteur contrôlé peut intervenir dans des conditions qui diffèrent : 1. soit par leur nature, 2. soit par leur intensité. De plus, le facteur contrôlé peut être : 1. soit à effets fixes, 2. soit à effets aléatoires.

La variable dépendante (V.D.) est une variable quantitative continue. Les variables indépendantes (V.I.) correspondent aux facteurs. De fait, l'étude d'un facteur est une analyse bivariée, tandis que, avec au moins deux facteurs, l'analyse est multivariée. L'An.O.Va. établit si la dépendance étudiée est significative pour le facteur considéré. Pour y répondre, il faut tester si la moyenne de la variable quantitative d'étude est homogène sur l'ensemble des modalités de la variable qualitative. Il faut rejeter l'hypothèse nulle H_0 d'égalité des moyennes par l'analyse de la variance. Le test utilisé est le test F de Fisher consistant à comparer la variance inter-échantillon à la variance intra-échantillon. On tente d'expliquer la **cause** de la diversité des informations par l'analyse de leur variance.

Il faut noter que l'analyse de la variance n'est valable en toute rigueur que pour des **échantillons tirés de populations normales et de même variance**. En général, le non-respect de ces conditions n'a pas trop d'influence sur la validité du test. Dit autrement, l'analyse de la variance est une **méthode robuste**. L'erreur introduite est toutefois d'autant plus forte que les effectifs des échantillons sont faibles et inégaux.

On distingue :

7. Analyse of Variance (An.O.Va.)

1. l'**analyse de la variance à simple entrée**. Un seul facteur est contrôlé, les autres facteurs étant regroupés sous le nom « facteurs non contrôlés » ;
2. l'**analyse de la variance à double entrée**. Elle étudie l'action simultanée de deux facteurs contrôlés, chacun agissant individuellement avec une possibilité d'interaction entre les deux ;
3. l'**analyse de la variance à entrées multiples** avec plusieurs facteurs contrôlés.

Remarque. Les points (2) et (3) sont des analyses multidimensionnelles.

4.6.7 Les tests d'indépendance vers les relations multivariées

Les tests d'indépendance sont simplement mentionnés ici, car ils interviennent dans le cas de deux ou de plusieurs variables aléatoires. Par exemple, pour tester l'indépendance entre deux variables aléatoires X et Y, par l'examen d'échantillons de taille n, on étudie différentes mesures de liaison dépendant de la nature des deux variables. Sont-elles qualitatives ou quantitatives ? Un chapitre concernant les statistiques d'ordre établira les premiers tests d'indépendance pour les variables qualitatives ordinales, tandis que le chapitre concernant la relation entre deux variables quantitatives établira les tests d'indépendance pour les variables quantitatives.

Test de Mann et de Whitney (1947)

Henry Mann (1905-2000)

Donald Ransom Whitney (1915-2007)

On dispose de deux échantillons, indépendants et non exhaustifs, E_1 et E_2, de tailles respectives n_1 et n_2. On veut comparer les deux moyennes expérimentales, c'est-à-dire tester l'hypothèse nulle $H_0 : \mu_1 = \mu_2$.

Le test est utilisé dans les cas de non normalité et si les échantillons sont de petites tailles. Il ne requiert d'aucune hypothèse et d'aucun paramètre exact.

La méthode consiste à remplacer les valeurs prises par la variable par les rangs de ces valeurs. L'échelle ordinale créée est celle qui est utilisée dans le test. Cette approche montre bien que les tests non paramétriques sont **universels**. Le test de Mann-Whitney est un test non paramétrique que l'on peut appliquer aux variables quantitatives et aux variables qualitatives.

Le test propose de remplacer les valeurs prises par la variable étudiée par les rangs de ces valeurs. De fait, avant de procéder au test en lui-même, il faut :

1. classer par ordre croissant l'ensemble des valeurs de deux échantillons en repérant l'origine de chaque valeur ;
2. affecter à chaque valeur de $E_1 \cup E_2$, son rang dans ce classement. S'il y a des scores *ex-aequo*, on attribue à chacun un rang égal à la moyenne des rangs qu'ils occupent ;

3. compter, pour tout élément x_i de E_1, le nombre d'éléments de E_2 situés après x_i ;
4. noter m_1 la somme de toutes les valeurs ainsi associées à tous les éléments de E_1 ;
5. définir de même m_2 en permutant les rôles de E_1 et de E_2 ;
6. poser $m = \min(m_1, m_2)$, c'est-à-dire que m est la plus petite des deux valeurs m_1 et m_2 obtenues.

On peut obtenir m_1 et m_2 de la façon suivante : soient r_1 et r_2 la somme des rangs des valeurs de chacun des deux échantillons. En cas de scores *ex-aequo*, les rangs sont déterminés comme indiqué ci-après :

$$m_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - r_1 \tag{4.227}$$

et

$$m_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - r_2 \tag{4.228}$$

Le test peut être construit. Soit M la variable aléatoire qui prend la valeur m à l'issue de l'expérience aléatoire. Les tables donnent, en fonction de n_1, n_2 et α la valeur m_α, telle que, sous H_0, $\Pr(M \leq m_\alpha) = \alpha$, dans les cas $\alpha = 0,05$ et $\alpha = 0,01$. On rejette alors l'hypothèse nulle si $m \leq m_\alpha$. Si n_1 et n_2 sont hors des tables, alors, si H_0 est vraie, H suit approximativement la loi normale $N(\mu, \sigma)$ avec :

$$\mu = \frac{n_1 n_2}{2} \tag{4.229}$$

et

$$\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{4.230}$$

On calcule la valeur de la variable normale réduite $u = \frac{m-\mu}{\sigma}$, et on conclut avec la table *ad hoc*, c'est-à-dire que l'on rejette H_0 si $|u| > u_\alpha$.

Exemple On s'intéresse à la surveillance d'une plage ayant subi une récente pollution à l'hydrocarbure. Un prélèvement de huit coques au hasard sur cette plage a permis de mesurer au laboratoire la concentration du polluant contenu dans leur chair en mg/kg de poids sec. Dix jours plus tard, un second échantillon de neuf coques a été collecté au hasard sur la plage pour évaluer le taux de polluant que leur chair renfermait. Les résultats sont reportés en mg/kg (Tab. 4.12).

Prélèvement 1	39	41	39	33	35	33	40	38	
Prélèvement 2	35	26	40	27	29	31	32	30	35

TABLE 4.12 – Prélèvements des coques sur la plage

En dix jours, la quantité de polluant a-t-elle significativement diminué sur la plage ?

La variable d'étude est quantitative et continue. Elle suit une distribution de nature inconnue. Les deux séries de valeurs indépendantes constituent des échantillons de petits tailles de populations mères inconnues P_1 et P_2.

Le test de Mann-Whitney permet de répondre à la question. Il s'intéresse à l'enchevêtrement des observations réunies des deux échantillons afin de savoir si la répartition des valeurs est homogène. Pour cela, on compare les médianes des deux distributions. Le test est unilatéral.

Sous H_0, les différences constatées dans les deux échantillons ne sont pas significatives. Elles peuvent être imputées au hasard des fluctuations d'échantillonnage. Sous H_1, les différences sont significatives.

$$\begin{aligned} H_0 &: m_{e_1} = m_{e_2} \\ H_1 &: m_{e_1} > m_{e_2} \end{aligned} \tag{4.231}$$

La variable de décision U est calculée par chaque échantillon. Pour chaque valeur x_i, on compte le nombre de valeur y_j qui lui sont strictement inférieures. Pour chaque valeur x_i, on compte le nombre de valeur y_j qui lui sont égales, et on le multiplie par $\frac{1}{2}$. U_x correspond à la somme de ces deux nombres. On fait de même avec y en calculant U_y. On considère le minimum entre les deux quantitées U_x et U_y que l'on confronte à une table des valeurs de U de Mann-Whitney pour les tailles des échantillons n_1 et n_2 considérés.

1. Ordonner les valeurs de deux séries et indiquer l'origine de chaque valeur.
2. Calculer U

Pour obtenir, le tableau 4.13, on utilise la **méthode des rangs avec rangs *ex-aequo* corrigés**. Il s'obtient huit étapes :

1. ordonner les valeurs des deux séries réunies ;
2. indiquer l'origine de chaque valeur. Par exemple, avec l'origine P2, au rang n° 15, la valeur mesurée est 40. On compare cette valeur à P1. On constate qu'il y a 1 valeur identique (le rang n° 16) et 6 valeurs inférieures (rang n° 7 : 33, rang n° 8 : 33, rang n° 11 : 35, rang n° 12 : 38, rang n° 13 : 39, rang n° 14 : 39), donc U_x vaut :

$$U_x = \frac{1}{2} \times 1 + 6 = 6,5 \tag{4.232}$$

3. créer deux colonnes destinées à recevoir les rangs pour chacune des séries ;
4. pour les rangs *ex-aequo*, considérer le rang moyen. Par exemple :

$$\begin{array}{l}\text{Pour } 33 \rightarrow \frac{7+8}{2} = 7,5 \\ \text{Pour } 35 \rightarrow \frac{9+10+11}{2} = 7,5\end{array} \tag{4.233}$$

5. réaliser la somme des rangs pour chacune des séries ;
6. retrancher à chacune des sommes le minimum des rangs $\frac{n_i(n_i+1)}{2}$ avec $i = \{1, 2\}$ pour obtenir la statistique U pour chaque série ;
7. retenir la valeur minimale des deux U obtenus $U_{\min}$;
8. comparer $U_{\min}$ avec la valeur seuil au risque α, U_{seuil}. H_0 est rejetée si $U_{\min} < U_{\text{seuil}}$.

Le résultat pour notre analyse est le suivant. Une table de Mann-Whitney est proposée pour chaque valeur de α. On choisit $\alpha = 10$ % si le test est bilatéral, et $\alpha = 5$ % si le test est unilatéral. La table se lit en croisant la taille des deux échantillons en ligne et en colonne – le sens n'ayant aucune importante puisque la matrice des valeurs est symétrique. Ici, $U_{\text{seuil}} = 18$, donc $U_{\min} < U_{\text{seuil}}$ $(15, 5 < 18)$. H_0 est rejeté, donc les différences sont significatives au risque $\alpha = 5$ %. Pour conclure, en dix jours, la quantité de polluant a significativement diminué sur la plage au risque $\alpha = 5$ %.

Par contre, pour $\alpha = 1$ %, la valeur seuil est $U_{\text{seuil}} = 11$, donc $U_{\min} > U_{\text{seuil}}$ $(11, 5 > 11)$. H_0 est vraie, donc les différences ne sont pas significatives au risque $\alpha = 1$ %. Pour conclure, en dix jours, la quantité de polluant n'a pas significativement diminué sur la plage au risque $\alpha = 1$ %.

Test de Wilcoxon

On dispose de deux échantillons appariés, c'est-à-dire que chaque valeur d'un échantillon est associée à une valeur de l'autre échantillon. Ils sont par conséquent de même taille. H_0 est l'égalité des moyennes des deux populations soit $\mu_1 = \mu_2$.

Avant de procéder au test en lui-même, il faut :

1. calculer les différences entre les valeurs appariées, supprimer les différences nulles et noter n le nombre de différences non nulles ;
2. classer ces différences par ordre croissant des valeurs absolues ;
3. affecter à chaque différence son rang dans ce classement. S'il y a des scores *ex-aequo*, on attribue à chacun un rang égal à la moyenne des rangs qu'ils occupent ;
4. calculer w_+ la somme des rangs des différences positives et w_- la somme des rangs des différences négatives.

Rang	Valeur mesurée ordonnée	Origine	Points	Rang du prélèvement 1 (P1)	Rang du prélèvement 2 (P2)
1	26	P2	0		1
2	27	P2	0		2
3	29	P2	0		3
4	30	P2	0		4
5	31	P2	0		5
6	32	P2	0		6
7	33	P1		7,5	
8	33	P1		7,5	
9	35	P2	2,5		10
10	35	P2	2,5		10
11	35	P1		10	
12	38	P1		12	
13	39	P1		13,5	
14	39	P1		13,5	
15	40	P2	6,5		15,5
16	40	P1		15,5	
17	41	P1		17	
Total			11,5	96,5	56,5
			Moyenne	8	9
			$\frac{n(n+1)}{2}$	36	45
			$U_{\min}$		11,5

TABLE 4.13 – Résultat de l'analyse de Mann-Whitney

On note $w = \min(w_-, w_+)$ la plus petite des valeurs w_- et w_+.

Le test peut être construit. Soit W variable aléatoire qui prend la valeur w à l'issue de l'expérience aléatoire.

— Pour $n \leq 25$, la table *ad hoc* donne, en fonction de n, la valeur w_α, telle que, sous H_0, $\Pr(W \leq w_\alpha) = \alpha$ dans les cas $\alpha = 0,05$ et $\alpha = 0,01$. On rejette l'hypothèse nulle si $w \leq w_\alpha$.

— Pour $n > 25$, lorsque H_0 est vraie, W suit approximativement la loi normale $N(\mu, \sigma)$ avec :

$$\mu = \frac{n(n-1)}{4} \tag{4.234}$$

et

$$\sigma = \sqrt{\frac{n\,(n+1)\,(2n+1)}{24}} \tag{4.235}$$

On calcule alors la valeur de la variable normale réduite : $u = \frac{w-\mu}{\sigma}$ et on conclut que l'on rejette H_0 si $|u| > u_\alpha$.

Test de Kruskal et Wallis

William Henry Kruskal (1919-2005)
Wilson Allen Wallis (1912-1998)

On dispose de k échantillons, indépendants et non exhaustifs, $E_1, \ldots, E_k$, de tailles respectives $n_1, \ldots, n_k$. On veut comparer globalement les k moyennes expérimentales, c'est-à-dire tester l'hypothèse nulle $H_0 : \mu_1 = \ldots = \mu_k$.

Avant de procéder au test en lui-même, il faut :

1. classer par ordre croissant l'ensemble des valeurs de ces k échantillons ;
2. détermine le rang de chaque valeur, de la même manière que les tests de Mann et de Kruskal, s'il existe des scores *ex-aequo* ;
3. noter, pour chaque échantillon E_i, r_i la somme des rangs des valeurs de cet échantillon.

Il est alors possible de calculer la quantité k :

$$k = \frac{12}{n\,(n-1)} \left(\sum_{i=1}^{k} \frac{{r_i}^2}{n_i} \right) - 3\,(n+1) \tag{4.236}$$

avec $n = \sum_{i=1}^{k} n_i$ l'effectif total.

Le test peut être construit. Soit H la variable aléatoire qui prend la valeur h à l'issue de l'expérience aléatoire.

— Si les n_i sont assez grands ($n_i > 5$ pour tout i), alors si H_0 est vraie, H suit à peu près une loi du χ^2 à $\upsilon = k - 1$ degrés de liberté. Dans la table, on lit la valeur $\chi^2\,(\alpha)$ telle que $\Pr\,(H \geq \chi^2\,(\alpha)) = \alpha$ et on rejette H_0 si $h \geq {\chi^2}_c$.

— Si les n_i ne sont pas assez grands, on dispose de tables qui donnent la valeur h_α, telle que $\Pr\,(h_\alpha \geq \chi^2\,(\alpha)) = \alpha$. On rejette H_0 si on obtient $h \geq h_\alpha$.

La table *ad hoc* donne h_α, pour $\alpha = 0,05$ et $\alpha = 0,01$, dans le cas de trois échantillons de tailles inférieures ou égales à 5.

Pour conclure, il faut rappeler que toutes les tables donnent la valeur des probabilités pour des **quantiles**.

4.7 Intervalles de confiance ou tests statistiques

Les tests statistiques suscitent de nombreuses controverses [Poinsot, 2004]. En général, ils sont mal utilisés et mal conçus. On dénombre sept critiques.

1. Les hypothèses H_0 du type « aucun effet » sont fausses dès le départ.
2. Il suffit d'un échantillon suffisamment grand pour montrer que n'importe quoi est statistiquement significatif.
3. Avec un échantillon suffisamment petit, on peut obtenir au contraire un résultat non significatif sur n'importe quoi, par simple manque de puissance du test.
4. Le fait que l'hypothèse H_0 ne soit pas rejetée est trop souvent abusivement interprété comme une confirmation (au moins implicite de l'hypothèse H_0).
5. Lorsque l'hypothèse H_0 est rejetée, beaucoup de chercheurs confondent la probabilité ρ du test avec la probabilité que l'hypothèse H_0 soit vraie.
6. La manière dont les résultats sont présentés dans les articles scientifiques rend souvent les méta-analyses (la synthèse entre plusieurs études, en combinant leurs résultats pour gagner en précision) très difficiles.
7. Les scientifiques donnent l'impression de s'hypnotiser sur la significativité statistique de leurs tests.

Mieux vaut un bon intervalle de confiance qu'un test bancal.

Bibliographie

[Poinsot, 2004] POINSOT, D. (2004). Statistiques pour statophobes. Une introduction au monde des tests statistiques à l'intention des étudiants qui n'y entravent que pouic et qui détestent les maths par-dessus le marché. Hyperlink[https ://perso.univ-rennes1.fr/denis.poinsot, https ://perso.univ-rennes1.fr/denis.poinsot].