Introduction to statistical and probabilistic method

Keita, Moussa

February 2016

Introduction à la méthode statistique et probabiliste

Par
Moussa Keita, PhD*
Février 2016
(Version 1)

*Ecole d'Economie, Université d'Auvergne Clermont Ferrand 1 ;
Ecole Nationale Supérieure de Statistique et d'Economie Appliquée ENSEA-C.I

Contact info: Email : keitam09@ymail.com

Codes JEL: C1

Mots clés: statistique, probabilités, estimateurs, sondage.

AVANT-PROPOS

Ce manuscrit propose une introduction à la méthode statistique et probabiliste. Il est structuré autour de trois grands thèmes développés en cinq chapitres. La première partie est une introduction aux calculs de probabilités dans laquelle nous abordons les notions de probabilités élémentaires, d'espaces probabilisés, de variables aléatoires, de fonctions de densité et de fonctions de répartition. Nous passons également en revue les principales lois de probabilités qui caractérisent les phénomènes courants de l'observation. La seconde partie est une introduction aux démarches d'analyses statistiques proprement dites. Dans cette partie, nous présentons, d'une part, les méthodes d'analyses statistiques à visées descriptives et exploratoires et d'autre part les méthodes d'analyses à visées inférentielles. Les notions abordées dans cette partie sont notamment celles de distributions statistiques, de liaisons entre variables mais aussi les notions d'estimateurs et de tests d'hypothèses. Quant à la troisième partie, elle propose une introduction aux techniques de sondage centrée sur l'étude des protocoles d'échantillonnage et des méthodes d'estimation à des fins d'extrapolation.

Le document étant toujours en cours de progrès, nous restons ouverts à toutes les critiques et suggestions de nature à améliorer son contenu*

* A ce propos, nous tenons ici à remercier les enseignants dont les notes de cours ont été d'un apport significatif à la rédaction de ce document. Il s'agit notamment de Phan T., de Aragon Y., Goga-Cardot C. et Ruiz-Gazen A. et de Desgraupes B.

Table des matières

Introduction générale

Le *déluge* d'information, né de la récente la révolution technologique, présenté sous le vocable de « Big Data » a mis à rude épreuve les approches classiques d'études statistiques fondées sur les données « structurées » prenant ainsi au dépourvu de nombreux statisticiens de notre temps. Désormais, la statistique évolue vers une science des données numériques où l'informatique joue un rôle de premier plan pour assurer une fouille automatique dans "l'océan de données non structurées", pour la visualisation des résultats et d'apprentissage automatique (machine-learning).C'est d'ailleurs à ce titre que le statisticien du 21ième siècle (capable de manier à la fois algorithmes avancés de programmation, outils mathématiques et concepts statistiques) été baptisé « Data Scientist ».

Cependant, ce nouveau contexte informationnel n'a nullement réussi à ébranler les fondements premiers de *la méthode statistique*. Plus que jamais la statistique reste "La Mecque" de l'empirisme où viennent recevoir leur bénédiction toutes les disciplines qui considèrent que les seuls critères de vérités scientifiques sont l'expérience et l'observation. Sur ce plan, plus personne ne semble échapper à la *loi statistique* ; de la Psychologie, au Marketing en passant par l'Economie ou même dans les tests de reconnaissance de voix ou dans les analyses d'imageries médicales.

Généralités

Selon une définition communément admise, la statistique est présentée comme l'ensemble des méthodes ayant pour objet la collecte, le traitement, l'analyse et l'interprétation de données d'observation sur un groupe d'unités.

Malgré sa simplicité cette définition résume assez bien l'objet de la statistique et permet de donner un aperçu général sur sa méthode. En effet, le but fondamental de la statistique reste le traitement et l'exploitation de l'information soit pour faire avancer l'état de la connaissance sur un phénomène d'observation considéré comme aléatoire ou pour éclairer les prises de décisions. Pour ce fait, elle est utile à la fois d'un point de vue micro (agents individuels) mais aussi d'un point de vue macro (ensemble de la société). Selon Pierre Dagnelie, un statisticien contemporain reconnu, les applications de la statistique se distinguent en trois grands axes: la statistique administrative ou « gouvernementale » assurée par les instituts de statistique ou les services statistiques nationaux à propos de grands ensembles de données ; la statistique « mathématique » ou « universitaire » faite avec peu de données et qui a pour but la novation de la méthode ; et la statistique « appliquée » ou la statistique « de terrain » faite notamment dans les instituts de sondage, dans le cabinets d'études, dans les entreprises etc… appliqués à des cas concrets de la réalité courante.

La statistique étant une discipline à part entière, elle dispose d'un vocabulaire et d'un jargon technique propre. Pour étayer les propos prenons un simple concret sur un phénomène statistique ordinaire décrit comme suit. Le Maire d'une petite commune souhaite étudier l'opportunité de la mise en circulation d'un nouveau bus de ramasse scolaire au profit des élèves de l'unique lycée de sa ville. Il compte alors se baser sur une étude statistique pour éclairer sa décision. Une petite enquête montre que : 46 élèves utilisent un deux-roues, 284 élèves utilisent les transports en commun, 163 élèves se déplacent à pied, 92 élèves sont déposés par leurs parents.

Pour analyser statistiquement cette situation, on utilise généralement des mots clés ou vocabulaire. Il s'agit notamment : population, individu, effectif, caractère ou variable.

On appelle **population** l'ensemble des sujets (**individus**) sur lesquels porte une étude statistique. La dénomination « population » est provient des premières applications de la statistique notamment la démographie. Dans l'exemple ci-dessus, la population statistique est l'ensemble des élèves du lycée. Les éléments de la population s'appellent **individus**. Il faut noter que les individus ne sont pas nécessairement des personnes. Ils peuvent être des sujets de toute nature (ensemble des habitations d'une ville, ensemble des espèces animales peuplant un parc zoologique, etc…).

L'étude de **tous les individus** d'une population s'appelle **recensement**. En revanche, une étude portant que sur une partie de la population est qualifiée de **sondage** ; la partie de la population choisie étant alors appelée échantillon.

On appelle **effectif total** de la population le nombre d'éléments de l'ensemble de cette population. Dans l'exemple ci-dessus, l'effectif total est 585 élèves

On appelle **variable statistique ou caractère**, la chose que l'on étudie et qui est commune à tous les individus de la population. Dans cet exemple, la variable statistique étudiée est le mode de transport utilisé.

On appelle **effectif associé** à une valeur de la variable, le nombre de fois où cette valeur apparaît parmi les individus étudiés. Dans cet exemple, l'effectif des individus qui utilisent les transports en commun est 284. L'effectif associé à la modalité transport en commun est 284.

L'assemblage (dans une table) de l'ensemble des valeurs renseignées pour un caractère forme une **série statistique**.

Une variable est dite **quantitative** si elle est représentée par un nombre (sur une échelle numérique de valeurs). Par exemple : âge, distance, durée, etc..

On distingue les variables quantitatives **discrètes** et les variables quantitatives **continues**. Une variable quantitative est dite discrète si elle ne prend que des valeurs isolées ; pas de valeurs décimales. Par exemple, le nombre d'enfants dans une famille, etc. Une variable quantitative est dite continue si elle peut prendre toutes les valeurs comprises entre 2 nombres. Elle autorise donc des valeurs décimales.

A la différence de la variable quantitative, une variable **qualitative** rend compte de la « qualité » du caractère étudié. Elle concerne par exemples que le sexe (homme, femme), ou encore la situation matrimoniale (célibataire, marié, divorcé, veuf) qui sont des caractères qu'on ne peut pas mesurer sur une échelle de valeurs numériques. Ces variables s'expriment donc en **modalités** c'est-à-dire des choix de réponses.

On distingue des variables qualitatives **nominales** et des variables qualitatives **ordinales**. Les variables qualitatives nominales sont des variables dont les modalités servent à qualifier les modalités sans établir un ordre hiérarchique. Par exemple, la couleur des yeux (noir, brun, gris, bleu, marron, vert, etc…). Ici, peu importe l'ordre dans lequel on établit cette liste. Quant aux variables qualitatives ordinales, ce sont des variables dont les valeurs traduisent un ordre dans les modalités. Par exemple, le degré de satisfaction des clients par rapport à la consommation d'un produit (très satisfait, satisfait, insatisfait, très insatisfait). Les variables qualitatives ordinales sont très souvent des degrés de satisfaction, d'approbation, etc...

Il faut noter que même si les variables qualitatives ne sont pas numériquement quantifiables, on peut tout de même les attribuer des codes numériques pour traduire les modalités. Par exemple, pour le sexe, on peut coder 1 pour masculin et 2 pour féminin. Pour le degré de satisfaction du client, on peut coder 1 : très satisfait, 2 : satisfait, 3 : insatisfait, 4 : très insatisfait. Même si ces codes sont numériques, il ne s'agira nullement de faire des calculs statistiques comme la moyenne ou des opérations arithmétiques comme la somme ou la différence. Celles-ci sont réservées aux variables quantitatives. Pour les variables qualitatives, on se limitera, le plus souvent, aux calculs de fréquence de chaque modalité. En reprenant l'exemple du lycée ci-haut, pour la variable moyen de transport, nous avons : « deux-roues », « transports en commun », « à pied », et « déposés par leurs parents ». Il s'agit là ici d'une variable qualitative nominale pour laquelle on ne peut calculer que la fréquence de chaque modalité (nous reviendrons un peu plus tard sur ces détails).

Plan de présentation

Ce document est structuré en cinq chapitres

Le premier chapitre est consacré à la modélisation de probabilités sur les phénomènes statistiques. Dans ce chapitre, nous présentons les théories de base sur les probabilités, nous introduisons la notion de variable aléatoire et nous étudions les principales lois de probabilité.

Le second chapitre est consacré à la présentation des méthodes de statistiques de statistiques descriptives et exploratoires. Dans ce chapitre, il s'agit de présenter les méthodes visant à synthétiser l'information sous forme d'indicateurs statistiques présentés dans des tableaux ou illustrés par de graphiques

Le troisième chapitre aborde les notions de statistique inférentielle en présentant les différentes méthodes d'estimations sur l'échantillon en vue d'une inférence à la population. Dans ce chapitre nous étudions la notion d'estimateurs en présentant les notions d'estimations ponctuelles et d'estimations par intervalles de confiance. Ces notions sont ainsi appliquées aux principaux estimateurs statistiques que sont la moyenne, la variance et la proportion.

Le quatrième chapitre est consacré aux tests d'hypothèses statistiques. C'est une continuité dans la démarche d'inférence statistique. Dans ce chapitre, nous passons en revue les grandes familles de tests statistiques notamment les tests de conformité, les tests de comparaison et les tests d'adéquation

Le cinquième chapitre est un *super-chapitre* qui exploite toutes les connaissances acquises dans les précédents chapitres pour mieux aborder les techniques de sondage. Dans ce chapitre, nous présentons les principales théories relatives à l'échantillonnage, à l'estimation et à l'extrapolation. Nous présentons à cet effet les principaux plans de tirage de l'échantillon notamment les plans de tirage à un degré mais aussi les plans à plusieurs degrés.

Le fait d'aborder l'échantillonnage dans le dernier chapitre peut sembler curieux à certains égards surtout quand on sait que la sélection de l'échantillon est la toute première phase de la chaîne d'une étude statistique. Mais, ce choix ici n'a rien de fortuit. En effet, l'un des grands paradoxes de la statistique c'est que « *pour pouvoir faire de la statistique, il faut d'abord faire de la statistique* ». Le sens de cette boutade est le suivant. Pour mener les analyses statistiques, il faut nécessairement disposer d'un échantillon "représentatif"; c'est-à-dire choisit selon les règles de l'art statistique basées sur des formules censées être utilisées après avoir choisi l'échantillon! Voyez ? Même nous, on s'y perd ! Mais, pas de panique, tout ceci deviendra plus clair lorsque nous atteindrons le chapitre 5.

En somme, on retiendra que le sondage, en général, et l'échantillonnage, en particulier, sont les plus gros consommateurs de formules et de propriétés statistiques et probabilistes. D'où la nécessité, de placer ce chapitre à la suite des autres afin de garder une cohérence dans l'évolution de la compréhension de la méthode statistique.

On peut aussi signaler que dans chaque chapitre, les notions abordées sont appuyées à la fois par des exemples d'applications mais aussi par des exercices pratiques dont les corrigés figurent à la fin de l'énoncé.

Les méthodes d'analyses multidimensionnelles notamment l'analyse factorielle, l'analyse en composantes principales, l'analyse en correspondances multiples, les méthodes de classification et les régressions ont fait l'objet d'un document à part pour des raisons de volumétrie.

Chapitre 1 : Statistiques descriptives et exploratoires

1.1. Analyse descriptive univariée

La description unidimensionnelle des données est la toute première phase de toute étude statistique. Elle est réalisée à travers le calcul des indicateurs statistiques de base, des tableaux de synthèse ainsi que des représentations graphiques sur chaque variable afin de résumer l'information contenues dans chaque variable.

1.1.1. Les indicateurs statistiques de base

Dans les analyses descriptives univariées, le choix du type d'indicateur statistique dépend de la nature de la variable à étudier. D'une manière générale, on utilise les caractéristiques de tendances centrales et de dispersion (moyenne, écart-type, etc..) lorsqu'il s'agit d'une variable quantitative. Et on utilise les fréquences absolues et relatives lorsqu'il s'agit des variables qualitatives.

NB : Il faut néanmoins noter qu'on peut aussi calculer les fréquences absolues et relatives sur les variables quantitatives mais on ne peut pas utiliser les indicateurs tels que la moyenne ou l'écart-type sur des variables qualitatives (définies par des modalités).

1.1.1.1. Indicateurs de valeur centrale

Moyenne arithmétique

La moyenne arithmétique $\bar{X}$ d'une variable discrète dont les valeurs sont x_1, x_2, …, x_n est définie par:

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

La moyenne arithmétique $\bar{X}$ d'une variable continue présentée en k classes telles que f_i est la fréquence et c_i le centre de la ième classe, est définie par :

$$\bar{X} = \sum_{i=1}^{k} f_i c_i$$

$$f_i = \frac{n_i}{N}$$

$$c_i = \frac{a_i + b_i}{2}$$

Où n_i est le nombre d'observations de la classe i, N est l'effectif total ; a_i et b_i représentent respectivement la borne inférieure et supérieure de la classe i.

Cette formule reste valable même s'il s'agit des valeurs discrètes où la classe n'est pas définie par une borne supérieure et inférieure mais par une valeur unique. Par exemple les variables discrètes comme le nombre d'accidents de route en un mois dans une commune, le nombre de passages de véhicules à un péage à une heure donnée, etc… Pour ces types de variables la moyenne se calcule comme suit :

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i$$

Il s'agit alors d'une moyenne pondérée (on utilisera les mêmes pondérations lorsqu'il s'agira de calculer la variance).

La moyenne est très utilisée car elle synthétise la série par une seule valeur qui tient compte de toutes les observations. Mais elle est peu robuste car très sensible aux valeurs extrêmes (contrairement à la médiane).

La médiane

La médiane partage la série des valeurs observées en deux sous-ensembles de même effectif.

Pour une variable discrète dont les valeurs ont été classées par ordre croissant, et dont la série des valeurs comporte un nombre impair de données égal à 2n +1, la médiane est la nième valeur.

Si la série comporte un nombre pair de données égal à 2n, la médiane est, par convention, la moyenne entre les deux valeurs de rang n et (n+1).

Dans le cas d'une variable continue pour laquelle on a une répartition en classes, on cherche la classe médiane $[e_{i-1}, e_i[$ telle que :

$$F(e_{i-1}) < 0{,}5 \text{ et } F(e_i) > 0{,}5$$

Où $F(e_{i-1})$ est la fonction de répartition (ou fréquence cumulée croissant) des classes. Puis on détermine la médiane M par interpolation linéaire à l'intérieur de la classe.

La médiane est un indicateur de position centrale insensible aux variations des valeurs extrêmes.

Mode ou classe modale

Le mode est la valeur la plus fréquente d'une variable discrète.

Il n'existe pas toujours pour une série ; et s'il existe il n'est pas nécessairement unique.

Pour une variable continue, la classe modale est la classe correspondant à la longueur la plus grande des rectangles de l'histogramme.

1.1.1.2. Indicateurs de dispersion

Etendue

L'étendue d'une distribution est définie par la différence entre la valeur maximale et la valeur minimale de la série. C'est est un indicateur instable car, par définition, il dépend des valeurs extrêmes.

Quartiles

Les trois quartiles Q1, Q2 et Q3 sont les valeurs partageant la série des observations en quatre parties égales. Ainsi 25% des valeurs de la série sont inférieures à Q1, 50% sont inférieures à Q2, enfin 75% sont inférieures à Q3.

Les quartiles sont des indicateurs de position. On remarque que le deuxième quartile est la médiane.

On appelle distance interquartile la quantité définie par : Q3 - Q1 .

Les quartiles d'une distribution en classes se font là encore, par interpolation linéaire.

On utilise aussi quelquefois les déciles qui partagent la série des observations en dix parties égales ou les centiles qui, eux, partagent la série en cent parties égales.

Variance et écart-type

Ce sont les deux mesures de dispersion les plus fréquemment utilisées.

On appelle variance de la série, le nombre positif défini par :

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

En développant la première expression on montre aisément que :

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

On appelle écart-type S le nombre réel positif égal à la racine carré de la variance.

Coefficient de variation

Le coefficient de variation est défini par :

$$CV = \frac{S}{\bar{x}} \times 100$$

Il ne dépend pas des unités choisies et permet d'apprécier :

- la représentativité de x par rapport à l'ensemble des données
- l'homogénéité de la distribution : un coefficient de variation d'une valeur inférieure à 15% permet de conclure à une bonne homogénéité de la distribution.

1.1.1.3. Indicateurs de forme

Une distribution parfaitement symétrique est telle que le mode, la moyenne et la médiane sont égales. Il est donc intéressant de les comparer quand on souhaite ajuster une distribution symétrique telle que celle de la loi de Gauss.

Coefficient d'asymétrie ou skewness

Il est défini par :

$$\gamma_1 = \frac{\mu_3}{S^3}$$

Où μ_3 représente le moment d'ordre 3 de la série. Et S^3 l'écart-type élevé à la puissance 3.

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^3$$

Le coefficient d'asymétrie illustre l'étalage de la série de part et d'autre de la moyenne de la variable. Pour une distribution plus étalée à droite de la moyenne, on a $\gamma_1 > 0$ et pour une distribution étalée à gauche, on a $\gamma_1 < 0$. Par exempe, il est égal à 0 pour une distribution parfaitement gaussienne et à 2 pour une distribution exponentielle

Coefficient d'aplatissement ou Kurtosis

Par définition, ce coefficient est égal à :

$$\gamma_2 = \frac{\mu_4}{S^4}$$

Où μ_4 représente le moment d'ordre 4 de la série. Et S^4 l'écart-type élevé à la puissance 4.

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Une distribution de Laplace-Gauss (loi normale) a un coefficient d'aplatissement égal à 3 et ce coefficient est égal à 9 pour une distribution exponentielle

Ces deux coefficients (asymétrie et aplatissement) fournissent un premier résultat concernant la comparaison entre la distribution de l'échantillon étudié et une distribution théorique (une distribution normale, par exemple).

Fréquence absolues et fréquences relatives

La fréquence absolue d'une valeur représente le nombre de fois que cette valeur se répète dans la série de données. Pour une variable, elle correspond au comptage de chaque valeur contenue dans cette variable.

Quant à la fréquence relative (ou proportion), elle correspond au rapport entre la fréquence absolue et l'effectif total (c'est-à-dire le nombre total de valeurs contenue dans la variable analysée). Ainsi quelle que soit la valeur i d'une variable, la fréquence absolue est :

$$fa_i = n_i$$

Et la fréquence relative est :

$$fr_i = \frac{n_i}{N}$$

Où n_i est le nombre de fois que la valeur i se répète dans les données et N le nombre total de données sur la variable étudiée.

1.1.2. Tableaux statistiques

Les tableaux statistiques ont pour but de synthétiser les informations fournies par les indicateurs statistiques de base (caractéristiques de tendance centrale, de dispersion et de forme mais aussi les fréquences).

Ces tableaux se présentent de manière différente suivant la nature des variables, en particulier selon si la variable est discrète ou continue.

D'une manière générale, on présente les caractéristiques de tendance centrale, de dispersion et de forme pour les variables quantitatives continue ou discrète. Et les tableaux de fréquence pour les variables qualitatives. Toutefois, on peut présenter les tableaux de fréquence pour les variables quantitatives discrètes ou les variables quantitatives continues regroupées en classes.

1.1.2.1. Variable discrète

Un tableau statistique décrivant une variable discrète présente usuellement, pour chaque valeur de la variable, la fréquence relative ou absolue (voire les deux) de cette valeur.

Exemple

On souhaite étudier la mobilité géographique des individus disposant d'un véhicule dans une agglomération. On sélectionne au hasard 60 individus et on observe le nombre de kilométrage réalisé en une semaine (valeurs arrondies). Les valeurs trouvées sont consignées dans la première ligne du tableau. Dans la deuxième ligne, on a indiqué le nombre d'automobilistes correspondant à chacune des mesures :

x_i	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110
$fa_i = n_i$	1	3	4	4	7	14	2	8	5	3	1	2	2	2	1	1
$fr_i = \frac{n_i}{N}$	0,02	0,05	0,07	0,07	0,12	0,23	0,03	0,13	0,08	0,05	0,02	0,03	0,03	0,03	0,02	0,02

La troisième ligne présente la fréquence relative de chaque valeur observée.

1.1.2.2. Variable continue

Les données statistiques d'un tableau de fréquence décrivant une variable continue sont regroupées en classes. Dans un tableau de ce type, figurent les extrémités e_i des classes ainsi que les effectifs de ces classes c'est à dire le nombre d'individus appartenant à chaque classe $[e_{i-1}, e_i[$. Par convention, l'extrémité droite de chaque classe est exclue de cette classe.

L'amplitude d'une classe est la longueur de l'intervalle correspondant. On peut aussi indiquer les fréquences relatives de chaque classe ainsi que les fréquences cumulées représentant les effectifs cumulés des classes d'extrémités inférieures.

Exemple

Le tableau présente la répartition, en pourcentage p, des 1800 ingénieurs d'un groupe industriel en fonction du nombre N d'années d'ancienneté dans le groupe :

N	[0, 1[	[1, 2[	[2, 3[	[3, 4[	[4, 6[	[6, 8[	[8,10[	[10,14[	[14,18[	[18,22[	≥22
p (%)	1,0	2,0	5,2	7,3	8,1	12,3	14,7	16,3	13,6	11,2	8,3

On peut remarquer que les classes ne sont pas toutes de même amplitude. La dernière classe est dite ouverte car n'y figure pas de borne supérieure.

1.1.2.3. Nombre de classes optimales

Il est fréquent d'avoir à « classer » une série de données. Il faut alors déterminer le nombre « optimale » de classes car :

-Un trop grand nombre de classes n'apportent pas de simplification notoire et les effectifs de chaque classe ne sont pas suffisamment représentatifs.

- Un très faible nombre de classes fait perdre des informations.

Usuellement et suivant le nombre de données, on recommande de prendre entre 5 et 20 classes.

Il existe cependant la formule de **Sturges** qui donne une valeur approchée du nombre k de classes souhaitable en fonction du nombre n d'observations :

$$k = 1 + 3.22 \times log_{10}(n)$$

Où n est le nombre d'observations total sur la variable quantitative à analyser.

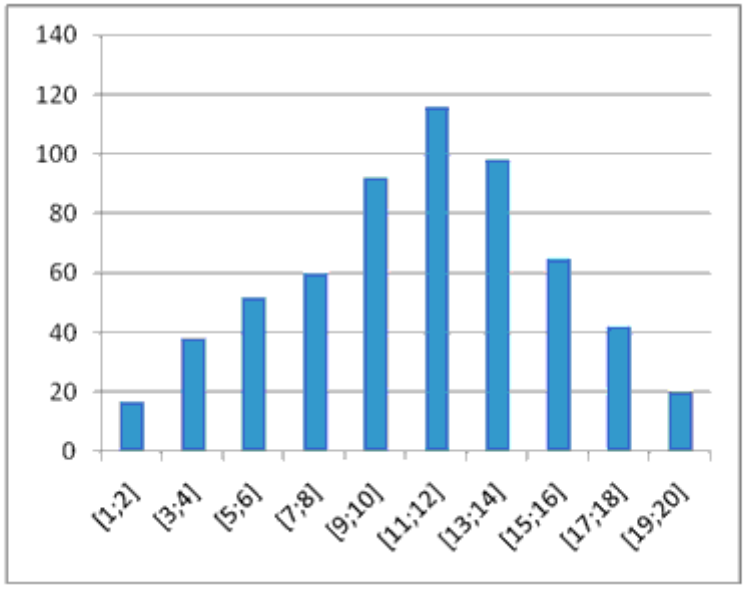
Ainsi, pour n = 100, on obtient k =7 classes et pour n = 10 000, on obtient k = 14 classes.

1.1.3. Les représentations graphiques

Les représentations graphiques sont des illustrations visuelles des informations fournies par les statistiques de base (présentées ou pas dans les tableaux statistiques). Dans le cadre de l'analyse univariée, on distingue plusieurs types de graphiques choisis (également en fonction de la nature de la variable étudiée).

1.1.3.1. Diagramme en bâtons

Un diagramme en bâtons est obtenu en portant en ordonnée, pour chaque valeur de la variable mise en abscisse, la fréquence relative correspondante. La variable mise en abscisse peut être une variable qualitative ou quantitative (voir exemple ci-dessous).



1.1.3.2. Diagramme en tige et feuille « *stem and leaf* »

Ce diagramme, appelé aussi diagramme « tige-feuille », dû à Tukey, réalise une sorte d'histogramme couché à partir des valeurs numériques constituant la série étudiée. Chaque valeur de donnée est décomposée en deux parties :

-**La tige (stem)** comprenant les chiffres principaux de chaque valeur numérique (usuellement le chiffre des centaines lorsque l'unité de mesure est en centaine et celui des dizaines si l'unité de mesure est en dizaine, etc…)

-**La feuille (leaf)** comprenant les autres chiffres (le chiffre des unités, par exemple, lorsque la tige est dizaine).

Exemple : Le tableau suivant présente le relevé des poids, en grammes, de 24 éprouvettes.

263	285	256	258	274	261	250	265	276	271	272	290
260	276	270	279	288	284	253	286	287	281	290	273

On va représenter ces valeurs en diagramme stem-leaf.

Ici la principale unité de mesure est en centaine, alors les tiges seront constituées par les centaines et les feuilles constitueront les chiffres restants.

Pour cela, il faut d'abord ordonner les données. Et distinguer les tiges et les feuilles. On obtient alors le tableau suivant :

Valeur	Centaines (tige)	Dizaine (feuilles)
250	25	0
253	25	3
256	25	6
258	25	8
260	26	0
261	26	1
263	26	3
265	26	5
270	27	0
271	27	1
272	27	2
273	27	3
274	27	4
276	27	6
276	27	6
279	27	9
281	28	1
284	28	4
285	28	5
286	28	6
287	28	7
288	28	8
290	29	0
290	29	0

On peut alors réaliser le diagramme comme suit :

25 | 0 3 6 8
26 | 0 1 3 5
27 | 0 1 2 3 4 6 6 9
28 | 1 4 5 6 7 8
29 | 0 0

Une lecture rapide du diagramme nous permet de dire que les poids sont tous entre 250 et 290 grammes et que la moyenne se situe entre 270 et 280 grammes.

1.1.3.3. Histogramme

Cette représentation est utilisée pour des séries continues dont les valeurs du caractère ont été regroupées en classes.

Un histogramme est composé de rectangles dont la largeur de chacun en abscisse est la largeur de la classe correspondante et dont la longueur en ordonnée est telle que la surface du rectangle est proportionnelle à l'effectif de la classe.

La manière de représenter dépend selon que les classes sont de même amplitude ou pas.

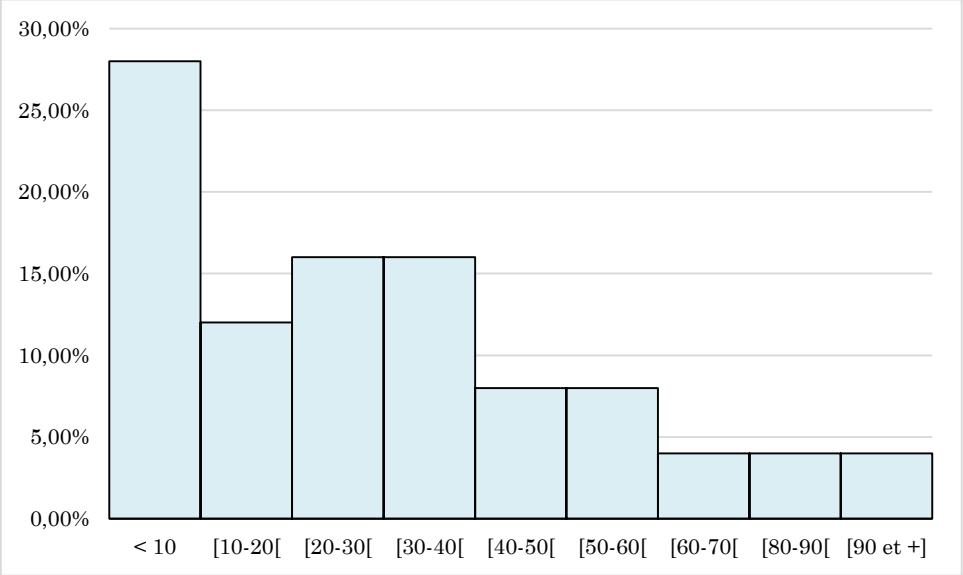
Exemple 1 : un cas particulier – classe de même amplitude

Répartition de la classe d'âge (en années) d'un échantillon de 1500 patients souffrant d'une pathologie donnée

Classes	< 10	[10-20[	[20-30[	[30-40[	[40-50[	[50-60[	[60-70[	[80-90[	[90 et +]	Total
Effectifs	420	180	240	240	120	120	60	60	60	1500

Dans le cas de classes de même amplitude, on reporte simplement en ordonnée la fréquence de chaque classe.

Tous les rectangles étant de même base, pour que l'aire soit proportionnelle à l'effectif, il suffit que la hauteur du rectangle le soit (voir figure).



Exemple 2 : cas général – classes d'amplitude quelconque

Classes	[0 ; 100[	[100 ; 150[	[150 ; 250[	[250 ; 400[	[400 ; 700[
Effectifs	100	80	120	90	60

Pour tracer un histogramme dans lequel les classes sont d'amplitudes quelconques, il faut que l'aire de chaque rectangle soit proportionnelle à l'effectif de la classe et non leur hauteur. Pour cela, il faut :

– dans un premier temps, commencer par graduer les axes c'est à dire choisir les échelles d'unité pour l'axe des abscisses (valeurs du caractère) et une unité d'aire (pour les effectifs)

– dans un second temps, déterminer la hauteur de chaque rectangle connaissant son aire et la longueur de sa base. Pour le cas d'un rectangle, la hauteur est égale à l'aire divisée par la base.

Application :

Unités choisies :

– 1 cm pour 100 en abscisse.

– 1 cm^2 pour un effectif de 20.

Calculs :

– Sur [0 ; 100[: base : 1 cm ; aire : 5 cm^2 car 5=100/20 ; hauteur : 5/1=5

On suit les mêmes étapes pour les autres classes.

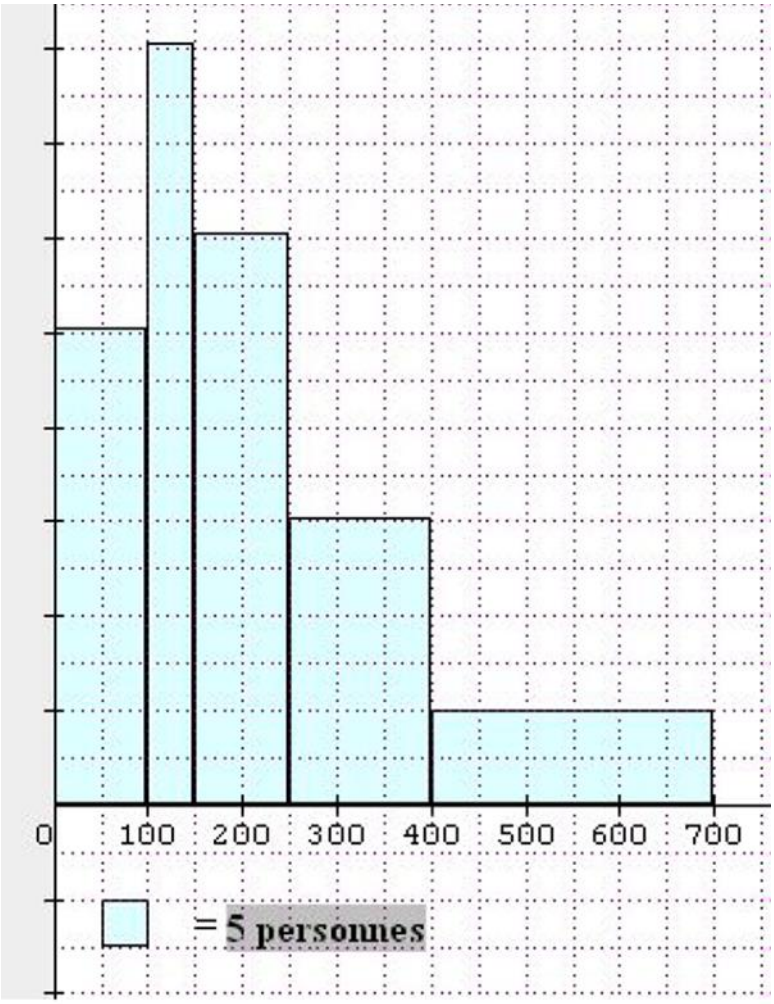
– Sur [100 ; 150[: base : 0, 5 cm ; aire : 4 cm^2 ; hauteur =4/0,5 =8 cm

– Sur [150 ; 250[: base : 1 cm ; aire : 6 cm^2 ; hauteur = 6/1= 6 cm

– Sur [250 ; 400[: base : 1, 5 cm ; aire : 4, 5 cm^2 ; hauteur= 4, 5/1,5 = 3 cm

– Sur [400 ; 700[: base : 3 cm ; aire : 3 cm^2 ; hauteur = 3/3 = 1 cm

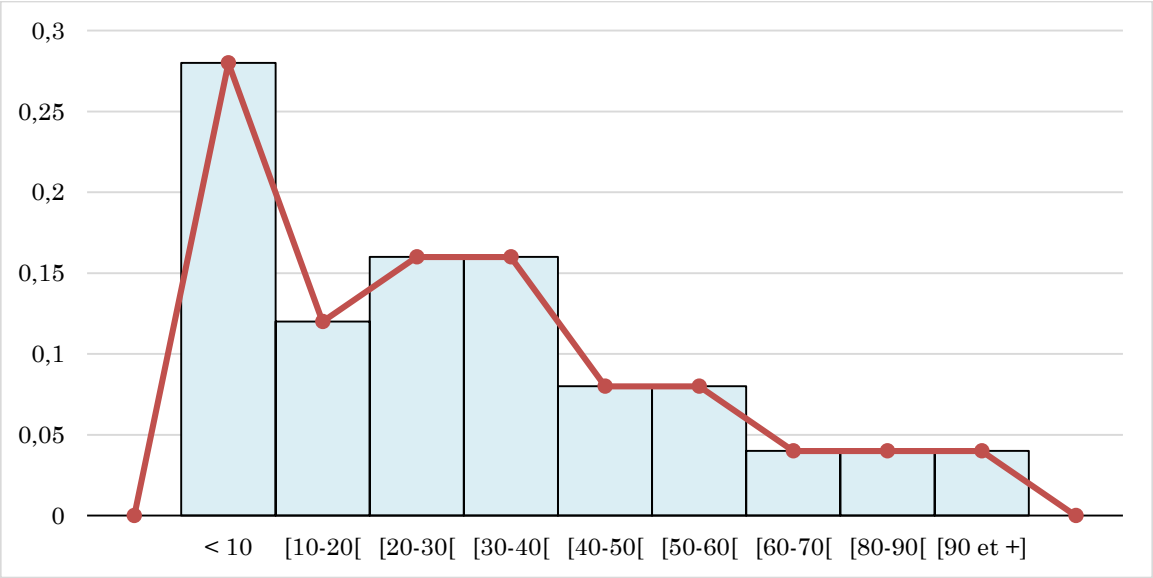
Ainsi, les bases et les hauteurs étant définies, on fait la représentation graphique suivante.



1.1.3.4. Polygone des fréquences

Le polygone de fréquence est obtenu en joignant par des segments de droite les milieux des cotés supérieurs des rectangles de l'histogramme. La courbe obtenue est fermée en créant deux classes fictives d'effectif nul à chaque extrémité.

Il permet de représenter la distribution des fréquences.



1.1.3.5. Courbes de fréquences cumulées

On considère les deux courbes de fréquences cumulées (croissante ou décroissante) construites à partir des fréquences relatives ou absolues.

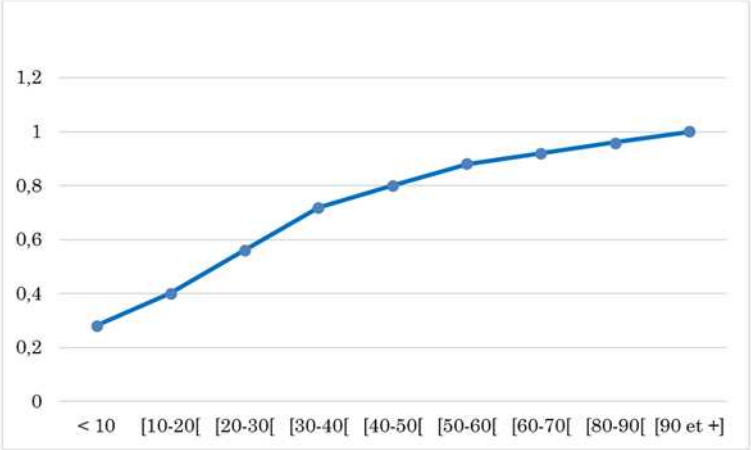
La courbe cumulative croissante est construite en joignant les points d'abscisses les limites supérieures des classes et d'ordonnées les fréquences cumulées croissantes.

La courbe cumulative décroissante joint, elle, les points ayant pour abscisses les limites inférieures des classes et pour ordonnées les fréquences cumulées décroissantes.

Exemple: Distribution des patients selon la classe d'âge

Classes	< 10	[10-20[	[20-30[	[30-40[	[40-50[	[50-60[	[60-70[	[80-90[	[90 et +]	Total
Effectifs	420	180	240	240	120	120	60	60	60	1500
Effectifs cumulés croissants	420	600	840	1080	1200	1320	1380	1440	1500	--
Fréquences	0,28	0,12	0,16	0,16	0,08	0,08	0,04	0,04	0,04	--
Fréquences cumulées croissantes	0,28	0,4	0,56	0,72	0,8	0,88	0,92	0,96	1	--

Représentation fréquence cumulée croissante



1.1.3.6. Diagramme en boîte à moustaches ou Box-Plot

Le diagramme en boîte ou boîte à moustaches représente schématiquement les principales caractéristiques d'une variable. Elle utilise 5 valeurs qui résument des données : le minimum, les 3 quartiles Q1, Q2(médiane), Q3, et le maximum.

La partie centrale de la distribution est représentée par une boîte de largeur arbitraire et de longueur la distance interquartile. La médiane est tracée à l'intérieur.

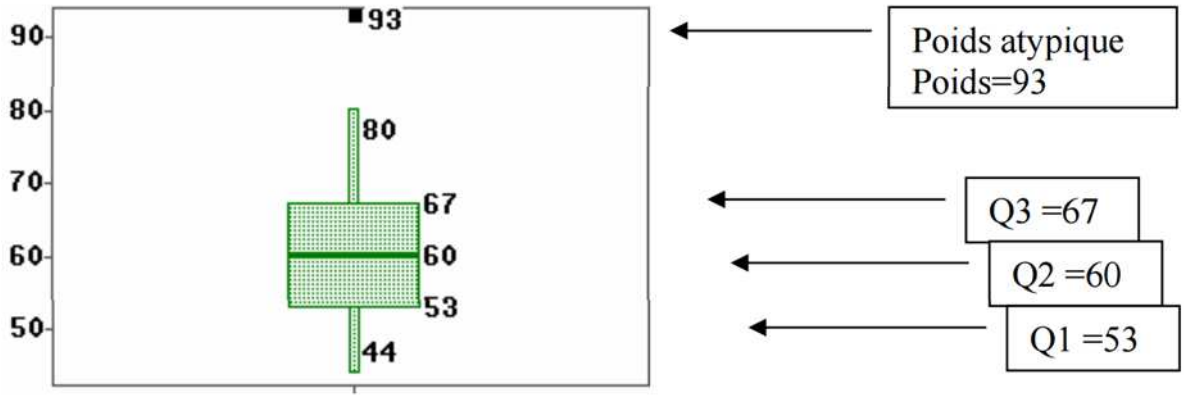
La boîte est complétée par des "moustaches" correspondant aux valeurs suivantes:

-**Valeur adjacente supérieure** : plus grande valeur inférieure à : Q3 + 1,5(Q3 – Q1)

-**Valeur adjacente inférieure** : plus petite valeur supérieure à : Q1 – 1,5(Q3 – Q1)

Les valeurs extérieures aux moustaches appelées valeurs aberrantes, sont représentées, en général, par des étoiles ou par des points.

Le graphique suivant illustre la boîte à moustaches sur les variables des individus d'un échantillon donné.



Sur la boîte à moustaches d'une variable, on distingue les éléments suivants :

• l'échelle des valeurs de la variable, située sur l'axe vertical.

• la valeur du 1er quartile Q1 (25% des effectifs), correspondant au trait inférieur de la boîte,

• la valeur du 2ème quartile Q2 (50% des effectifs), représentée par un trait horizontal à l'intérieur de la boîte,

• la valeur du 3ème quartile Q3 (75% des effectifs), correspondant au trait supérieur de la boîte,

• les 2 « moustaches» inférieure et supérieure, représentées ici par les petits rectangles verticaux de part et d'autre de la boîte. Ces 2 moustaches, délimitent les valeurs dites adjacentes qui sont déterminées à partir de l'écart interquartile (Q3-Q1).

• les valeurs dites extrêmes, atypiques, exceptionnelles, (outliers) situées au-delà des valeurs adjacentes sont individualisées. Elles sont représentées par des marqueurs (carré, ou étoile, etc.).

L'extrémité de la moustache inférieure est la valeur minimum dans les données qui est supérieure à la valeur frontière basse : Q1-1,5*(Q3-Q1) soit 32 pour la variable POIDS.

L'extrémité de la moustache supérieure est la valeur maximum dans les données qui est inférieure à la valeur frontière haute : Q3+1,5*(Q3-Q1) soit 88 pour la variable POIDS.

Dans le schéma suivant deux valeurs sont atypiques car situées au-delà de la frontière haute.

La « box plot » est aussi souvent utilisée pour comparer deux séries de mesure de la même variable. Elle permet, par exemple, de comparer les moyennes entre deux ou plusieurs groupes, ou bien d'étudier la dispersion de valeurs entre les groupes.

1.1.3.7. Courbe de concentration

Elle est très utilisée en statistique économique pour l'étude d'une variable positive cumulative telle que le revenu ou le chiffre d'affaire ou la consommation…

Considérons une variable X correspondant, par exemple, à une distribution de revenu, de fonction de répartition F et de masse totale M. La courbe de concentration est définie par l'ensemble des points tels que, pour chaque valeur x de la variable, l'abscisse est $F(x)$, proportion des individus gagnant moins que x, et l'ordonnée $G(x)$ définie par : $G(x) = \frac{Masse\ des\ revenus < x}{Masse\ totale\ de\ revenu}$

Remarque :

Pour une distribution quelconque : $F(x) > G(x)$.

La courbe de concentration est donc en dessous de la première bissectrice. Son premier point est l'origine des axes, le dernier le point de coordonnées (1, 1). Elle est toujours située dans le carré de longueur 1 dont deux cotés sont les axes.

L'indice de concentration, ou indice de Gini, est le double de l'aire définie entre la courbe de concentration et la première bissectrice.

Cet indice est compris entre 0 et 1. Plus il est petit, plus la courbe est proche de la bissectrice et donc les valeurs de F et G peu éloignées

1.2. Analyse descriptive bi(multi)variée

Dans les analyses descriptives univariées, il s'agit d'étudier la distribution d'une seule variable à la fois notamment à travers les statistiques de base, les tableaux et les graphiques. A présent, il s'agit de mener des analyses descriptives multivariées.

D'une manière générale, les analyses descriptives multivariées sont des analyses exploratoires qui permettent d'étudier les associations entre les variables. Plusieurs indicateurs statistiques permettent en effet de mesurer les associations entre les variables. Tout comme pour les analyses univariées, le choix d'un indicateur dépend de la nature des variables en jeux. Par exemple, pour examiner l'association entre variables quantitatives, on utilise généralement le coefficient de corrélation linéaire ; pour les variables qualitatives, on peut utiliser les coefficients de cramer, etc…

Cette section a pour but de revisiter les principaux indicateurs permettant de mesurer les associations entre les variables.

1.2.1. Liaison entre deux variables quantitatives

1.2.1.1. Etude graphique de la liaison

Supposons que l'on ait observé deux variables X et Y sur un ensemble de n individus. On a obtenu n couples (x_i, y_i) soit encore deux vecteurs X et Y de R_n : $X = x_i \; et \; Y = y_i$.

On peut représenter l'ensemble des points de coordonnées x_i, y_i dans un repère du plan. Cette représentation fournit une première indication sur d'éventuelles liaisons entre les deux variables.

1.2.1.2. Liaison linéaire entre deux variables quantitatives

Généralement, on examine la liaison linéaire entre deux ou plusieurs variables en utilisant soit la covariance, soit le coefficient de corrélation linéaire ou le coefficient de détermination.

Covariance entre deux variables

Soit X et Y deux variables aléatoires. On mesure la covariance entre ces deux variables est mesurée comme suit :

$$COV(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Où $\bar{X}$ et $\bar{Y}$ représentent les moyennes empiriques.

Notons aussi qu'en développant cette expression, on retrouve une nouvelle expression de la covariance telle que :

$$COV(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_iY_i) - \bar{X}\bar{Y}$$

A travers cette expression, on peut donner l'expression développée de la variance S_Y^2 telle que :

$$COV(X,X) = \frac{1}{n}\sum_{i=1}^{n}(X_iX_i) - \bar{X}\bar{X} = S_X^2$$

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

Coefficient de corrélation linéaire

Le coefficient de corrélation linéaire mesure le degré de dépendance linéaire entre deux variables. Il est égal à la covariance des deux variables divisée par le produit leur écart-type empirique :

$$r_{XY} = \frac{COV(X,Y)}{S_XS_Y}$$

Le coefficient est compris entre -1 et 1. Pour le cas de deux variables indépendantes, la corrélation est égale à 0.

Propriétés

1. $|r| \leq +1$, r représente le cosinus de l'angle formé par les vecteurs X - $\bar{X}$ et Y - $\bar{Y}$.
2. Si $|r| = 1$, il y a relation linéaire entre les deux variables et réciproquement.
3. Le coefficient de corrélation r n'est pas robuste car très sensible aux valeurs extrêmes.
4. Si les variables sont indépendantes, le coefficient de corrélation r est nul.
5. Si le coefficient r est nul, on ne peut conclure qu'à l'indépendance **linéaire** entre les variables et non à leur indépendance non linéaire.

Le coefficient de détermination

Le coefficient de détermination est le carré du coefficient de corrélation. Il représente dans quelle proportion la variance de Y est expliquée par la variation de X et réciproquement. C'est le carré de la covariance divisée par le produit des variances.

$$R_{xy}^2 = \frac{COV^2(X,Y)}{S_X^2 S_Y^2}$$

Le coefficient de détermination est compris entre 0 et 1.

1.2.2. Liaison entre deux variables qualitatives nominales

1.2.2.1. Tableaux de contingence (tableaux croisés)

Les tableaux croisés ou tableaux de contingence sont les tableaux statistiques de données pour deux variables qualitatives.

Désignons par X et Y les deux variables étudiées, par p le nombre de modalités de X et q le nombre de celles de Y. Au croisement de la ligne x_i et de la colonne y_j du tableau, figure le nombre n_{ij} de données telles que $X = x_i$ et $Y = y_j$

Y	y_1	…	y_i	…	Total ligne
X					
x_1					
…			…		
x_i		…	n_{ij}	…	$n_{i\bullet}$
…			…		
Total colonne			$n_{\bullet j}$		

Fréquences marginales et fréquences conditionnelles

On appelle fréquences marginales, les effectifs totaux par ligne ou par colonne. On les obtient à partir des expressions suivantes

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$$

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$$

Où $n_{i\bullet}$ représente la fréquence marginale-ligne i et $n_{\bullet j}$ la fréquence marginale-colonne j.

On appelle fréquence conditionnelle, les fréquences relatives par ligne ou par colonne. On les obtient à partir des expressions suivantes

$$n_{i/j} = \frac{n_{ij}}{n_{i\bullet}}$$

$$n_{j/i} = \frac{n_{ij}}{n_{\bullet j}}$$

Où $n_{i/j}$ représente la fréquence conditionnelle-ligne i obtenue en divisant l'effectif dans chaque cellule de ligne i par le total des effectifs de cette ligne. et $n_{j/i}$ la fréquence conditionnelle-colonne j obtenue en divisant l'effectif dans chaque cellule de la colonne j par le total des effectifs de cette colonne.

On peut aussi définir de la même façon :

la fréquence relative de la modalité (i, j) :

$$f_{ij} = \frac{n_{ij}}{N}$$

ainsi que les fréquences relatives marginales $f_{i\bullet}$et $f_{\bullet j}$ et les fréquences relatives conditionnelles.

On appelle alors tableau des profils-lignes le tableau des fréquences conditionnelles $n_{j/i}$ et tableau des profils-colonnes les tableau des fréquences conditionnelles $n_{j/i}$.

NB :

$$\sum_{i=1}^{p} n_{i\bullet} = \sum_{j=1}^{q} n_{\bullet j} = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = N$$

Où N est l'effectif total

Ainsi, après avoir établi les profils-lignes et colonnes ainsi que les fréquences marginales et conditionnelles, on peut maintenant proposer plusieurs mesures d'association entre variables qualitatives.

1.2.2.2. La distance de khi-deux : D

On appelle mesure de liaison D entre deux variables qualitatives l'expression suivante :

$$D = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{N}\right)^2}{\frac{n_{i\bullet} \times n_{\bullet j}}{N}}$$

Lorsque cette valeur tend vers 0, cela démontre une absence de lien entre les deux variables. Pour les variables indépendantes, la valeur de D est 0.

NB : Une valeur de D différente ne signifie pas nécessairement une association significative entre les deux variables. Il faut alors mener un test statistique pour

vérifier si cette statistique est significative. On utilise pour cela le test d'indépendance de khi-2. Nous aborderons les tests dans le quatrième chapitre. Pour l'instant, nous proposons des mesures d'association entre les variables sans examiner si c'est associations sont statistiquement significatives ou pas.

1.2.2.3. Autres mesures d'association entre variables qualitatives nominales

D'autres coefficients ont été définis à partir du D de khi-deux pour mesurer le degré d'association entre les variables qualitatives. Ce sont notamment :

- le coefficient de contingence de Pearson :

$$P = \sqrt{\left(\frac{D}{D+N}\right)}$$

Où N est l'effectif total et D la distance de khi-deux définie précédemment.

- le coefficient de Cramer :

$$C = \sqrt{\frac{D}{N \times min(p-1; q-1)}}$$

Ce coefficient a pour avantage d'être compris entre 0 et 1. L'indépendance est représentée par la valeur 0 alors la valeur 1 correspond à une liaison fonctionnelle parfaite entre les deux variables.

1.2.3. Liaison entre deux variables qualitatives ordinales

Il est assez fréquent de disposer de deux classements d'un même ensemble d'objets. L'exemple le plus usuel est celui du classement par deux membres de jury d'un festival de n films de courts métrages. On dispose alors de ces deux classements :

Films	1	2	…	p	…	n
Classement du 1er juré	r_1	r_2	…	r_p	…	r_n
Classement du 2ième juré	s_1	s_2	…	s_p	…	s_n

Dans cette expérience, chaque classement est une permutation des n premiers entiers. Dès lors, étudier la liaison entre les deux variables revient à comparer les classements générés par ces deux variables.

Plusieurs indicateurs ont été alors proposés pour évaluer cette liaison.

1.2.3.1. Coefficient de corrélation des rangs de Spearman

Définition

Le coefficient de corrélation de rang de Spearman se présente comme suit :

$$r_s = \frac{Cov(r,s)}{s_r s_s}$$

Les rangs étant des permutations des n premiers entiers, on utilise alors les résultats sur la distribution discrète de la loi uniforme sur [1, n] :

$$\bar{r} = \bar{s} = \frac{n+1}{2}$$

$$s_r^2 = s_s^2 = \frac{n^2-1}{12}$$

Le coefficient s'écrit alors :

$$r_s = \frac{\frac{1}{n}\sum_{i=1}^{n} r_i s_i - \left(\frac{n+1}{2}\right)^2}{\frac{n^2-1}{12}}$$

En posant

$$d_i = r_i - s_i$$

$$\frac{1}{n}\sum_{i=1}^{n} r_i s_i = \frac{1}{2}\sum_{i=1}^{n} {u_i}^2 + \frac{1}{2}\sum_{i=1}^{n} {v_i}^2 - \frac{1}{2}\sum_{i=1}^{n} {d_i}^2$$

Or les u_i et les v_i sont des entiers variant entre 1 et n :

$$\sum_{i=1}^{n} {u_i}^2 = \sum_{i=1}^{n} {v_i}^2 = \frac{n(n+1)(2n+1)}{6}$$

On en déduit :

$$\frac{1}{n}\sum_{i=1}^{n} r_i s_i - \left(\frac{n+1}{2}\right)^2 = \frac{1}{2n}\sum_{i=1}^{n} {d_i}^2 + \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2$$

Le calcul du deuxième terme est simple :

$$\frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

D'où l'expression du coefficient de Spearman

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} {d_i}^2$$

Propriétés

$-1 \leq r_s \leq 1$

$r_s = 1$ les classements sont identiques, $d_i = 0$ pour tout i.

$r_s = -1$ les classements sont inverses l'un de l'autre.

$r_s = 0$ les classements sont indépendants.

1.2.3.2. Coefficient de corrélation des rangs de Kendall

Définition

Soit r_i , r_j et s_i , s_j les rangs, dans les deux classements, d'un couple de deux objets (i, j).

Au couple (i, j) on attribue :

+1 si les deux objets sont dans le même ordre : $r_i < r_j$ et $s_i < s_j$

-1 si les deux classements sont discordants : $r_i < r_j$ et $s_i > s_j$

On somme les valeurs obtenues pour les $\frac{n(n-1)}{2}$ couples (i, j) distincts. Soit S le résultat, le coefficient de Kendall est défini par :

$$\tau = \frac{2S}{n(n-1)}$$

Propriétés

$-1 \leq \tau \leq 1$

En effet d'après la définition de S, on a :

$$-\frac{n(n-1)}{2} \leq S \leq \frac{n(n-1)}{2}$$

$\tau = 1$ les classements sont identiques.

$\tau = -1$ les classements sont inverses l'un de l'autre.

$\tau = 0$ les classements sont indépendants.

1.2.4. Liaison entre une variable qualitative et une variable quantitative : le rapport de corrélation

Lorsqu'on recherche une liaison éventuelle entre une variable quantitative et une variable qualitative, on a alors besoin du rapport de corrélation.

Le rapport de corrélation de la variable Y (quantitative) en la variable X (qualitative) est la mesure de liaison, non symétrique, définie par :

$$\eta^2_{Y/X} = \frac{V[E(Y/X)]}{V(Y)}$$

$\eta^2_{Y/X}$ étant généralement inconnu, on utilise son équivalent empirique définit comme suit :

Supposons que la variable X présente k modalités d'effectifs n_1, n_2, …, n_k. Notons $\bar{y}_i$ la moyenne de Y pour la catégorie k et $\bar{y}$ la moyenne totale de Y. On définit le rapport de corrélation empirique entre la variable qualitative X et la variable quantitative Y par :

$$e^2 = \frac{\frac{1}{n}\sum_{j=1}^{k}(\bar{y}_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_i^j - \bar{y}\right)^2}$$

Propriétés

$$0 \leq e^2 \leq 1$$

Si $e^2 = 0$ alors il n'y a pas de dépendance en moyenne. En effet : $e^2 = 0$ implique, $\bar{y}_i = \bar{y}$ pour toutes les valeurs de i.

Si $e^2 = 1$, pour une modalité i de X, tous les individus ont la même valeur et ceci pour toutes les valeurs de l'indice.

Chapitre 2 : Variables aléatoires et modèles de probabilités

2.1. Notions de probabilité

2.1.1. Expérience aléatoire et évènements aléatoires

Une expérience est dite d'aléatoire si on ne peut pas déterminer à l'avance avec certitude son résultat c'est à dire si répétée dans les mêmes conditions, elle aboutit à des résultats différents. Toutefois, le nombre de résultats possible est supposé être fini bien qu'il ne soit pas nécessairement dénombrable. Par exemples les lancers répétés d'un dé, la mesure du diamètre d'une pièce mécanique, etc… Les résultats possibles d'une expérience aléatoire constituent l'ensemble fondamental Ω appelé aussi **univers des possibles**.

Un événement aléatoire est une assertion relative au résultat de l'expérience. L'ensemble des événements liés à cette expérience est l'ensemble des parties de Ω noté $P(\Omega)$

Exemple : dans le lancer simultané de deux dés, on s'intéresse aux chiffres inscrits sur les faces supérieures.

- L'ensemble fondamental est Ω = {(1, 1), (1, 2),… , (6, 6)}. Il possède 36 éléments.
- Le fait que « la somme des deux chiffres soit inférieure ou égale à 5 » est un événement aléatoire possible ; la partie Ω′ de Ω correspondante est : Ω′ = {(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2)}.

2.1.2. Vocabulaire sur les ensembles d'événements

Nous allons nous référer à des propriétés ensemblistes usuelles :

1. A tout événement , on associe son contraire $\bar{A}$,
2. Aux événements A et B , on associe leur union $A \cup B$ ou leur intersection $A \cap B$
3. L'événement certain est représenté par Ω,
4. L'événement impossible est représenté par ∅.
5. A et B sont deux événements incompatibles si la réalisation de l'un exclut celle de l'autre, c'est à dire si les parties A et B sont disjointes.
6. A_1, A_2, …, A_n forment un système complets d'événements si elles constituent une partition de Ω : elles sont disjointes deux à deux et leur réunion forme Ω tout entier.

2.1.3. Quelques axiomes du calculs de probabilités

On considère dans ce chapitre que les ensembles sont finis. Considérons une expérience aléatoire dont l'univers des possibles est Ω. On a vu que l'ensemble des événements liés à cette expérience est l'ensemble des parties de Ω noté $P(\Omega)$.

2.1.3.1. Probabilités élémentaires

On qualifie de probabilité toute application P de $P(\Omega)$ dans [0, 1] telle que :

$$P(\Omega) = 1$$

Propriétés

De ces axiomes on déduit les propriétés suivantes :

$$P(\emptyset) = 0$$

$$P(\bar{A}) = 1 - P(\mathrm{A})$$

$$A \subset B \Rightarrow P(\mathrm{A}) \leq P(\mathrm{B})$$

pour tout couple (A, B) d'événements incompatibles :

$$P(A \cup B) = P(\mathrm{A}) + P(\mathrm{B})$$

pour tout couple (A, B) d'événements non incompatibles :

$$P(A \cup B) = P(\mathrm{A}) + P(\mathrm{B}) - P(A \cap B)$$

2.1.3.2. Théorème des probabilités totales

Soit $(A_1, A_2, \ldots, A_n)$ un système complet d'événements de Ω, alors $\forall A_i \in \Omega, \mathrm{i} = 1, \ldots, \mathrm{n}$, on a :

$$P(A_i) = \sum_{j=1}^{n} P\left(A_i \cap B_j\right) \; ; i = 1, \ldots, n \; ; j = 1, \ldots, n$$

2.1.3.3. Espaces probabilisés

Considérons un espace Ω fini. A tout élément a_i de $\Omega, i = 1, \ldots, N$ on associe un nombre réel positif ou nul $P(a_i)$ tel que :

$$\sum_{i=1}^{n} P(a_i) = 1$$

A toute partie A_i de Ω constitué de k éléments a_i on associe le nombre :

$$P(A_i) = \sum_{j=1 | a_j \in A_i}^{k} P(a_j) \leq 1$$

On définit ainsi une probabilité sur l'ensemble des parties de Ω.

Supposons que l'on attribue le même poids à chaque événement élémentaire, on a :

$$P(a_i) = \frac{1}{card(\Omega)} = \frac{1}{N}$$

$$\forall\, i \in 1,..,\mathrm{N}$$

On définit une probabilité uniforme sur cet ensemble et la probabilité associée à une partie A quelconque de Ω est alors définie par :

$$P(A_i) = \frac{card(A_i)}{card(\Omega)} = \frac{n_i}{N} \;\forall\, i \in 1,..,\mathrm{n}$$

Exemple

On réalise l'expérience de lancée de deux dés. Et on s'intéresse à la probabilité de l'événement A « la somme des deux chiffres est inférieure ou égale à 5 ».

En supposant que les dés ne sont pas pipés (toutes les faces ont le même poids), la probabilité de cet évènement est le suivant :

$$P(A) = \frac{card(A)}{card(\Omega)}$$

Avec Ω = {(1, 1), (1, 2),… , (6, 6)} soit 36 éléments $card(\Omega) = 36$.

A = {(1, 1), (1, 2), (1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2),…,} $\Rightarrow card(A) = 9$

$$P(A) = \frac{10}{36}$$

On peut, par exemple, remarquer que l'événement aléatoire B : « au moins un des chiffres est égal à 5 » est incompatible avec l'événement A… puisqu'aucune face des dés ne porte l'évènement « 0 » (évènement impossible).

2.1.4. Probabilités conditionnelles et évènements indépendants

Considérons deux événements A et B. Supposons qu'on ne s'intéresse à la réalisation de A que sous la condition que B soit déjà réalisé. Cela revient à rechercher la réalisation de A ∩ B par rapport à B.

2.1.4.1. Définition

Soit Ω l'univers des possibles d'une expérience aléatoire et B un événement de probabilité non nulle.

On appelle probabilité conditionnelle de A sachant B l'application de $P(A)$ dans [0, 1], définie par :

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

Exemple

Dans le jeu du lancer de deux dés, on rappelle que A est l'événement élémentaire : « la somme des deux chiffres est inférieure ou égale à 5 ». On définit l'événement B comme étant l'événement : « les deux chiffres obtenus lors des lancers sont pairs ».

On a :

$$P(A) = \frac{10}{36}$$

Pour l'évènement B, on a :

$$P(B) = \frac{9}{36}$$

On peut remarquer que seul le couple (2, 2) satisfait aux deux exigences c'est-à-dire « la somme des deux chiffres est inférieure ou égale à 5 » et « les deux chiffres obtenus lors des lancers sont pairs ». On a alors :

$$P(A \cap B) = \frac{1}{36}$$

On en déduit la probabilité conditionnelle :

$$P(A|_B) = \frac{\frac{1}{36}}{\frac{9}{36}} = \frac{1}{9}$$

2.1.4.2. Evènement indépendants

L'événement A est indépendant de l'événement B si :

$$P(A|_B) = P(A)$$

·Conséquences :

Si A est indépendant de B, alors B est indépendant de A.

A et B sont indépendants si et seulement si :

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

Les événements (A_1, A_2, …, A_n) sont dits mutuellement indépendants si, pour toute partie I de l'ensemble {1, 2, …, n}, on a :

$$P\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} P(A_i)$$

Cette propriété est plus forte que l'indépendance deux à deux qui en est une simple conséquence.

2.1.4.3. L'expression Bayésienne des probabilités conditionnelles

La formule de Bayes est une expression généralisée des probabilités conditionnelles.

Considérant que :

$$P(A \cap B) = P(A|_B) \times P(B) = P(B|_A) \times P(A)$$

On en déduit la première formule de Bayes :

$$P(A|_B) = \frac{P(B|_A) \times P(A)}{P(B)}$$

$$P(B|_A) = \frac{P(A|_B) \times P(B)}{P(A)}$$

Soit (B_i) un système complet d'événements. Le théorème des probabilités totales de A s'écrit :

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P\left(A|_{B_i}\right)P(B_i)$$

On en déduit la deuxième formule de Bayes :

$$P(B_i|_A) = \frac{P\left(A|_{B_i}\right) \times P(B_i)}{\sum_{j=1}^{n} P\left(A|_{B_j}\right) P\left(B_j\right)}$$

2.2. Notion de variables aléatoires

Définition

Soit Ω l'univers des possibles d'une expérience aléatoire ayant un nombre fini d'issues. Toute application X de $P(\Omega)$ dans R définit une **variable aléatoire réelle**. L'ensemble $X(\Omega)$ des valeurs prises par la variable aléatoire s'appelle univers-image.

Exemple

Dans l'exemple du lancer des deux dés, on a vu que l'univers des possibles Ω est constitué des 36 couples. La probabilité d'obtenir un quelconque de ces couples est de 1/36.

Si on s'intéresse à la somme des deux chiffres obtenus, on définit une application X de Ω dans R. L'ensemble d'arrivée est l'ensemble E = {2, 3, …, 12}. E est appelé univers-image. X est une variable aléatoire définie sur $P(\Omega)$ ensemble des parties de Ω.

Pour obtenir la probabilité d'une valeur quelconque de E, il suffit de chercher la partie de A antécédent de cette valeur par X.

Par exemple pour une variable aléatoire définie telle que la somme des deux chiffres est égale à 6, la probabilité est:

$$P_X(6) = P[X^{-1}(6)] = P[\{(1,5),(2,4),(3,3),(4,2),(5,1)\}] = 5/36$$

2.3. Notion de loi de probabilité

Définir une loi de probabilité d'une variable aléatoire X c'est associer à chacune des valeurs possibles a_i de la variable X une valeur probabilité correspondante de sorte

$$\sum_{i=1}^{n} P(a_i) = 1$$

Plus généralement la loi de probabilité de X, notée P_X est définie sur R par :

$$\forall\, A \subset R \quad P_X(A) = P\left(a|_{X(a)\in A}\right) = P[X^{-1}(A)]$$

On définit bien ainsi une probabilité sur R.

2.3.1. Notion de fonction de répartition

On appelle fonction de répartition d'une variable aléatoire X l'application F de R dans [0, 1] définie par :

$$F(x) = P(X < x)$$

Propriétés :

F est monotone croissante avec :

$$F(-\infty) = 0$$

$$F(+\infty) = 1$$

Pour tout intervalle [a, b] de R.

$$P(a < X < b) = F(b) - F(a) = P(X < b) - P(X < a)$$

2.3.2. Notion de densité de probabilité

On appelle fonction densité de probabilité de la variable aléatoire X la quantité $f(x)$ définie telle que :

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

pour une variable continue ;

$$F(x) = \sum_{k=1|x_k<x}^{n} f(x_k)$$

pour une variable discrète.

A travers ces deux expressions, on peut distinguer deux grandes catégories de lois de probabilités : les lois de probabilités **continues** et les lois de probabilités **discrètes**. Le reste de chapitre sera consacrée à l'étude de ces deux catégories de lois.

La figure 2.1 illustre graphiquement la relation entre la fonction de densité de probabilité et fonction de répartition.

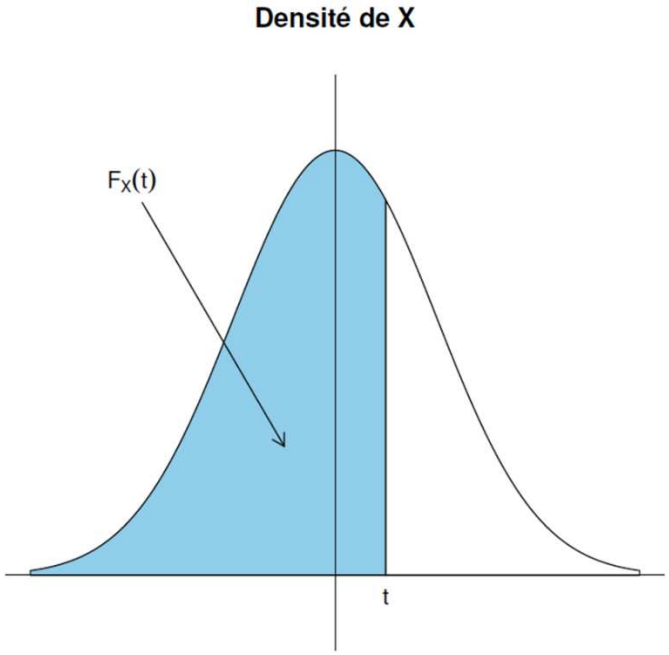


Figure 2.1: Relation densité / fonction de répartition

2.3.3. Notion de fonction quantile

Étant donné une variable aléatoire réelle X de fonction de répartition $F(X)$ on appelle fonction quantile de X, que l'on notera Q_X, la fonction définie sur]0; 1[par la relation :

$$Q_X(p) = F^{-1}(x) = q \Leftrightarrow F(q) = p$$

La fonction quantile est donc la réciproque de la fonction de répartition.

La figure 2.2 illustre graphiquement la relation entre fonction de répartition et fonction quantile pour une valeur $p = \beta$ et une valeur $q = t$

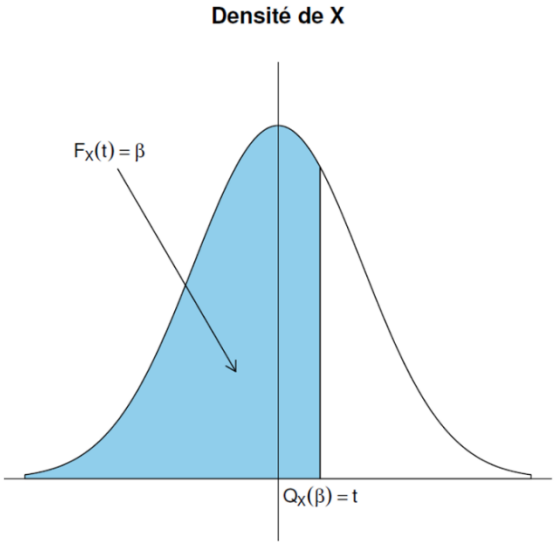


Figure 2.2: Relation fonction de répartition / fonction quantile

2.4. Les lois de probabilités discrètes

2.4.1. Définition

Une variable aléatoire X est dite discrète si elle ne prend qu'un nombre fini ou dénombrable de valeurs x_1, x_2,… x_n.

La loi de probabilité d'une variable discrète est dite loi de probabilité discrète. Elle est caractérisée par :

- Possibilité d'énumérer toutes les valeurs x_i
- les densités de probabilité se présentent comme suit :

$$p_i = P(X = x_i)$$

Elles correspondent généralement aux fréquences relatives définies telle que :

$$p_i = P(X = x_i) = \frac{card(X = x_i)}{card(\Omega)} = \frac{n_i}{N}$$

2.4.2. Caractéristiques d'une loi discrète : espérance et variance

Les principales caractéristiques de tendance et de dispersion d'une loi discrète sont les deux premiers moments à savoir l'espérance et la variance. Celles-ci s'expriment comme suit :

$$E(X) = \sum_{i=1}^{n} p_i x_i$$

$$var(X) = \sum_{i=1}^{n} p_i (x_i)^2 - \big(E(X)\big)^2$$

Avec

$$p_i = P(X = x_i) = \frac{card(X = x_i)}{card(\Omega)} = \frac{n_i}{N}$$

Exemple

Dans le jeu du lancer de deux dés, on a vu que la somme S des deux chiffres inscrits sur les faces supérieures des dés peut prendre une des valeurs de E = {2, 3, 4,…,12}. Les probabilités associées sont simples à calculer et présentées dans le tableau suivant :

S	2	3	4	5	6	7	8	9	10	11	12
P(S=s)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

On calcule alors aisément l'espérance et la variance grâce aux expressions ci-dessus. On trouve E(S)= 7 et VAR(S)=5,83.

2.4.3. Etudes de quelques lois de probabilités discrètes

2.4.3.1. Loi uniforme discrète

Définition

Soit X une variable aléatoire prenant chaque valeur de l'ensemble {1, 2, …, n}. La loi de X est dite uniforme sur [1, n] si :

$$\forall\, k \in [1, \mathrm{n}] \quad P(X = k) = \frac{1}{n}$$

Ainsi tous les éléments de l'intervalle $[1, \mathrm{n}]$ ont donc la même probabilité $\frac{1}{n}$. On parle alors **d'équiprobabilité**.

La loi uniforme discrète est généralement noté $U(1, n)$

Espérance et Variance de la loi uniforme discrète

En appliquant les formules précédentes, on peut facilement calculer l'espérance et la variance. En effet :

$$E(X) = \sum_{i=1}^{n} p_i x_i = \sum_{i=1}^{n} \frac{1}{n} x_i = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n}(1 + 2 + \cdots + n) = \frac{1}{n} \times \frac{n(n+1)}{2} = \frac{(n+1)}{2}$$

(Propriété somme des n premiers termes d'une suite arithmétique de raison r=1)

$$E(X) = \frac{(n+1)}{2}$$

On montre aisément que la variance est $Var(X) = \frac{n^2-1}{12}$ car :

$$var(X) = \sum_{i=1}^{n} p_i (x_i)^2 - \big(E(X)\big)^2 = \frac{1}{n} \sum_{i=1}^{n} {x_i}^2 - \big(E(X)\big)^2 = E(X^2) - \big(E(X)\big)^2$$

$$E(X^2) = \frac{1}{n}(1^2 + 2^2 + \cdots + n^2) = \frac{(n-1)(2n+1)}{6}$$

$$var(X) = \frac{(n-1)(2n+1)}{6} - \left(\frac{(n+1)}{2}\right)^2$$

$$Var(X) = \frac{n^2 - 1}{12}$$

2.4.3.2. Loi de Bernoulli

Définition

C'est la loi d'une variable aléatoire X ne pouvant prendre que deux valeurs notées 1 ou 0, appelées communément succès ou échec. Soit p la probabilité que la variable X soit égale à 1 (probabilité du succès), alors $1 - p$ est la probabilité que X soit égale à 0 (probabilité de l'échec) :

$$P(X = 1) = p \; et \; P(X = 0) = 1 - p$$

La loi de Bernoulli est généralement noté $B(p)$

En situation d'équiprobabilité comme dans le cas du lancer d'un jeton pile ou face (non truqué), on a :

$$P(X = 1) = P(X = 0) = \frac{1}{2}$$

Espérance et variance de la loi de Bernoulli

L'espérance et la variance d'une loi de Bernoulli dont la probabilité de succès est p se calculent comme suit :

$$E(X) = \sum_{i=1}^{n} p_i x_i = 1 \times p + 0 \times (1-p) = p$$

$$E(X) = p$$

$$var(X) = \sum_{i=1}^{n} p_i (x_i)^2 - \big(E(X)\big)^2 = [p(1^2) + (1-p)(0^2)] - p^2 = p - p^2 = p(1-p)$$

$$var(X) = p(1-p)$$

2.4.3.3. La loi binomiale

Définition

D'une manière sommaire, la loi binomiale est une loi qui naît de plusieurs répétitions d'une expérience de Bernoulli.

Supposons que l'on répète n fois, dans des conditions identiques, une expérience aléatoire dont l'issue se traduit par l'apparition ou la non-apparition d'un événement A de probabilité p. Le résultat de la ième expérience est noté X_i et la loi de X_i est la loi de Bernoulli de paramètre p. On suppose que chaque résultat est indépendant du précédent. On appelle Y la variable aléatoire égale au nombre d'apparitions de A dans les n répétitions (épreuves). Y peut prendre les valeurs allant de 0 (aucune apparition de A dans les n répétitions) à n apparitions (l'évènement a est apparu dans les n répétitions). Bien entendu Y peut prendre des valeurs intermédiaires situées entre 0 et n. On dit alors que Y est distribuée selon une loi binomiale de paramètres n et p. Elle est généralement notée B(n, p).

Ainsi de façon formelle, on peut définir une loi binomiale comme suit :

Soit $X_i\{i = 1, \dots, n\}$ n variable indépendantes suivant toutes une loi de Bernoulli de paramètre p. La variable Y définie telle que $Y = \sum_{i=1}^{n} X_i$ est distribuée selon une loi binomiale de paramètres n et p.

Ainsi pour chaque valeur k prise par Y telle que $k = 0, 1, \dots, n$, on a l'expression générale de la probabilité suivante :

$$P(Y = k) = C_n^k p^k (1-p)^{n-k}$$

On peut distinguer deux valeurs particulières dans cette expression généralisée.

D'abord lorsqu'il n'y a aucune apparition de A, on a k=0 et ainsi

$$P(Y = 0) = (1-p)^n$$

Ensuite lorsqu'il y a n apparitions de A, on a $k = n$; on a alors :

$$P(Y = n) = p^n$$

Espérance et variance de la loi binomiale

Pour calculer l'espérance de la loi binomiale on se sert de la propriété de linéarité de l'opérateur d'espérance (se référer à la fin de cette section pour plus de détails sur l'opérateur d'expérance). En effet, puisque :

$$Y = \sum_{i=1}^{n} X_i$$

On a

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

On a montré que :

$$E(X_i) = p$$

Alors

$$E(Y) = \sum_{i=1}^{n} p = np$$

$$E(Y) = p$$

Pour calculer la variance, on applique le principe de l'indépendance. On sait que la variance d'une somme de variables aléatoires indépendante est la somme des variances. Dès lors on a :

$$Var(Y) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} var(X_i)$$

$$Var(X_i) = p(1-p)$$

Alors

$$Var(Y) = np(1-p)$$

Propriétés :

Si Y suit la loi binomiale $B(n,p)$, la variable Z telle que $Z = n - Y$ suit la loi $B(n,1-p)$. Cette formule est importante lorsqu'il plus facile de calculer la probabilité de Z pour ensuite en déduire la probabilité de Y.

Pour deux variables aléatoires Y_1 et Y_2 suivant respectivement des lois B(n_1, p) et B(n_2, p). Si Y_1 et Y_2 sont indépendantes alors la variable Z telle que $Z = Y_1 + Y_2$ suit une loi B($n_1 + n_2$, p).

2.4.3.4. Loi géométrique

Définition

La loi géométrique est une loi qui sert à modéliser le nombre de répétitions nécessaire d'une expérience de Bernoulli pour faire apparaître un évènement A.

Par exemple on souhaite jouer à un jeu de loterie de lancer d'un dé où le gain est matérialisé par la sortie de l'évènement « 4 ». On soupçonne que le dé est pipé c'est-à-dire que la probabilité de sortie de sortie des différents cotés n'est pas égale à 1/6. On veut maintenant étudier nombre minimum de fois qu'il faut jouer pour gagner la première fois à ce jeu c'est-à-dire faire apparaitre le côté « 4 ». On peut alors utiliser une loi géométrique. Pour cela, on transforme cette expérience en une répétition d'expériences de Bernoulli décrit comme suit. On définit l'événement A « le côté sortie est 4 » noté 1 et l'évènement contraire « le côté sortie n'est pas 4 » noté 0. Soit p la probabilité que le chiffre sortie soit 4. Dans ces conditions on peut maintenant considérer que nombre d'essais nécessaires pour faire apparaître 4 pour la première fois est une loi géométrique.

Soit k le nombre fois minimum de répétition, la fonction de densité d'une loi géométrique se présente comme suit :

$$P(Y = k) = p(1-p)^{k-1} \quad k \geq 1$$

Espérance et variance d'une loi géométrique

On démontre facilement que :

$$E(Y) = \frac{1}{p}$$

$$Var(Y) = \frac{1-p}{p^2}$$

2.4.3.5. Loi hypergéométrique

Définition

Dans une population de taille N, une proportion p d'individus possèdent une caractéristique C_0 particulière (par exemple la proportion d'ordinateurs défectueux dans une série de fabrication d'usine). On prélève un échantillon de taille n dans cette population, le tirage s'effectuant sans remise. Soit X la variable aléatoire égale au nombre d'individus de l'échantillon possédant la propriété C_0 . La variable X suit une loi de probabilité hypergéométrique notée $H\ (N, n, p)$ si et seulement si la fonction de densité s'écrit comme suit :

$$P(X = k) = \frac{C_{Np}^{k} C_{N-Np}^{n-k}}{C_N^n}$$

Le nombre possible d'échantillons de taille n dans la population est C_N^n

Dans un quelconque échantillon de taille n, il y a k individus possédant la propriété C_0, choisis parmi les Np individus de la population entière possédant C_0 ; (n –k) autres individus ne la possédant pas ; ils sont choisis parmi les (N–Np) individus de la population ne possédant pas C_0.

Remarques

X est la somme de n variables de Bernoulli, **non indépendantes**, correspondant aux tirages successifs des n individus.

Soit X_1 la variable aléatoire correspondant au tirage du premier individu. X_1 suit une loi de Bernoulli de paramètre p. $E(X_1) = p$ et $V(X_1) = p(1- p)$.

Soit X_2 la variable aléatoire correspondant au tirage du deuxième individu (tirage sans remise). On va chercher la loi de X_2. En appliquant le théorème des probabilités totales :

$$P(X_2 = 1) = P\left(X_2 = 1|_{X_1=1}\right)P(X_1 = 1) + P\left(X_2 = 1|_{X_1=0}\right)P(X_1 = 0)$$

$$P(X_2 = 1) = \frac{Np - 1}{N - 1}p + \frac{Np}{N - 1}(1 - p) = p$$

X_2 ne prenant que les deux valeurs 1 et 0, on en déduit que $P(X_2 = 0) = 1 - p$. X_2 suit donc une loi de Bernoulli de paramètre p.

On procède ainsi pour toutes les variables X_i.

Au final on aboutit à la conclusion que chaque X_i suit bien une loi de Bernoulli de paramètre p. mais qu'en revanche, elles ne sont pas indépendantes.

Espérance et variance d'une loi hypergéométrique

En utilisant les propriétés vues précédemment, on démontre aisément que :

$$E(X) = np$$

$$Var(X) = \left(\frac{N-n}{N-1}\right) np(1-p)$$

Pour la cas de la variance penser à la propriété :

$$Var(X) \sum_{i=1}^{n} var(X_i) + 2COV\left(\sum_{j=1|i\neq j}^{n} var(X_i, X_j)\right)$$

2.4.3.6. La loi de poisson

Définition

Une variable aléatoire discrète suit une loi de poisson lorsqu'elle peut prendre toutes les valeurs entières positives ou nulles dont les probabilités s'expriment telles que :

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Espérance et variance d'une loi de poisson

Espérance

$$E(X) = \sum_{i=1}^{n} p_i x_i = \sum_{k} \left(\frac{e^{-\lambda}\lambda^k}{k!}\right) k = e^{-\lambda}\lambda \sum_{k} \left(\frac{\lambda^{k-1}}{(k-1)!}\right) = \lambda$$

$$E(X) = \lambda$$

Variance

$$Var(X) = \sum_{i=1} p_i (x_i)^2 = \sum_{k} \left(\frac{e^{-\lambda}\lambda^k}{k!}\right) k^2 - \lambda^2$$

$$= \sum_{k} \left(\frac{e^{-\lambda}\lambda^k}{k!}\right) k(k-1) + \sum_{k} \left(\frac{e^{-\lambda}\lambda^k}{k!}\right) k - \lambda^2$$

$$= \lambda^2 + \lambda - \lambda^2 = \lambda$$

$$Var(X) = \lambda$$

On peut donc remarquer que pour la loi de poisson

$$E(X) = Var(X) = \lambda$$

2.4.4. Quelques rappels sur l'opérateur d'espérance E(.)

2.4.4.1. Définition et propriétés de l'opérateur d'espérance

De façon simple, l'espérance correspond à la moyenne de cette variable lorsque n est grand. Elle se calcule par la formule suivante :

$$\mathrm{E(X)} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Dans certains cas, au lieu d'utiliser $\bar{X}$, on utilise E(X). Par exemple, pour calculer la variance, on écrit :

$$\mathrm{VAR(X)} = \frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathrm{E(X)}\big)^2$$

De plus, sachant que $\mathrm{VAR(X)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$, on peut réécrire cette expression comme suit :

$$\mathrm{VAR(X)} = \mathrm{E(X^2)} - \big(\mathrm{E(X)}\big)^2$$

Avec

$$\mathrm{E(X^2)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Cette formulation de la variance s'avère d'une importance capitale dans beaucoup de démonstrations. Elle peut se généraliser quelle que soit la variable considérée. Soit une variable Z telle que $Z = XY$, on a :

$$\mathrm{E(Z)} = \mathrm{E}(XY) = \frac{1}{n}\sum_{i=1}^{n} X_i\, Y_i$$

$$\mathrm{VAR(Z)} = \frac{1}{n}\sum_{i=1}^{n}\big(X_i Y_i - \mathrm{E(Z)}\big)^2$$

Or on sait que :

$$\mathrm{VAR(Z)} = \frac{1}{n}\sum_{i=1}^{n}(X_i Y_i)^2 - \big(\mathrm{E(Z)}\big)^2$$

Par conséquent :

$$\mathrm{VAR(Z)} = \mathrm{E(Z^2)} - \big(\mathrm{E(Z)}\big)^2$$

Ainsi de façon explicite, on peut écrire :

$$\mathrm{VAR(XY)} = \mathrm{E(X^2Y^2)} - \big(\mathrm{E(XY)}\big)^2$$

Et lorsque les deux variables X et Y sont indépendantes alors $\mathrm{E(XY)} = \mathrm{E(X)} * \mathrm{E(Y)}$. Ainsi, on a :

$$\mathrm{VAR(XY)} = \mathrm{E(X^2Y^2)} - \big(\mathrm{E(X)E(Y)}\big)^2$$

On peut élargir ce type de raisonnement au cas de la covariance entre X et Y. En effet,

$$\mathrm{COV(X,Y)} = \frac{1}{n}\sum_{i=1}^{n}\big(X_i - \mathrm{E(X)}\big)\big(Y_i - \mathrm{E(Y)}\big) = \frac{1}{n}\sum_{i=1}^{n}(X_i Y_i) - \mathrm{E(X)E(Y)}$$

$$\mathrm{COV(X,Y)} = \mathrm{E(XY)} - \mathrm{E(X)E(Y)}$$

Si les deux variables sont indépendantes $\mathrm{E(XY)} = \mathrm{E(X)E(Y)}$, Par conséquent

$$\mathrm{COV(X,Y)} = 0$$

Aussi, la formule de la covariance peut être généralisée quelle que soit la puissance de X et de Y. On a :

$$\mathrm{COV}(\mathrm{X}^r, \mathrm{Y}^r) = \mathrm{E}(\mathrm{X}^r\mathrm{Y}^r) - \mathrm{E}(\mathrm{X}^r)\mathrm{E}(\mathrm{Y}^r)$$

Exemple :

$$\mathrm{COV}(\mathrm{X}^2, \mathrm{Y}^2) = \mathrm{E}(\mathrm{X}^2\mathrm{Y}^2) - \mathrm{E}(\mathrm{X}^2)\mathrm{E}(\mathrm{Y}^2)$$

2.4.4.2. Quelques utilisations de l'opérateur d'espérance $E(.)$

Calcul de l'espérance dans le cas de la somme ou du produit de deux variables aléatoires

- ***Espérance d'une somme*** $\mathrm{E(X+Y)}$**:**

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

- ***Espérance d'un produit*** $\mathrm{E(XY)}$**:**

On sait que :

$$\mathrm{COV}(X,\mathrm{Y}) = \mathrm{E}(\mathrm{XY}) - \mathrm{E}(\mathrm{Y})\mathrm{E}(\mathrm{Y}) \Rightarrow$$

$$\mathrm{E}(\mathrm{XY}) + \mathrm{COV}(X,\mathrm{Y}) + \mathrm{E}(\mathrm{Y})\mathrm{E}(\mathrm{Y})$$

Si les deux variables sont indépendantes $\mathrm{COV}(X,\mathrm{Y}) = 0$. Ainsi, on a :

$$\mathrm{E}(\mathrm{XY}) = \mathrm{E}(\mathrm{X})\mathrm{E}(\mathrm{Y})$$

L'espérance du produit de 2 variables aléatoires indépendantes est le produit des espérances.

Calcul de la variance dans le cas de la somme ou du produit de deux variables aléatoires

- ***Variance d'une somme*** $\mathrm{VAR}(\mathrm{X}+\mathrm{Y})$**:**

$$VAR(X+Y) = E((X+Y)^2) - [E(X+Y)]^2$$

$$= E(\mathrm{X}^2 + \mathrm{Y}^2 + 2XY) - [E(X) + E(Y)]^2$$

$$= E(\mathrm{X}^2) + E(\mathrm{Y}^2) + 2E(XY) - [E(X)]^2 - [E(Y)]^2 - 2E(X)E(Y)$$

$$= E(\mathrm{X}^2) - [E(X)]^2 + E(\mathrm{Y}^2) - [E(Y)]^2 + 2[E(XY) - E(X)E(Y)]$$

$$VAR(X) + VAR(Y) + 2\mathrm{COV}(\mathrm{X},\mathrm{Y})$$

$$VAR(X+Y) = VAR(X) + VAR(Y) + 2\mathrm{COV}(\mathrm{X},\mathrm{Y})$$

Si les deux variables sont indépendantes, on a $2\mathrm{COV}(\mathrm{X},\mathrm{Y}) = 0$ ainsi on a:

$$VAR(X+Y) = VAR(X) + VAR(Y)$$

- ***Variance d'un produit*** $\mathrm{VAR}(\mathrm{XY})$**:**

$$\mathrm{VAR}(\mathrm{XY}) = \mathrm{E}(\mathrm{X}^2\mathrm{Y}^2) - [\mathrm{E}(\mathrm{XY})]^2$$

Or $\quad \mathrm{COV}(X,\mathrm{Y}) = \mathrm{E}(\mathrm{XY}) - \mathrm{E}(\mathrm{X})\mathrm{E}(\mathrm{Y}) \Rightarrow$

$$\mathrm{E}(\mathrm{XY}) = \mathrm{COV}(X,\mathrm{Y}) + \mathrm{E}(\mathrm{X})\mathrm{E}(\mathrm{Y})$$

$$\mathrm{COV}(\mathrm{X}^2,\mathrm{Y}^2) = \mathrm{E}(\mathrm{X}^2\mathrm{Y}^2) - \mathrm{E}(\mathrm{X}^2)\mathrm{E}(\mathrm{Y}^2) \Rightarrow$$

$$\mathrm{E}(\mathrm{X}^2\mathrm{Y}^2) = \mathrm{COV}(\mathrm{X}^2,\mathrm{Y}^2) + \mathrm{E}(\mathrm{X}^2)\mathrm{E}(\mathrm{Y}^2)$$

Ainsi, on a : $\mathrm{VAR}(\mathrm{XY}) = \mathrm{COV}(\mathrm{X}^2,\mathrm{Y}^2) + \mathrm{E}(\mathrm{X}^2)\mathrm{E}(\mathrm{Y}^2) - \left[COV(X,Y) + \mathrm{E}(\mathrm{X})[\mathrm{E}(\mathrm{Y})]\right]^2$

On sait que : $\mathrm{E}(\mathrm{X}^2) = \mathrm{VAR}(\mathrm{X}) + [\mathrm{E}(\mathrm{X})]^2$ et $\mathrm{E}(\mathrm{Y}^2) = \mathrm{VAR}(\mathrm{Y}) + [\mathrm{E}(\mathrm{Y})]^2$

Ainsi : $\mathrm{VAR}(\mathrm{XY}) = \mathrm{COV}(\mathrm{X}^2,\mathrm{Y}^2) + [\,\mathrm{VAR}(\mathrm{X}) + [\mathrm{E}(\mathrm{X})]^2].[\,\mathrm{VAR}(\mathrm{Y}) + [\mathrm{E}(\mathrm{Y})]^2] - \left[COV(X,Y) + \mathrm{E}(\mathrm{X})[\mathrm{E}(\mathrm{Y})]\right]^2 \qquad (1.13a)$

A noter que dans le cas de l'indépendance:

$$\mathrm{COV}(\mathrm{X}^2, \mathrm{Y}^2) = \mathrm{COV}(\mathrm{X}, \mathrm{Y}) = 0$$

Ce qui permet donc de réduire la formule à:

$$\mathrm{VAR}(\mathrm{XY}) = [\,\mathrm{VAR}(\mathrm{X}) + [\mathrm{E}(\mathrm{X})]^2].\,[\,\mathrm{VAR}(\mathrm{Y}) + [\mathrm{E}(\mathrm{Y})]^2] - \big[\mathrm{E}(\mathrm{X})[\mathrm{E}(\mathrm{Y})]\big]^2$$

Ainsi, en développant cette expression, on retrouve la formule initiale

$$\mathrm{VAR}(\mathrm{XY}) = \mathrm{VAR}(\mathrm{X})\mathrm{VAR}(\mathrm{Y}) + \mathrm{VAR}(\mathrm{X})[\mathrm{E}(\mathrm{Y})]^2 + \mathrm{VAR}(\mathrm{Y})[\mathrm{E}(\mathrm{X})]^2$$

Autres formules particulières

$$\mathrm{VAR}(\mathrm{aX} + \mathrm{b}) = \mathrm{a}^2\mathrm{VAR}(\mathrm{X})$$

$$\mathrm{COV}(\mathrm{aX}; \mathrm{Y}) = \mathrm{aCOV}(\mathrm{X}; \mathrm{Y})$$

2.5. Lois de probabilité continues

2.5.1. Définitions

Une variable aléatoire X est dite continue en loi s'il existe une fonction f non négative définie sur R telle que, pour toute partie A de R on a :

$$P(X \in A) = \int_A f(x)dx$$

Cette fonction f s'appelle densité de probabilité de la variable aléatoire X. L'ensemble des valeurs prises par une variable aléatoire continue est infini indénombrable.

La fonction de répartition F d'une variable aléatoire X est alors définie par :

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

Nous avons déjà présenté les propriétés générales d'une fonction de densité et d'une fonction de répartition.

2.5.2. Espérance et variance d'une loi de probabilité continue

L'espérance et la variance d'une variable aléatoire continue X sont définies par :

$$E(X) = \int_R f(x)xdx$$

$$Var(X) = \int_R f(x)x^2dx = \big(E(X)\big)^2$$

2.5.3. Etude de quelques lois de probabilité continues

2.5.3.1. La loi uniforme continue U(0,a)

Définition

Une variable aléatoire réelle X suit une loi uniforme sur l'intervalle [0, a] si sa loi de probabilité admet pour densité :

$$f(x) = \frac{1}{a} \quad \forall\, x \in [0, a];$$

$$f(x) = 0 \quad x \in]-\infty, 0[$$

$$f(x) = 0 \quad \forall\, x \in]\, a, +\infty[$$

Sa fonction de répartition se présente alors comme suit :

$$\begin{cases} F(x) = \dfrac{x}{a} \quad \forall\, x \in [0, a] \\ F(x) = 0 \quad \forall\, x \in]-\infty, 0[\\ F(x) = 1 \quad \forall\, x \in]\, a, +\infty[\end{cases}$$

Espérance et variance d'une loi uniforme continue

$$E(X) = \int_0^a f(x)xdx = \int_0^a \left(\frac{1}{a}\right)xdx = \frac{1}{a}\int_0^a xdx = \frac{1}{a}\left[\frac{x^2}{2}\right]_0^a = \frac{a^2}{2a} - 0 = \frac{a}{2}$$

$$E(X) = \frac{a}{2}$$

$$Var(X) = \int_0^a f(x)x^2dx - \big(E(X)\big)^2 = \int_0^a \frac{1}{a}x^2dx - \left(\frac{a}{2}\right)^2 = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12}$$

$$Var(X) = \frac{a^2}{12}$$

2.5.3.2. Loi exponentielle

Définition

Une variable aléatoire réelle positive suit une loi exponentielle de paramètre λ, positif, si sa densité de probabilité est définie par :

$$f(x) = \lambda e^{-x\lambda} \quad \forall\, x \geq 0$$

$$f(x) = 0 \quad \forall\, x < 0$$

Sa fonction de répartition est égale à :

$$F(x) = \int_0^x f(x)dx = \int_0^x \left(\lambda e^{-\lambda t}\right)dt = 1 - e^{-x} \quad \forall\, x \geq 0$$

$$F(x) = 0 \quad \forall\, x < 0$$

Espérance et variance d'une loi exponentielle

L'espérance et la variance d'une loi exponentielle se présente comme suit :

$$E(X) = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

(Facilement démontrable)

On peut remarquer que :

$$E(X) = \sqrt{Var(X)} = \sigma_x = \frac{1}{\lambda}$$

L'espérance de la loi exponentielle est égale à l'écart-type. On se souvient que dans le cas de la loi de poisson (loi discrète) l'espérance est égale à la variance.

La loi exponentielle est généralement utilisée dans les études de fiabilité. Elle permet, par exemple, de modéliser la durée de vie de circuits électroniques ou de matériel subissant des défaillances brutales. Dans ces cas $E(X) = \frac{1}{\lambda}$ est appelée MTBF (Mean Time Between Failure) et λ s'appelle le taux de défaillance.

2.5.3.3. La loi gamma

Définition

Une variable aléatoire positive X suit une loi Gamma de paramètre r, notée γ_r, si sa densité est donnée par :

$$f(x) = \frac{1}{\Gamma(r)} e^{-x} x^{r-1}$$

Où $\Gamma(.)$ est la fonction d'Euler définie telle que :

$$\Gamma(r) = \int_0^{+\infty} e^{-y} t^{r-1} dy$$

Une loi gamme définie avec $r > 1$ est appelée loi d'**Erlang**

Espérance et Variance d'une loi gamma

L'espérance et la variance d'une loi gamma est constantes et égales à r. On a :

$$E(X) = Var(X) = r$$

Sur ce point la loi gamma partage cette même propriété d'égalité de l'espérance à la variance avec la loi de poisson bien que cette dernière soit une loi discrète.

En termes d'application pratique, la loi Gamma permet de modéliser les temps de défaillance de matériels en études de fiabilité. Dans la théorie des files d'attente, la loi Gamma représente la loi de probabilité d'arrivée de t événements dans un processus poissonnien. Elle donc utilisée dans l'étude du passage des appels sur un réseau téléphoniques.

2.5.3.4. Loi normale ou loi de Laplace-Gauss

Cette loi joue un rôle incontournable en statistique. C'est la loi de référence pour étudier les phénomènes courants (« normaux »). Elle s'applique alors dans un grand nombre de domaines. De plus, elle apparaît comme la loi limite pour de nombreuses lois lorsque les échantillons sont de très grandes tailles. *On comprend alors qu'avec un recul d'observation suffisant on verra que tous les phénomènes tendent à se « normaliser ».*

Définition

Une variable aléatoire réelle X suit une loi de Laplace-Gauss de paramètres m et σ^2 si sa densité, définie sur R, a pour expression :

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

Cette loi est notée : LG (m, σ^2) ou N (m, σ^2).

Sa fonction de répartition s'écrit alors :

$$F(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dt$$

Elle n'a pas d'expression analytique plus simple et mais elle a été tabulée pour les valeurs $m = 0$ et $\sigma^2 = 1$ alors appelée loi normale centrée et réduite.

La fonction de densité de la loi normale centrée et réduite est :

$$f(x)_{\substack{m=0 \\ \sigma^2=1}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Espérance et Variance d'une variable normale N (m, σ^2)

$$E(x) = \int_R f(x)xdt = \frac{1}{\sqrt{2\pi}\sigma} \int_R e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} xdx$$

Posons :

$$\frac{x-m}{\sigma} = z$$

On a :

$$E(X) = \frac{\sigma}{\sqrt{2\pi}} \int_R ze^{-\frac{1}{2}z^2} du + m \int_R f(x)dx$$

La première intégrale est nulle (intégration sur R d'une fonction impaire) et la deuxième est égale à 1 donc on a :

$$E(X) = 0 + m = m$$

$$E(X) = m$$

De la même façon on démontre que :

$$Var(X) = \sigma^2$$

D'où alors la dénomination « loi normale de moyenne m et de variance σ^2 »

Cas particulier : Lorsque $m = 0$ et $\sigma^2 = 1$, on a alors une loi normale centrée réduite.

La figure 2.3 illustre l'allure d'une loi normale centrée réduite.

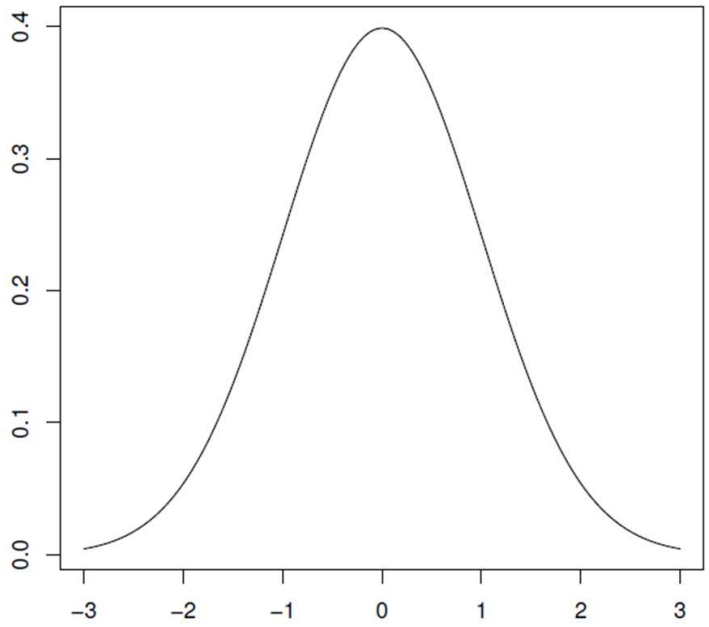


Figure 2.3: Loi normale de paramètres 0 et 1 : N(0; 1)

Étant donnée une variable aléatoire réelle X distribuée selon une loi normale d'espérance m et de variance σ^2, alors la version centrée réduite de X suit une loi normale d'espérance 0 et de variance 1.

$$X \rightsquigarrow N(m, \sigma^2) \Rightarrow Z = \frac{X - m}{\sigma} \rightsquigarrow N(0,1)$$

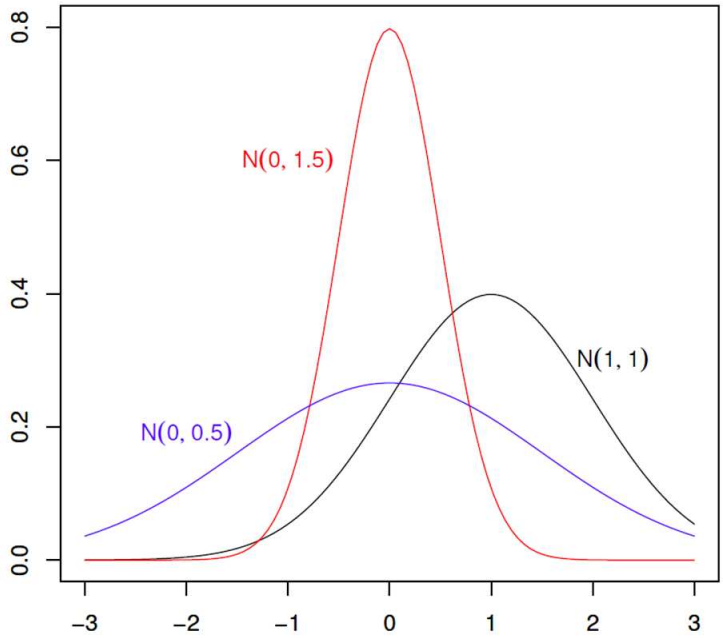


Figure 2.4: Allure de la fonction de répartition de la loi normale selon la valeur des paramètres

Symétrie de la loi normale

La loi normale est une loi symétrique autour de la moyenne. En d'autres termes, tous les quantiles qui se trouvent à la même distance de la moyenne (qu'ils négatifs ou positifs) ont la même probabilité. Par exemple, sur la figure 2.3, les quantiles $x = 2$ a la même probabilité que $x = -2$.

La symétrie de la loi normale vient de la propriété suivante :

$$X \rightsquigarrow N(m, \sigma^2) \Rightarrow 2m - X \rightsquigarrow N(\mu, \sigma^2)$$

On déduit de cette propriété fondamentale les deux résultats suivants :

$$P(X \leq t) = P(X \geq 2m - t)$$

$$Q_X(p) = 2m - \sigma Q_X(1 - p)$$

Un cas particulier : $m = 0$, $\sigma^2 = 1$, alors on a :

$$X \rightsquigarrow N(0,1) \Rightarrow -X \rightsquigarrow N(0,1)$$

$$P(X \leq t) = P(X \geq -t)$$

$$Q_X(p) = -Q_X(1 - p)$$

2.5.3.5. Les lois dérivables de la loi normale

La loi de Student

Étant donné un entier $l > 1$, une variable aléatoire X est dite distribuée selon une loi de Student à l degrés de liberté si sa densité de probabilité est :

$$f(x) = C_l \left(1 + \frac{x^2}{l^2}\right)^{-\frac{l+1}{2}}$$

Où C_l est une constante de normalisation pour faire en sorte que la probabilité totale soit égale à 1. Elle vaut $C_l = \frac{1}{\sqrt{\pi l}} \frac{\Gamma(\frac{l+1}{2})}{\Gamma(\frac{l}{2})}$ avec Γ (.) représentant la fonction d'Euler. On écrit alors $X \rightsquigarrow T_l$.

NB : Pour toute valeur de l, la variable T est centrée c'est-à-dire E(X)=0.

Symétrie de la loi de Student.

A l'instar de la loi normale, la loi de Student est une loi symétrique. Ainsi, pour tout $\beta \in]0; 1[$, on a :

$$Q_X(p) = -Q_X(1-p)$$

La figure 2.5 illustre l'allure de la loi de Student en comparaison de la loi normale centrée réduite.

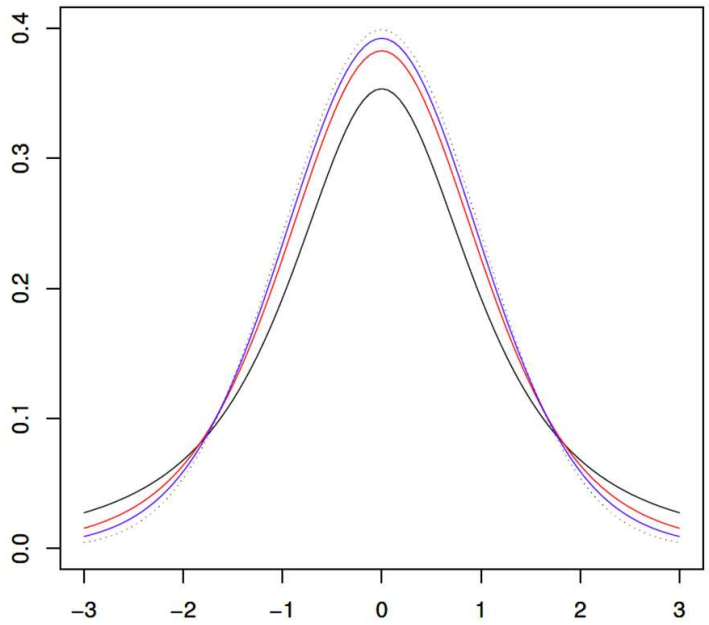


Figure 2.5: Plusieurs lois de Student (en pointillés, loi N(0; 1))

Loi du Khi-deux χ^2

Étant donné un entier $l > 1$, une variable aléatoire X est dite distribuée selon une loi du χ^2 (khi-deux) à l degrés de liberté si sa densité de probabilité est :

$$f(x) = C_l x^{\frac{l}{2}-1} exp\left(-\frac{x}{2}\right)$$

Où C_l est une constante de normalisation égale à $\frac{1}{2^{\frac{l}{2}}\Gamma(\frac{l}{2})}$ avec $\Gamma(.)$ représentant la fonction d'Euler. On écrit alors $X \rightsquigarrow \chi^2{}_l$.

NB : Pour la loi de χ^2, on a :

$$E(X) = l$$

$$V(X) = 2l$$

La figure 2.6 illustre l'allure de la loi de Khi-deux.

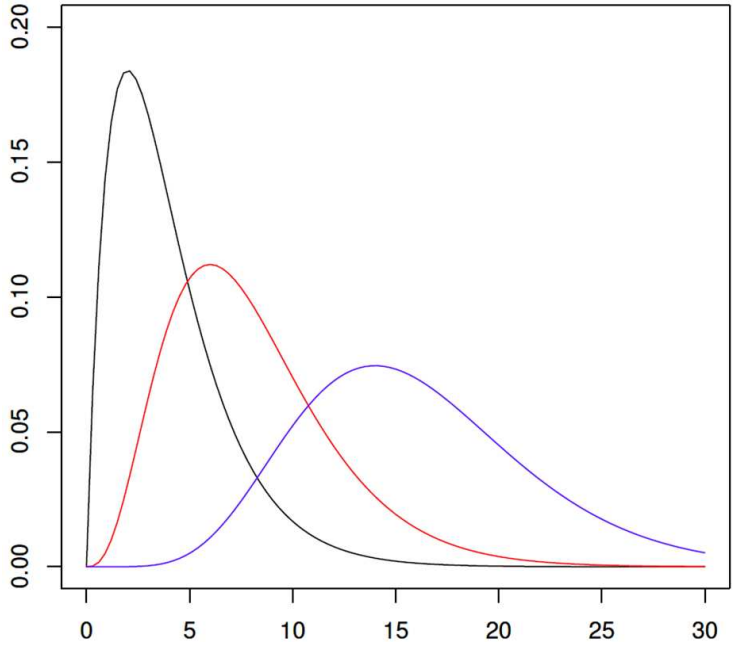


Figure 2.6 : Différentes allures de la loi de khi-deux.

Relation entre la loi normale centrée réduite, la loi de Khi-deux et la loi de Student.

1. Si X suit une loi normale N(0; 1) alors X^2 suit une loi du χ^2 à 1 degré de liberté :

$$X \rightsquigarrow N(0,1) \Rightarrow X^2 \rightsquigarrow \chi^2(1)$$

2. Pour généraliser cette relation, il en vient que, si X_1, X_2,. . . , X_n sont des variables aléatoires indépendantes distribuées selon la loi normale N(0; 1) alors la somme des carrés suit une loi du χ^2 à n degrés de liberté.

$$X_1^2 + X_2^2 + \cdots + X_n^2 = \sum_{i=}^{n} X_i^2 \rightsquigarrow \chi^2(\mathrm{n})$$

3. Pour deux variable aléatoire X_1 et X_2 , lorsque X_1 suit une loi normale centrée réduite $N(0,1)$et X_2 suit une loi de khi-deux à n degrés de liberté, alors le rapport défini par :

$$\frac{X_1}{\sqrt{\frac{(X_2)^2}{n}}} \rightsquigarrow T(n)$$

Cette expression montre que la loi de Student s'obtient à un rapport près de la loi normale centrée réduite et de la loi de khi-deux.

La loi de Fisher

La loi de Fisher s'obtient par le rapport de deux variables aléatoires suivant chacune une loi de khi-deux. Soient X_1 et X_2 deux variables aléatoires suivant une loi de khi-deux de degrés de liberté respective n_1 et n_2 alors le rapport défini par

$$\frac{\left(\frac{X_1}{n_1}\right)}{\left(\frac{X_2}{n_2}\right)} \rightsquigarrow F(n_1, n_2)$$

NB : En utilisant toutes ces relations d'équivalence, on peut aisément montrer que le carré d'une variable de Student à n degrés de liberté est une variable de Fisher à 1 et n degrés de liberté. Car :

$$(T(n))^2 = \frac{\frac{(X_2)^{2^2}}{1}}{\frac{(X_2)^2}{n}} \rightsquigarrow F(1, n_2)$$

Bien que la loi de Fisher soit le rapport entre deux lois de khi-deux, elle dispose de sa propre fonction de densité qui se présente comme suit :

$$f(x) = \frac{\Gamma(n)}{\Gamma^2(n)} \frac{\left(\frac{n}{p}\right)^{\frac{n}{2} \times \frac{n}{x^2} - 1}}{\left(1 + \frac{n}{p}x\right)^{\frac{n+p}{2}}}$$

Où Γ est la fonction d'Euler définie précédemment.

On peut aussi montrer que :

$$E(X) = \frac{p}{p-2} \quad p > 2$$

$$V(X) = 2\frac{p^2}{n}\frac{n+p-2}{(p-2)^2(p-4)} \quad p > 4$$

2.6. Théorèmes d'approximation des lois

2.6.1. Théorème centrale limite : convergence vers la loi normale

Soit (X_1 X_2,…, X_n) une suite de variables aléatoires indépendantes et identiquement distribuées iid (c'est-à-dire suivant toutes la même loi connue ou inconnue) d'espérance m et de variance σ^2, la quantité définie par :

$$\frac{(X_1 + X_n + \cdots + X_n) - nm}{\sigma\sqrt{n}}$$

converge en loi vers une variable aléatoire Z distribuée selon une loi normale (0, 1) pour n suffisamment grand.

Ce théorème est qualité de théorème central-limite ou théorème de la limite centrale. Il montre que toutes les lois convergent vers la loi normale N(0,1) lorsque n est suffisamment grand. C'est pourquoi la loi normale N(0,1) est considérée comme la loi statistique universelle pour n suffisamment grand.

Réarrangeons légèrement cette formule en divisant le numérateur et le dénominateur par n. On a :

$$\frac{\frac{(X_1 + X_n + \cdots + X_n) - nm}{n}}{\frac{\sigma\sqrt{n}}{n}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - m}{\frac{\sigma}{\sqrt{n}}}$$

On reconnait que :

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

Ainsi on a :

$$\frac{\frac{(X_1 + X_n + \cdots + X_n) - nm}{n}}{\frac{\sigma\sqrt{n}}{n}} = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$$

Nous allons le voir dans le prochain chapitre que $\bar{X}$ (la moyenne empirique) est une variable aléatoire telle que :

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

$$var(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Dès lors $\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}$ à la valeur centrée et réduite de la moyenne qu'on peut réécrire comme suit :

$$\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}\left(\frac{\bar{X}-m}{\sigma}\right)$$

On peut ainsi reformuler le théorème central limite comme suit :

$$\frac{(X_1 + X_n + \cdots + X_n) - nm}{\sigma\sqrt{n}} = \sqrt{n}\left(\frac{\bar{X}-m}{\sigma}\right)$$

Cette quantité tend vers une loi normale centrée réduite. Cela correspond donc à la deuxième reformulation du théorème qui s'énonce comme suit :

Soit (X_1 X_2,…, X_n) une suite de variables aléatoires indépendantes et identiquement distribuées iid d'espérance m et de variance σ^2, la quantité définie par $\sqrt{n}\left(\frac{\bar{X}-m}{\sigma}\right)$ converge en loi vers une variable aléatoire Z distribuée selon une loi normale (0, 1) pour n suffisamment grand.

$$\sqrt{n}\left(\frac{\bar{X}-m}{\sigma}\right) \xrightarrow{L} N(0,1)$$

Exemple d'application : Approximation d'une loi binomiale par la loi normale

Soit X une variable aléatoire suivant la loi binomiale de paramètres B(n, p) où $n = 40$ et $p = 0{,}3$. Montrer en prenant $X = 11$ que la loi normale est une bonne approximation de la loi binomiale.

En effet, l'espérance de la loi binomiale est $E(X) = np = 40 \times 0{,}3 = 12$; sa variance est $Var(X) = np(1-p) = 40 \times 0{,}3(1-0{,}3) = 8{,}4$.

En approximant la loi binomiale B(n,p) par la loi normale N(m, σ^2), on considère : $m = E(X) = 12$ et $\sigma^2 = Var(X) = 8{,}4$.

A présent calculons la probabilité de $X = 11$ dans chaque loi.

Pour la loi binomiale, on sait que :

$$P(X = k) = C_n^k p^k (1-p)^{n-k}$$

$$P(X = 11) = C_{40}^{11} (0{,}3)^{11} (1-0{,}3)^{40-11}$$

$$P(X = 11) = 0{,}132$$

Pour la loi normale, la fonction de densité est :

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

$$f(11) = \frac{1}{\sqrt{2\pi}\sqrt{8{,}4}} e^{-\frac{1}{2}\left(\frac{11-12}{\sqrt{8{,}4}}\right)^2} = 0{,}130$$

L'approximation étant assez satisfaisante, on peut donc considérer que X suit approximativement une loi normale N(12, 8,4) .

2.6.2. Les approximations entre les lois

Au-delà de l'approximation qui considère la loi normale comme la loi limite, il existe aussi d'autres approximations entre différentes lois définies sous certaines conditions. Ces possibilités d'approximations sont résumées comme suit :

1- On peut approximer une loi hypergéométrique H(N,n,p) par une loi binomiale B(n,p) lorsque $n < 0{,}10N$
2- On peut approximer une loi binomiale B(n,p) par une loi de poisson P(λ) lorsque p<0,10 et N>50.
3- On peut approximer une loi de poisson P(λ) par une loi normale N(m, σ^2) lorsque λ est définie telle que λ=np avec λ>18.
4- Bien entendu, on peut directement approximer une loi binomiale B(n,p) par une loi normale N(m, σ^2) avec m=np et σ^2=np(1-p). avec la condition que np>5 et np(1-p)>5.

Remarque

D'une manière générale, l'approximation d'une loi A par une loi B consiste à estimer les paramètres de la loi A (ex : espérance et variance) et à attribuer ces valeurs aux paramètres de la loi B pour ensuite calculer les probabilités des évènements.

Chapitre 3 : Estimations et inférences statistiques

3.1. Introduction

Considérons un échantillon de taille n extrait d'une population de taille N et pour laquelle on s'intéresse à un caractère X mesuré pour chaque individu de la population.

Le caractère X est considéré comme une variable aléatoire et l'échantillon de valeurs est constitué de n réalisations de cette variable.

On représente cette situation au moyen d'un modèle statistique qui comporte une famille de lois de probabilités parmi lesquelles se trouve la loi suivie par la variable X. Ces lois de probabilité dépendent en général d'un ou plusieurs paramètres notés θ. Dans ce cas, on dit qu'on a un modèle statistique paramétrique.

Par exemple : Pour une loi normale, les paramètres sont la moyenne m et l'écart-type σ^2. Pour une loi de Bernoulli ou une loi binomiale, c'est la probablilité p.

Un des problèmes les plus courants en statistique consiste à trouver la valeur du ou des paramètres pour la population. Mais comme on ne peut pas en général avoir l'information nécessaire, on doit se contenter des valeurs fournies par l'échantillon.

À partir de l'échantillon de valeurs, on essaie de résoudre divers types de problèmes :

1. les problèmes de test : choix entre deux éventualités dont une seule est vraie.

2. les problèmes d'estimation ponctuelle : choisir une valeur du paramètre θ.

À partir des données de l'échantillon, il faut définir une fonction (appelée aussi une statistique) dont la valeur estime θ.

3. les problèmes d'estimation ensembliste : déterminer un sous-ensemble de l'ensemble des paramètres représentant un ensemble d'éventualités. Cela conduit à la détermination d'intervalles de confiance.

On considère que chaque X_i est une variable aléatoire et on suppose qu'elles sont indépendantes entre elles. D'autre part, elles sont identiquement distribuées (puisque distribuées comme X elle-même).

En abrégé, on dit que les X_i sont i.i.d. qui est l'abréviation de *indépendantes et identiquement distribuées*.

On a donc :

$$E(X_i) = m \text{ et } Var(X_i) = \sigma^2 \ \forall \ i = 1, 2, \dots, n$$

3.2. Les estimateurs

On notera (X_1, …, X_n) l'échantillon de valeurs. Les statistiques calculées à partir de cet échantillon dépendent évidemment de l'échantillon lui-même. On dit alors que les paramètres sont empiriques puisqu'elles résultent de l'expérience (tirage, observation, mesure, etc.).

On va voir en particulier les quantités empiriques les plus couramment utilisées :

1. la moyenne empirique ;

2. la variance empirique ;

3. la fréquence empirique.

La valeur empirique d'un paramètre est également une variable aléatoire car, non seulement, il est calculé à partir d'un variable aléatoire mais aussi d'un échantillon qui lui-même aléatoirement choisi. En effet, si on avait tiré un autre échantillon, la statistique empirique aurait certainement une autre valeur. Il s'agit donc d'une valeur aléatoire et on s'intéressera à son **espérance** et à sa **variance**.

Exemple

On se propose d'estimer la taille moyenne des étudiants dans une université donnée. On a tiré au hasard 10 étudiants et en mesurant la taille moyenne parmi ces dix étudiants. Notre estimateur serait ici la moyenne de l'échantillon.

Deux questions viennent à l'esprit :

1. la taille de l'échantillon est-elle importante ? Car, intuitivement, plus l'échantillon sera grand et meilleure sera l'estimation.

2. le nombre d'échantillons tirés est-il important ? L'idée est qu'en accumulant beaucoup de valeurs de la statistique qui sert d'estimateur (en tirant beaucoup d'échantillons), à la fin, en moyenne, on aura une "bonne" estimation de la vraie valeur du paramètre qui nous intéresse.

On verra plus loin, à travers les propriétés asymptotiques, dans quelle mesure la théorie vient confirmer ces intuitions.

On peut construire beaucoup d'estimateurs différents pour estimer un paramètre donné. Certains seront considérés comme meilleurs que d'autres selon différents critères.

3.3. Propriétés d'un estimateur : espérance et variance

Un estimateur est toujours caractérisé par deux éléments : sa variance et son espérance. Ils permettent de juger de la qualité de l'estimateur notamment (sans biais et de variance minimale).

3.4. Estimateur sans biais

On dit que l'estimateur $\hat{\theta}$ est sans biais lorsque :

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

Où θ représente la vraie valeur du paramètre.

La quantité $b(\theta)$ définie telle que $b(\theta) = E(\hat{\theta}) - \theta$ s'appelle biais de l'estimateur. Si $b(\theta) \neq 0$, on dit que l'estimateur est biaisé.

Un estimateur est donc sans biais lorsque son espérance est égale à la vraie valeur du paramètre.

3.5. La moyenne empirique

La moyenne empirique de l'échantillon noté $\bar{X}$ se définit comme suit :

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

Sa valeur constitue une estimation ponctuelle de la moyenne m de la population lorsque celle-ci est inconnue.

On va calculer son espérance et sa variance.

3.5.1. Esperance

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

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

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

Etant donné que $E(X_i) = m$, alors on a :

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} m$$

$$E(\bar{X}) = \frac{nm}{n}$$

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

L'espérance de la moyenne empirique donc est la vraie moyenne de la population. On peut aussi dire que $\bar{X}$ est un estimateur sans biais de m.

3.5.2. Variance

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

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

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

Or $Var(X_i) = \sigma^2$ alors, on a :

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2$$

$$= \frac{n\sigma^2}{n^2}$$

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

La variance de la moyenne empirique est la variance de la population divisée par la taille de l'échantillon. Plus la taille de l'échantillon est grande plus la variance de l'estimateur est faible.

Par ailleurs, si l'on suppose que X suit une loi normale, alors on peut écrire ce qui suit :

$$X \rightsquigarrow N(m, \sigma^2) \Rightarrow \bar{X} \rightsquigarrow N\left(m, \frac{\sigma^2}{n}\right)$$

Et en considérant leur version centrée réduite, on a :

$$\left(\frac{X-m}{\sigma}\right) \rightsquigarrow N(0,1) \Rightarrow \left(\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}\right) \rightsquigarrow N(0,1)$$

3.6. La fréquence empirique

Considérons une variable aléatoire qui suit une loi de Bernoulli, c'est-à-dire d'une variable aléatoire X qui ne peut prendre que deux valeurs 0 (échec) ou 1 (succès). Dans cette expérience, il s'agit d'étudier la probabilité de succès. On écrit $X \rightsquigarrow B(p)$.

Dans une expérience de Bernoulli, la moyenne empirique est appelée fréquence empirique. Elle représente l'estimateur du paramètre p. On la note F_n plutôt.

$$F_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Mais puisque les valeurs de l'échantillon prennent la valeur 0 ou 1, leur somme est le nombre de fois où la valeur est 1. En divisant par n, on obtient donc la proportion des valeurs égales à 1.

$$F_n = \frac{1}{n}\sum_{i=1}^{n_1} 1_{X_i=1}$$

$$F_n = \frac{n_1}{n}$$

La fréquence empirique correspond donc à la moyenne empirique calculée sur une variable binaire 0 et 1. Sa valeur est un estimateur de la proportion de 1.

Esperance et variance de la fréquence empirique

L'espérance d'une loi de Bernoulli B(p) est égale à p et la variance à $p(1-p)$. On déduit donc des calculs sur la moyenne empirique que :

$$E(F_n) = p$$

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

F_n est donc un estimateur sans biais de la proportion p dans la population.

3.7. La variance empirique

On appelle variance empirique d'une échantillon (X_1,…,X_n) la quantité :

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

La variance empirique correspond donc à la moyenne des écarts à la moyenne empirique.

Notons qu'en développant cette expression, on retrouve une nouvelle expression dite expression développée.

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

La racine carrée de la variance empirique notée S_n correspond à l'écart-type empirique.

3.7.1. Esperance

Mais signalons-le tout de suite. La variance empirique telle que présentée ci-dessus est un estimateur biaisé de la variance de la population σ^2. En effet, calculons son espérance.

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

On sait que :

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

$$= \sum_{i=1}^{n}\left((X_i - m) - (\bar{X} - m)\right)^2$$

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

$$= \sum_{i=1}^{n}(X_i - m)^2 \; - 2(\bar{X} - m)\sum_{i=1}^{n}(X_i - m) \; + \sum_{i=1}^{n}(\bar{X} - m)^2$$

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

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

Ainsi,

$$E(S_n^2) = \frac{1}{n} E\left(\sum_{i=1}^{n} (X_i - m)^2 \ - n(\bar{X} - m)^2 \right)$$

$$E(S_n^2) = \frac{1}{n} \left(\sum_{i=1}^{n} E(X_i - m)^2 \right) - E(\bar{X} - m)^2$$

Or $\sum_{i=1}^{n} E(X_i - m)^2 \ = n\sigma^2$ et $E(\bar{X} - m)^2 = Var(\bar{X})$, alors, on a :

$$E(S_n^2) = \frac{n\sigma^2}{n} - Var(\bar{X})$$

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

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

$$E(S_n^2) \neq \sigma^2$$

On voit donc qu'à un coefficient près, l'espérance de la variance empirique est différente de la variance de la population. Cet estimateur est donc biaisé. D'où la nécessité de trouver un estimateur non biaisé. C'est là qu'intervient la notion de variance empirique modifiée.

3.7.2. Variance empirique modifiée

Soit S_n^{*2} la variance empirique modifiée. Elle se calcule comme suit :

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

On peut aisément montrer que :

$$S_n^{*2} = \left(\frac{n}{n-1} \right) S_n^2$$

Et que :

$$E(S_n^{*2}) = \sigma^2$$

Variance de la variance empirique S_n^2 :

L'expression de la variance de S_n^2 se présente comme suit :

$$Var(S_n^2) = \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\mu_2^2\right)$$

Où μ_2 et μ_4 représentent respectivement les moments centrés d'ordre 2 et 4 de l'échantillon.

Lorsque n est plus grand (n→ +∞), on a :

$$Var(S_n^2) = \frac{\mu_4 - \mu_2^2}{n}$$

Un peu plus loin, en utilisant le fait que $n\frac{S_n^2}{\sigma^2} \rightsquigarrow \chi^2(n)$, on montrera que la variance de la variance empirique peut aussi s'écrire :

$$var(S_n^2) = \frac{2\sigma^4}{n}$$

3.8. Comportement asymptotique

L'étude du comportement asymptotique des estimateurs est l'étude des propriétés probabilistes lorsque la taille des échantillons n augmente et tend vers l'infini. On cherche à savoir s'il existe une limite et comment sont distribuées les valeurs empiriques calculées.

3.8.1. Comportement de la moyenne empirique $\bar{X}$

Il y a deux résultats importants qui précisent le comportement asymptotique de la moyenne empirique lorsque n→ +∞:

- La loi des grands nombres justifie l'intuition selon laquelle plus l'échantillon est grand, plus la moyenne empirique se rapproche de l'espérance ;
- Le théorème central limite indique comment sont réparties les valeurs obtenues à partir de différents échantillons (voir chapitre 2).

3.8.2. Loi des grands nombres

Théorème 3.1. Si $\{X_i\}_{i\geq 1}$ est une suite de variables aléatoires réelles indépendantes et identiquement distribuées, alors la moyenne empirique $\bar{X}$ tend presque sûrement vers la moyenne m lorsque n→ +∞.

Ce théorème stipule donc que plus l'échantillon est grand et plus (il est probable que) la moyenne empirique se rapproche de la moyenne de la population.

Remarque :

La moyenne m est l'espérance des variables aléatoires X_i.

La notion de "*convergence presque sûre*" évoquée par ce théorème signifie que ce résultat est probabiliste presque certaine. C'est à dire qu'il y a une probabilité de presque 100% que la limite de $\bar{X}$ soit m lorsque n→ +∞.

3.8.3. Rappels sur le théorème central limite

Si $\{X_i\}_{i\geq 1}$ est une suite de variables aléatoires réelles indépendantes et identiquement distribuées avec E(Xi) = m et Var(Xi) = σ^2 pour tout i, alors la loi de probabilité de la quantité $\sqrt{n}\left(\frac{\bar{X}-m}{\sigma}\right)$ tend vers une loi normale centrée réduite lorsque n→ +∞.

$$\sqrt{n}\left(\frac{\bar{X}-m}{\sigma}\right) \rightsquigarrow N(0,1)$$

En fait, une autre manière de présenter les choses est de dire que "si n est assez grand" alors la moyenne empirique suit "approximativement" une loi normale $N\left(m,\frac{\sigma}{\sqrt{n}}\right)$.

Remarque : Le point le plus important dans ce théorème est qu'il reste valable quelle que soit la loi de probabilité suivie par les variables X_i , la seule condition étant qu'elles soient i.i.d c'est à dire aient la même loi avec une espérance et une variance finies.

3.8.4. Comportement de la variance empirique S_n^2

La loi des grands nombres (LGN) et le théorème central limite (TCL) précédemment présentés décrivent le comportement asymtotique de la moyenne empirique $\bar{X}$. On peut tout de même les utiliser pour étudier le comportement de S_n^2. On a alors les deux théorèmes suivants.

Propriétés

1. Si X est telle que $E(X^2)$ est finie, alors S_n^2 et S_n^{*2} tendent presque sûrement vers σ^2 lorsque n tend vers l'infini.

2. Si X est telle que E(X^4) est finie, alors la quantité $\sqrt{n}\left(\frac{S_n^2-\mu_2}{\sqrt{\mu_4-\mu_2^2}}\right)$ converge en loi vers la loi normale N(0; 1)

Toutefois, il faut ajouter deux théorèmes supplémentaires sur la variance empirique modifiée qui est très important dans les tests statistiques.

3. Dans le cas d'un échantillon gaussien (loi normale), la moyenne empirique $\bar{X}$n et la variance empirique S_n^{*2} sont des variables aléatoires indépendantes l'une de l'autre.

4. Si (X_1; … ;X_n) est un échantillon de variables gaussiennes (loi normale), alors les variables $\sqrt{n}\left(\frac{\bar{X}-m}{\sigma}\right)$ et $(n-1)\frac{S_n^{*2}}{\sigma^2}$ sont indépendantes et suivent respectivement la loi normale N(0; 1) et la loi de khi-deux à n-1 degrés de liberté.

Comme on a : $S_n^{*2} = \left(\frac{n}{n-1}\right)S_n^2$, la propriété est aussi vraie pour $n\frac{S_n^2}{\sigma^2}$ qui suit une loi de khi-deux à n degrés de liberté.

Une conséquence directe de ce théorème est de donner une expression simple de la variance de S_n^{*2} dans le cas d'un échantillon gaussien. En effet, puisque $(n-1)\frac{S_n^{*2}}{\sigma^2} \rightsquigarrow \chi^2(n-1)$, on a (sachant que la variance de la loi du 2 vaut deux fois le nombre de degrés de liberté) :

$$Var\left((n-1)\frac{S_n^{*2}}{\sigma^2}\right) = 2(n-1)$$

$$\frac{(n-1)^2}{\sigma^4}var(S_n^{*2}) = 2(n-1)$$

Ainsi, en tirant S_n^{*2}, on a :

$$var(S_n^{*2}) = \frac{2\sigma^4}{(n-1)}$$

De même pour S_n^2

$$Var\left(n\frac{S_n^2}{\sigma^2}\right) = 2n$$

$$\frac{n^2}{\sigma^4}var(S_n^2) = 2n$$

$$var(S_n^2) = \frac{2\sigma^4}{n}$$

3.9. Estimations par intervalles de confiance

3.9.1. Intervalles de confiance de la moyenne

Exemple introductif

Dans une chaîne de fabrication de conserves de légumes, les boîtes de conserve sont remplies par une machine. Le directeur de l'usine souhaiterait contrôler le poids des boîtes à la sortie de la chaîne de fabrication.

Contexte statistique :

- Population : ensemble des boîtes de conserve produites par la chaîne de fabrication ;
- Individu : une boîte de conserve produite par la chaîne de fabrication ;
- Variable : poids X ;
- m = poids moyen des boîtes de conserve produites par la chaîne de production.

L'objectif du directeur est de déterminer une estimation la valeur de m à partir d'un échantillon de X en se donnant un intervalle de valeurs possible avec un certain niveau de confiance.

Etant donnée une valeur m, déterminer un intervalle dépendant d'un échantillon $X_1, \dots, X_n$ de X tel que

$$P\big(m \in IC(X_1, \dots, X_n)\big) = 1 - \alpha$$

Où α est appelé risque d'erreur. Et $1 - \alpha$ est appelé seuil de confiance. $IC(X_1, \dots, X_n)$ représente l'intervalle de confiance pour la moyenne m. Il dépend de l'échantillon choisi.

3.9.1.1. Cas où la variance de la population est connue

Supposons que le poids X d'une boîte de conserve produite par la chaîne de production soit une variable aléatoire continue distribuée selon une loi normale de moyenne m et d'écart-type $\sigma = 25$. X$\rightsquigarrow$ N(m,25).

L'objectif est de proposer une estimation par intervalle de confiance de m à 95% de confiance.

D'abord, il faut calculer la moyenne empirique à partir de l'échantillon $\bar{X}$ qui est une estimation ponctuelle de m.

Le but étant de déterminer $P\big(m \in IC(X_1, \dots, X_n)\big) = 1 - \alpha$. Le point de départ est de chercher un intervalle J tel que $P(\bar{X} \in J) = 1 - \alpha$

On sait d'une part que

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

Et d'autre part, on sait que si $X \rightsquigarrow N(m, \sigma^2)$ alors $\bar{X} \rightsquigarrow N\left(m, \frac{\sigma^2}{n}\right)$. Ce qui équivaut à écrire que $X \rightsquigarrow N(m, \sigma) \Rightarrow \bar{X} \rightsquigarrow N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

Ainsi, on peut écrire :

$$\bar{X} \rightsquigarrow N\left(m, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \left(\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}\right) \rightsquigarrow N(0,1) \Leftrightarrow \sqrt{n}\left(\frac{\bar{X} - m}{\sigma}\right) \rightsquigarrow N(0,1) = Z_{1-\frac{\alpha}{2}}$$

$$P\big(m \in IC(X_1, \dots, X_n)\big) = P\left(-Z_{1-\frac{\alpha}{2}} \leq \sqrt{n}\left(\frac{\bar{X} - m}{\sigma}\right) \leq Z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$= P\left(-Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq \bar{X} - m \leq Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$= P\left(\bar{X} - Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\big(m \in IC(X_1, \dots, X_n)\big) = P\left(\bar{X} - Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Ainsi

$$IC \ = \ \bar{X} \pm Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

Où $Z_{1-\frac{\alpha}{2}}$ est le quantile de la loi normale pour un niveau α donné. Par exemple pour $\alpha = 5\%$ (0,05), on a $Z_{1-\frac{\alpha}{2}} = Z_{0,975} = 1{,}96$

En conclusion, lorsque la variance est connue, l'intervalle de confiance de la moyenne m est calculé en fonction de la moyenne empirique et du quantile de la loi normale.

3.9.1.2. Cas où la variance de la population n'est pas connue

Lorsque la variance de la population n'est pas connue, pour proposer une estimation de l'intervalle de confiance de la moyenne, on se sert à la fois de la moyenne empirique et de la variance empirique.

Pour rappel, la variance empirique (modifiée) se présente comme suit (en omettant l'Astérix sur le S) :

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

Sachant que les $X \rightsquigarrow N(m, \sigma^2)$ alors S_n^2 sera distribuée selon une loi de Khi-deux (car selon la définition c'est la somme des carrés d'une loi normale). En tenant compte de cette situation, on a alors ce qui suit :

$$X \rightsquigarrow N(m, S_n^2) \Rightarrow \bar{X} \rightsquigarrow N\left(m, \frac{S_n^2}{n-1}\right)$$

$$\Leftrightarrow$$

$$X \rightsquigarrow N(m, S_n) \Rightarrow \bar{X} \rightsquigarrow N\left(m, \frac{S_n}{\sqrt{n-1}}\right)$$

Ainsi, on a :

$$\bar{X} \rightsquigarrow N\left(m, \frac{S_n}{\sqrt{n-1}}\right) \Rightarrow \left(\frac{\bar{X} - m}{\frac{S_n}{\sqrt{n-1}}}\right) \rightsquigarrow T(n-1)$$

Cette relation montre qu'en centrant et en réduisant la moyenne empirique respectivement par sa moyenne et son écart-type, on obtient une loi de Student (car il s'agit du rapport entre une loi normale et la racine carrée d'une loi de khi-deux, voir définition des lois au chapitre 2). Ainsi, on a

$$\sqrt{n-1}\left(\frac{\bar{X} - m}{S_n}\right) \rightsquigarrow T(n-1) = t_{1-\frac{\alpha}{2}}(n-1)$$

Où $t_{1-\frac{\alpha}{2}}(n-1)$ est le quantile de la loi Student pour un niveau α donné avec le nombre de degré de liberté égale à n-1. Par exemple pour $\alpha = 5\%$ (0,05) et pour une taille d'échantillon n=500, alors on a $t_{1-\frac{\alpha}{2}}(n-1) = t_{0,975}(199) = 1{,}96$.

Notons que lorsque la taille de l'échantillon est suffisamment importante, la loi de Student converge vers la loi normale.

3.9.2. Intervalle de confiance d'une proportion

Exemple

Un maire s'inquiète au sujet de sa réélection lors des prochaines élections municipales. On choisit au hasard 426 électeurs que l'on interroge sur leurs intentions de vote en faveur du Maire sortant ; 201 des 426 électeurs déclarent avoir l'intention de voter pour le maire sortant lors des prochaines élections municipales. On se propose alors de calculer un intervalle de confiance pour la proportion de votants pour le Maire sortant lors des échéances électorales à seuil de confiance de 95%.

Dans cet exemple, la population étudiée est l'ensemble des votants de la commune. Et la variable étudiée est de type Bernoulli tel que $X = 1$ lorsque le votant déclare vouloir voter pour le Maire (Succès) et $X = 0$ si le votant déclare ne pas vouloir voter pour lui (Echec).

$$X \rightsquigarrow B(p)$$

Le paramètre de cette loi est p la proportion dans la population qui reste inconnue. Il faut alors partir de la proportion empirique et utiliser les propriétés de convergence vers la loi normale.

La proportion empirique est :

$$F_n = \frac{1}{n}\sum_{i=1}^{n_1} 1_{X_i=1} = \frac{n_1}{n}$$

Avec $E(F_n) = p$ et $Var(F_n) = \frac{p(1-p)}{n}$ où n_1 est le nombre de succès, n la taille de l'échantillon et p est la proportion dans la population.

En utilisant la loi des grands nombres, on peut écrire :

$$\left(\frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}\right) = \sqrt{n}\left(\frac{F_n - p}{\sqrt{p(1-p)}}\right) \rightsquigarrow N(0,1)$$

Ainsi, à un risque d'erreur , on peut alors définir l'intervalle de confiance de p tel que :

$$P\big(p \in IC(X_1, \dots, X_n)\big) = P\left(-Z_{1-\frac{\alpha}{2}} \leq \sqrt{n}\left(\frac{F_n - p}{\sqrt{p(1-p)}}\right) \leq Z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$= P\left(-Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}} \leq F_n - p \leq Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

$$= P\left(F_n * \sqrt{\frac{p(1-p)}{n}} \leq p \leq F_n + Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

$$P\big(p \in IC(X_1, \dots, X_n)\big) = P\left(F_n - Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}} \leq p \leq F_n + Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}}\right)$$
$$= 1-\alpha$$

Ainsi

$$IC \ = \ F_n \pm Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}}$$

Où $Z_{1-\frac{\alpha}{2}}$ est le quantile de la loi normale pour un niveau α donné.

Application numérique :

$F_n = \sqrt{\frac{201}{426}} = 0{,}472$; $\alpha = 5\%$, alors, on a $Z_{1-\frac{\alpha}{2}} = Z_{0{,}975} = 1{,}96$. Ainsi

$$IC \ = \ 0{,}472 \ \pm 1{,}96 * \sqrt{\frac{0{,}472(1-0{,}472)}{426}}$$

$$IC \ = [0{,}425 \, ; 0{,}519]$$

L'intention de vote pour le Maire sortant étant compris entre 42,25% et 51,19%, cela montre qu'il n'est pas sûr qu'il soit réélu.

3.10. Exercices d'application des notions abordées dans le chapitre

Exercice 1

Une machine fabrique des comprimés en grande série. On considère la population constituée de tous les comprimés produits en une journée. Pour étudier le caractère « poids du comprimé » sur cette population, on prélève au hasard et de manière non exhaustive un échantillon de 10 comprimés que l'on pèse. On obtient les résultats suivants :

0,79	0,80	0,81	0,79	0,84	0,78	0,79	0,80	0,83	0,81

1) Calculer la moyenne $\bar{x}$ et l'écart-type s de cet échantillon.
2) Donner une estimation ponctuelle de la moyenne m et de l'écart-type σ du poids des comprimés.
3) Soit $\overline{X}$ la variable aléatoire qui, à chaque échantillon, de taille constante n, tiré au hasard et de manière non exhaustive de la population des comprimés fabriqués par cette machine associe la moyenne des poids des comprimés de l'échantillon. On suppose que l'écart-type de la population est de 0,02g.

Quelle doit être la taille de l'échantillon pour que la variable $\overline{X}$ ait un écart-type égal à deux milligrammes ?

Corrigé :

1) $\bar{x}$ = 0,80g et s = 0,018g
2) $\bar{x}$ est une estimation ponctuelle de m ; s est une estimation ponctuelle de σ.

   Une autre estimation ponctuelle de σ est $\sqrt{\frac{n}{n-1}}s$ = 0,02g.
3) On sait que la variable $\overline{X}$ a un écart-type σ ($\overline{X}$) égal à $\frac{\sigma}{\sqrt{n}}$. Comme on veut que cet écart-type soit égal à 0,002, on obtient $\sqrt{n} = \frac{0,02}{0,002}$ = 10. D'où n doit être égal à 100.

Exercice 2

Un atelier fabrique des pièces cylindriques. La variable aléatoire qui associe à chaque pièce son diamètre suit une loi normale de moyenne m = 15mm et d'écart-type σ = 0,35mm.

On prélève un échantillon non exhaustif de 200 pièces.

1) Calculer la probabilité que la moyenne de l'échantillon soit comprise entre 14,95mm et 15,05mm.

2) Quel devrait être l'effectif de l'échantillon pour que sa moyenne soit comprise entre les limites précédentes avec une probabilité au moins égale à 99% ?

Corrigé :

1) L (X) = N(15, 0,35) donc la loi de $\overline{X}$ est une N(15, 0,35/√200)

On fait d'abord une transformation de $\overline{X}$ selon le théorème Central-limite

$$U = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$$

Mais comme P (14,95 < $\overline{X}$ < 15,05) alors, en transformant on a :
P (14,95 < $\overline{X}$ < 15,05) = P (-2 < U < +2) = 0,95

Mais on sait que $P\ (a < U < b) = P(U < b) - P(U < a)$
Dans le cas d'une loi symétrique comme la loi normale on a souvent: $P\ (-b < U < b)$
D'où $P\ (-b < U < b) = P(U < b) - P(U < -b)$
Or $P(U < -b) = P(U > b)$ propriété loi symétrique
Mais on sait aussi que $P(U > b) = 1 - P(U < b)$.
D'où finalement $P\ (-b < U < b) = P(U < b) - [\ 1 - P(U < b)]$

Ainsi $P\ (-b < U < b) = 2P(U < b) - 1$

$$P\ (-2 < U < +2) = 2P(U < +2) - 1$$

Cherchons dans la table de la loi normale centrée et réduite P(U<+2)
Cela correspond à 0,978. On peut aussi utiliser la fonction excel:
=LOI.NORMALE.STANDARD.N(2;VRAI)

$$P\ (-2 < U < +2) = 2P(U < +2) - 1 = 2 * 0{,}978 - 1 = 0.95$$

$$P\ (14{,}95 < \overline{X} < 15{,}05) = P\ (-2 < U < +2) = 0{,}95$$

La probabilité est égale à 95%.

2) Ici la probabilité est fixée à 99%, le but ici est de formuler U, ensuite de déterminer les fractiles correspondant à 99% et d'égaliser la borne sup à la fractile de droite (on peut aussi égaliser la borne inf à la fractile de gauche, puisque la loi normale est symétrique).

D'abord, c'est le fractile $t_{\alpha/2}$ qu'on cherche d'abord. Puisque $1-\alpha = 99\%$ alors $\alpha = 1\%$ Ce qui veut dire que $\alpha/2 = 0.5\% = 0.005$
Ainsi la fractile à lire dans la table est 99% +0.5%=0.99+0.005=0.995
Ou bien on utilise la fonction excel :
=LOI.NORMALE.STANDARD.INVERSE(0,995)
$t_{0.995}$=2.5758.

En égalisant U à la borne sup on a :

$$\frac{14.95 - \mathrm{m}}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq \frac{15{,}05 - \mathrm{m}}{\frac{\sigma}{\sqrt{n}}}$$

P (14,95 < $\overline{X}$ < 15,05) = 0,99 = P (-2, 57 < U < +2, 57)

$$\frac{15{,}05 - \mathrm{m}}{\frac{\sigma}{\sqrt{n}}} = 2.57 \text{ ou bien } \frac{14.95 - \mathrm{m}}{\frac{\sigma}{\sqrt{n}}} = -2.57$$

On utilise l'une de ces deux équations pour tirer n. Par exemple, en égalisant la borne supérieure de U à 2.57, on obtient n=325

Exercice 3

Une étude préalable a montré que, dans une production en grande série, une machine fabrique des joints d'étanchéité avec un pourcentage p = 2% de joints défectueux.

Une entreprise de plomberie commande 400 joints. On supposera que cette commande correspond à un échantillon non exhaustif.

1) Par quelle loi peut-on approcher la loi suivie par la variable F de la distribution d'échantillonnage ?
2) Quelle est la probabilité pour que, dans cet envoi, on trouve au plus 3% de joints défectueux ?

Corrigé :

F a pour espérance : E (F) = p = 0,02 et pour écart-type : σ = $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}02 \times 0{,}98}{400}} = 0{,}007$

Puisque la taille de l'échantillon est supérieure à 30, nous pouvons approcher la variable F par une variable normale de paramètres 0,02 et 0,007. On cherche la probabilité que F soit au plus égale à 0,03

$$\Pr(F \leq 0{,}03) = \Pr(\frac{F - 0{,}02}{0{,}007} \leq \frac{0{,}03 - 0{,}02}{0{,}007}) = \Pr(\frac{F - 0;02}{0{,}007} \leq 1{,}43) = 0{,}9236 \text{ .}$$

Il y a donc 92 chances sur 100 pour que la commande contienne moins de 3% de joints défectueux.

La correction de continuité, dont on peut se poser la question de savoir si elle est utile ici, est de 0,5 pour 400×F c'est à dire que la longueur du pas pour F est de 0,0012 ce qui est tout à fait négligeable.

Exercice 4

Dans un établissement de construction mécanique, une machine produit en série des tiges métalliques dont la longueur X présente des imperfections dues au fonctionnement de la machine. Il est admis que X est une variable gaussienne d'espérance m et d'écart-type σ = 2 mm.

1) On a mesuré 100 tiges et on a trouvé une moyenne empirique $\overline{X} = 23{,}65mm$. Construire un intervalle de confiance pour m au niveau 95%.
2) Pour un même niveau, on souhaite réduire la largeur de l'intervalle de confiance trouvé dans la question précédente en choisissant un échantillon de taille supérieure. En souhaitant une largeur d'intervalle de 0,5 mm, quelle doit être la taille du nouvel échantillon ?
3) Une tige est utilisable si sa longueur X est comprise entre 22,65 mm et 24,65 mm. En supposant que m = 23,65 mm, calculer la probabilité qu'une tige soit acceptable.

Corrigé :

1) Comme n est suffisamment grand, la loi de $\overline{X}$ est N (m, $\frac{\sigma}{\sqrt{n}}$). L'intervalle de confiance recherché est donc :

$$\overline{X} - \frac{1{,}96\sigma}{\sqrt{n}} < m < \overline{X} + \frac{1{,}96\sigma}{\sqrt{n}} \text{ soit} \qquad 23{,}258 < \text{m} < 24{,}042$$

2) La largeur de l'intervalle de confiance pour m doit être au plus égale à 0,5. La largeur se calcule comme la différence entre les deux bornes (sup-inf). Donc :

$$2 \times 1{,}96 \times \frac{\sigma}{\sqrt{n}} = 0{,}5 \text{ c'est à dire } n \cong 250$$

3) Il s'agit de calculer : P (22.65 < X< 24.65)=?

On centre et réduit:

$$\text{P}\left(\frac{22.65 - \text{m}}{\sigma} < \frac{X - m}{\sigma} < \frac{24.65 - \text{m}}{\sigma}\right)$$

Nb : On prend σ et $\frac{\sigma}{\sqrt{n}}$ parce qu'il s'agit de la loi de X et non la loi de $\overline{X}$

Ainsi on a : $\left|\frac{X - m}{\sigma}\right| < 0{,}5mm$

$$P\left(\left|\frac{X - m}{\sigma}\right| < 0{,}5\right) = P(|U| < 0{,}5) = P(-0{,}5 < U < 0{,}5) = 2P(U < 0{,}5) - 1$$

Dans la table (ou avec Excel la fonction =LOI.NORMALE.STANDARD.N(0,5;VRAI)), on a : $P(U < 0{,}5) = 0{,}691462461$

D'où $P(\left|\frac{X-m}{\sigma}\right| < 0{,}5) = 2 * 0{,}69146246 - 1 \cong 0{,}38$

Exercice 5

On suppose que la durée de vie d'ampoules électriques suit une loi normale d'écart-type 100 heures.

Quelle est la taille minimum de l'échantillon à prélever pour que l'intervalle de confiance, à 95%, de la durée de vie moyenne ait une longueur inférieure à 20 heures.

Corrigé :

La durée de vie moyenne m est estimée par $\overline{X}$, durée de vie moyenne sur l'échantillon.

Si n est suffisamment grand, la loi de $\overline{X}$ est N (m, $\frac{\sigma}{\sqrt{n}}$).

Sachant la formule de l'intervalle de confiance $\overline{X} \pm 1{,}96 \times \frac{\sigma}{\sqrt{n}}$. En faisant la différence entre les deux bornes, on retrouve la largeur de l'intervalle de confiance pour m doit être au plus donc égale à : $2 \times 1{,}96 \times \frac{\sigma}{\sqrt{n}} = 20$

Ensuite, on tire n :

n >= 385.

Exercice 6

Une machine fabrique des engrenages en grande série. Les mesures des diamètres de 200 engrenages issus d'un échantillon non exhaustif pris dans la fabrication journalière de la machine ont une moyenne de 0 ,824 cm et un écart-type de 0,042 cm.

Déterminer un intervalle de confiance au risque 5% de la moyenne des diamètres des engrenages.

Corrigé :

Avec le théorème central limite, nous savons que la loi de la moyenne de l'échantillon est :

$\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}$ suit une loi normale 0 et 1

Mais comme σ est inconnu, on utilise le deuxième théorème central limité fondé sur la variance empirique : $\frac{\overline{X}-m}{\frac{S}{\sqrt{n-1}}}$ qui suit une loi de student à n-1 degrés de libertés.(notons que : $S^2=\frac{1}{n-1}\sum(X_i-m)^2$

Ceci étant, la mesure du diamètre suit une loi normale de paramètres m et σ, m et σ étant inconnus.

On sait alors que la variable $T=\frac{\overline{X}-m}{s/\sqrt{n-1}}$ suit une loi de Student à (n-1) degrés de liberté.

Application numérique : ici $\overline{X}=0.824; S=0.042$. Ainsi :

$$L(\frac{0,824-m}{0,042/\sqrt{199}})=T_{199}$$

Grâce aux tables de la loi de Student (et même avec la table de la loi normale de Laplace-Gauss car en fait n très grand), on sait que :

$$P(\left|\frac{0,824-m}{0,042/\sqrt{199}}\right|<1,96)=95\%$$

Pour trouver la valeur, on lit dans la table de student où on utilise la fonction excel suivante : =LOI.STUDENT.INVERSE(0,05;199)
Si on veut utiliser la loi normale de Laplace-Gauss, on fait :
=LOI.NORMALE.STANDARD.INVERSE(0,975)
Remarque sur les fonctions excel : dans la loi de student, on spécifie uniquement le seuil α pour déterminer la fractile. . Alors que pour la loi normale, on addition la moitié de α au seuil de confiance, c'est-à-dire : $\frac{\alpha}{2}+(1-\alpha)$. Ici 0.975=(0.05/2)+ 0.95

On en déduit l'intervalle pour m : $|0,824-m|<1,96\times\frac{0,042}{\sqrt{199}}$

$$0,818<m<0,830$$

Exercice 7

On dispose de 10 prises de sang recueillies dans les mêmes conditions chez un même sujet. On obtient pour chacune un dosage du cholestérol en grammes :

2,45	2,48	2,50	2,47	2,47	2,49	2,47	2,46	2,46	2,48

Chaque mesure peut être considérée comme une réalisation particulière d'une variable aléatoire X :« taux de cholestérol », suivant une loi normale N (m, σ).

1) Donner un intervalle de confiance pour m (seuil de confiance 5%).
2) Donner un intervalle de confiance pour σ^2 (même seuil).

Corrigé :

On obtient les résultats suivants sur l'échantillon : $\bar{x} = 247{,}3\,cg$, s = 1,418 cg et s* = 1,494 cg

1) X suit une loi normale de paramètres m et σ inconnus donc $L(U = \frac{\overline{X} - m}{s/\sqrt{n-1}}) = T_{n-1}$

Par lecture de la table de la loi de Student avec n = 10 (ou avec excel : =LOI.STUDENT.INVERSE(0,05;9), on trouve 2.26. on a donc: $P(|U| < 2{,}26) = 95\%$

D'où l'intervalle sur m : $|247{,}3 - m| < 2{,}26 \times 1{,}418/3$ $\boxed{246{,}33 < m < 248{,}37}$

2) $L(\frac{nS^2}{\sigma^2}) = \chi^2_{n-1}$. La table de la loi χ^2_9 nous donne : $P(2{,}7 < \frac{nS^2}{\sigma^2} < 19{,}023) = 95\%$

Attention : la lecture de la table se fait de manière différente par rapport à d'autres tables puisque les deux bornes ne sont pas symétriques. Il faut donc définir deux probabilités : l'une correspondant à la borne inférieure et l'autre correspondant à la borne supérieure. A chacune de ces probabilités correspond donc une fractile dans la table de Student. Ici avec le seuil $\alpha = 0.05$, on a $\frac{\alpha}{2} = 0.025$.

Pour lire la table de khi-deux, la première borne est égale à $\frac{\alpha}{2} = 0.025$. En lisant le fractile correspondant (à 9 degrés de libertés), on trouve 2.70. Avec excel : on fait

=LOI.KHIDEUX.INVERSE(0,025;9)

La seconde borne se calcule comme pour la loi normale en faisant :

$\frac{\alpha}{2} + (1 - \alpha)$

Ce qui donne =0,025 + 0,95= 0,975. La fractile correspondante est 19.023. Ce qui se retrouve aussi avec excel avec la fonction :

=LOI.KHIDEUX.INVERSE(0,975;9) . Ces deux bornes, on peut maintenant encadrer. D'où l'expression :

$$P(2{,}7 < \frac{nS^2}{\sigma^2} < 19{,}023) = 95\%$$

On en déduit donc que : $2{,}7 < \frac{nS^2}{\sigma^2} < 19{,}023$. Et ainsi on a l'intervalle de confiance pour σ^2 :

$\boxed{1{,}06 < \sigma^2 < 7{,}44}$

Exercice 8

Lors d'un contrôle qualité sur une population d'appareils électro-ménagers d'un certain type, au cours d'un mois de fabrication, on prélève d'une manière non exhaustive un échantillon de 1000 appareils. Après un test de conformité, on constate que 60 appareils ont un défaut.
Donner un intervalle de confiance du pourcentage p d'appareils défectueux, au risque de 5%.

Corrigé :

Le pourcentage d'appareils défectueux dans l'échantillon, estimation ponctuelle de p, est f = 0,06.

n est suffisamment grand pour que l'approximation par une loi normale soit justifiée. La loi de f est donc N (p , $\sqrt{\frac{p(1-p)}{n}}$). On a donc Pr (-1,96 < $\frac{f-p}{\sqrt{\frac{p(1-p)}{n}}}$ < +1,96) = 0,95. L'intervalle de confiance, à 95%, à risques symétriques, est:

$$[\, f - 1{,}96\sqrt{\frac{p(1-p)}{n}}, f + 1{,}96\sqrt{\frac{p(1-p)}{n}}\,]$$

Le paramètre à estimer apparaît dans les bornes de l'intervalle ! Plusieurs méthodes sont proposées :

- On donne à p la valeur 0,5 qui rend maximum p(1-p). L'intervalle est alors : [0,06-1,96×0,022 , 0,06+1.96 × 0,022] = [0,017 , 0,103]
- On donne à p l'estimation 0,06 : [0,06 – 1,96 × 0,0075, 0,06 +1,96 × 0,0075] = [0,045 , 0,075]
- On résout l'inéquation : $(f-p)^2 \le (1{,}96)^2 \frac{p(1-p)}{n}$ on obtient [0,046 , 0,078].

Exercice 9

Afin d'étudier le salaire horaire, en francs, des ouvriers d'un secteur d'activité, on procède à un tirage non exhaustif d'un échantillon de taille n = 16. On a les résultats suivants :

41	40	45	50	41	41	49	43	45	52	40	48	50	49	47	46

1) Proposer des estimateurs non biaisés de la moyenne et de la variance.
Calculer leurs valeurs pour cet échantillon.
2) On suppose que la loi suivie par le salaire horaire est une loi normale.
   - Déterminer un intervalle de confiance à 95% pour la moyenne.
   - Déterminer un intervalle de confiance à 95% pour la variance.

Corrigé :

1) Estimateur de m : $\bar{x}$ = 45,44F Estimateur non biaisé de σ^2 : $S^{*2} = \frac{1}{n-1}\sum(X_i - \overline{X})^2 = 16{,}26$

2)

- Intervalle de confiance pour la moyenne :

$(\overline{X} - m) \times \frac{\sqrt{n}}{S^*}$ suit une loi de Student à (n-1) degrés de liberté. La table donne : $t_{0,975}(15) = 2{,}131$

D'où l'intervalle pour m : $[45{,}44 - 2{,}131 \times 4{,}03 \times \frac{1}{4}\ ,\ 45{,}44 + 2{,}131, \times \frac{1}{4}] = [43{,}28\ ,\ 47{,}59]$

- Intervalle de confiance pour la variance :

$\frac{nS^2}{\sigma^2}$ suit une loi du χ^2 à (n-1) degrés de liberté. La table de la loi du χ^2 nous donne :

Pr $(6{,}26 < \chi^2(15) < 27{,}50) = 0{,}95$. D'où l'intervalle: $\frac{15 \times 16{,}26}{27{,}50} < \sigma^2 < \frac{15 \times 16{,}26}{6{,}26}$

$[8{,}87\ ,\ 38{,}97]$

Exercice 10

Dans un échantillon pris au hasard de 100 automobilistes, on constate que 20 d'entre eux ne mettent pas leurs ceintures de sécurité. Donner un intervalle de confiance à 99% pour la proportion d'automobilistes qui ne mettent pas leurs ceintures. (On explicitera différentes méthodes).

Corrigé :

Nous cherchons un intervalle de confiance sur une proportion. Ici la proportion p d'automobilistes sans ceinture est inconnu et sur un échantillon de taille n = 100, on a une estimation f = 0,20.

La taille de l'échantillon est suffisamment grande pour que l'on puisse dire que :

$$L\left(\frac{f-p}{\sqrt{\frac{p(1-p)}{n}}}\right) = LG(0,1)$$

$$P\left(\left|\frac{f-p}{\sqrt{\frac{p(1-p)}{n}}}\right| < 2,5758\right) = 99\% \qquad |f-p| < 2,5758\sqrt{\frac{p(1-p)}{100}}$$

Cette équation se résout de plusieurs façons :

- On remplace p par 0,5 sous la racine carrée : $|f-p| < 0,129$
- On remplace p par f = 0,2 sous la racine carrée : $|f-p| < 0,103$
- On résout l'inéquation du second degré : $(f-p)^2 < (2,5758)^2 \times \frac{p(1-p)}{100}$

$f^2 - 2fp + p^2 < 0,066p(1-p)$ $\quad$ $1,066p^2-0,466p+0,04 < 0$…

Exercice 11

Pour déterminer le poids moyen d'épis de blé appartenant à une variété particulière, on a procédé à 10 pesées réalisées sur des épis tirés au hasard. On suppose que le poids des épis appartenant à cette variété est une variable aléatoire suivant une loi normale de moyenne m et d'écart-type σ, ces deux paramètres étant inconnus.

1) Les observations sont les suivantes :

194,46	183,16	171,57	177,38	155,37	205,61	171,24	207 ,73	175,54	188,30

- Donner un intervalle de confiance au niveau 95% pour la moyenne m.
- Donner un intervalle de confiance au niveau 95% pour la variance σ^2.

2) La proportion d'utilisateurs de cette variété dans la région était égale à 15%. Soit $\hat{p}_n$ la proportion d'utilisateurs de cette variété parmi n agriculteurs.
   - Déterminer, à partir du théorème central limite, le nombre minimal d'agriculteurs que l'on doit interroger pour que :

Pr ($|\hat{p}_n$ - p|≤ 0,01) > 0,95

- A la suite d'une campagne publicitaire, on a constaté que, sur un échantillon de 5000 agriculteurs, 1125 utilisent la variété considérée. Donner un intervalle de confiance au niveau 95% pour la nouvelle proportion p* d'utilisateurs de la variété dans la région.
- Peut-on dire, au risque de 5% de se tromper, que la publicité a influencé les agriculteurs ?

Corrigé :

Les calculs de la moyenne et de l'écart-type donnent : $\bar{x} = 183{,}04$ et s = 15,50

- Intervalle de confiance pour la moyenne m : $L(\frac{\overline{X} - m}{s/\sqrt{n-1}}) = T_{n-1}$

La table de Student nous donne $t_{0,05}(9)$=2,262 et l'intervalle de confiance :

$183{,}04 - 2{,}262 \times \frac{15{,}5}{3} < m < 183{,}04 + 2{,}262 \times \frac{15{,}5}{3}$ 171,35 < m < 194,73

- Intervalle de confiance pour la variance : $L(\frac{nS^2}{\sigma^2}) = \chi^2_{n-1}$

$P(2{,}7 < \frac{nS^2}{\sigma^2} < 19) = 0{,}95$ l'intervalle de confiance pour σ^2 : $126{,}45 < \sigma^2 < 889{,}81$

p = 15%.

- Le théorème central-limite donne : $L(\hat{p}_n) \cong LG(p, \sqrt{\frac{p(1-p)}{n}})$

$|\hat{p}_n - p| < 1{,}96 \times \sqrt{\frac{0{,}15 \times 0{,}85}{n}} \leq 0{,}01$ On en déduit : $n > 4898$

- Intervalle pour la proportion :

$\left|\hat{p}^*_n - p^*\right|^2 < 1{,}96^2 \times \frac{p^* \times (1-p^*)}{5000}$

Pour résoudre cette inéquation on peut soit remplacer p* par 0,5 dans le second membre soit remplacer p* par 0,225 dans le second membre soit résoudre l'inéquation :

$5003{,}8p^{*2} - 2253{,}8p^* + 253{,}12 < 0$ 0,213<p*<0,237

- Au risque 5%, on peut conclure à l'influence de la publicité puisque p* > 0,2 alors que, avant la publicité, p était égal à 15%.

Chapitre 4 : Les tests d'hypothèses statistiques

4.1. Généralités sur les tests d'hypothèses

La plupart des problèmes de statistique sont des problèmes de décision ou de choix. On dispose d'un modèle statistique dans lequel les variables suivent des lois de probabilité. Ce modèle est en général un modèle d'échantillonnage car on travaille sur un échantillon plutôt que sur la population entière.

Le choix (ou la décision), dans le cas des tests d'hypothèses, consiste à déterminer, parmi deux éventualités, laquelle est la bonne. On formule une éventualité et on cherche à décider s'il faut l'accepter ou la rejeter.

C'est un problème de prise de décision et il faut donc disposer d'une règle de décision.

Les lois de probabilités du modèle statistique ne sont pas forcément connues.

On distingue donc deux situations :

1. si on connaît les lois suivies par l'échantillon extrait de la population, on réalise des tests paramétriques. Cela provient du fait que les lois sont souvent définies en fonctions de paramètres (comme par exemple le m et le σ^2 d'une loi normale et que la formulation du test porte sur un de ces paramètres.

2. si on ne connaît pas les lois de l'échantillon, on parle de test non-paramétrique. La procédure du test ne repose alors pas sur des propriétés de lois de probabilité.

4.1.1. Formulation de l'hypothèse d'un test

Afin de pouvoir décider entre plusieurs hypothèses possibles, on met en avant une hypothèse particulière que l'on appelle l'hypothèse nulle (notée H0).

On formule aussi une hypothèse alternative qui est notée H1.

Souvent l'hypothèse H1 est le contraire de l'hypothèse H0 mais ce n'est pas nécessairement le cas. Il arrive que l'hypothèse H1 soit plus restrictive.

Par exemple, si l'hypothèse H0 est

$$H_0: a = b$$

l'hypothèse H1 pourrait être

$$H_1: a \neq b$$

$$H_1: a < b$$

$$H_1 : a > b$$

Dans le premier cas, on parle de test bilatéral et dans les deux autres cas de test unilatéral (à gauche ou à droite).

Le fait que les hypothèses H0 et H1 ne soient pas l'opposé l'une de l'autre peut sembler surprenant mais il faut bien comprendre deux choses :

1. le choix fait par le test est basé sur l'hypothèse H0 : il consiste à décider s'il faut rejeter ou accepter H0 et non pas à choisir entre H0 et H1.

2. l'hypothèse H1 sert à formuler la règle de décision du test.

Le résultat d'un test est "rejeter H0" ou bien "ne pas rejeter H0". On ne conclut jamais par "rejeter H1" et encore moins par "accepter H1".

4.1.2. Risques d'erreur

On dispose en général d'une information insuffisante puisqu'on a des informations sur un échantillon seulement et non pas sur toute la population.

La prise de décision encourt donc un double risque d'erreur:

- On peut décider que H0 est fausse alors qu'elle est vraie. C'est le risque de première espèce, noté α.
- On peut décider que H0 est vraie alors qu'elle est fausse. C'est le risque de deuxième espèce, noté β.

Le risque α intéresse l'utilisateur du test : pour lui, H0 est rejetée ou pas au risque α. Le risque est le risque de rejeter à tort. Il est courant de fixer α = 0,05 (5%) ou α = 0,01 (1%).

Le risque β intéresse le concepteur du test : c'est le risque d'accepter à tort.

Par conséquent, pour lui, 1- β représente la puissance du test puisque c'est le risque d'accepter à juste titre.

Le tableau suivant résume la situation :

Décision	**H0 Vraie (H1 Fausse)**	**H0 Fausse (H1 Vraie)**
Accepter H0	$1-\alpha$ Décision correcte	β Erreur de 2ème espèce
Rejeter H0	α Erreur de 1ère espèce	1- β Décision correcte

Remarque :

1. Les procédures de tests fixent une limite supérieure au risque de première espèce. On prend souvent la valeur 5% (significatif) ou 1% (très significatif).

Cette valeur (aussi appelée seuil) représente le niveau de signification du test.

2. on souhaite minimiser à la fois les risques α et β mais, pour un échantillon donné, une diminution du risque α conduit à une augmentation du risque β.

3. l'erreur de seconde espèce diminue si la taille de l'échantillon augmente.

4. un test unilatéral est plus puissant qu'un test bilatéral.

4.1.3. Région de rejet et région d'acceptation

Les tests procèdent tous schématiquement de la même manière : on dispose d'une variable de décision X suivant une loi théorique donnée lorsque l'hypothèse H0 est vraie.

La région de rejet est l'ensemble des valeurs de la variable de décision qui sont "improbables" lorsque H0 est vraie. Ce sont des valeurs qui conduisent à rejeter cette hypothèse. Mais, si H0 était vraie, ce serait un rejet à tort. Or justement c'est la probabilité que mesure le risque α.

Par conséquent, le seuil α définit la taille de la région de rejet. C'est une région située sous la courbe de la distribution d'échantillonnage. Cette région peut prendre deux formes différentes :

Si on fait un test unilatéral, elle est entièrement à une extrémité de la distribution de probabilité ;

Si on fait un test bilatéral, elle est en deux morceaux de surface $\frac{\alpha}{2}$ à chaque extrémité de la distribution.

Voici deux illustrations graphiques dans le cas où la variable de décision X suit une loi normale centrée réduite N(0; 1).

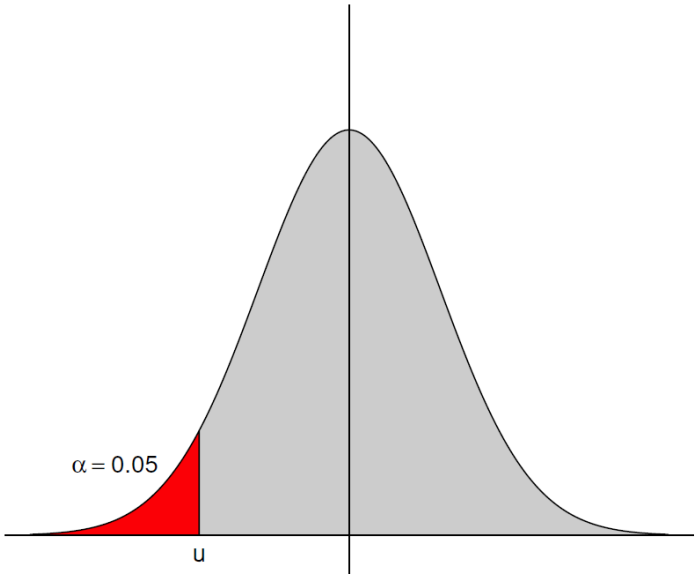


Figure 3.1: Région de rejet dans le cas du test unilatéral à droite

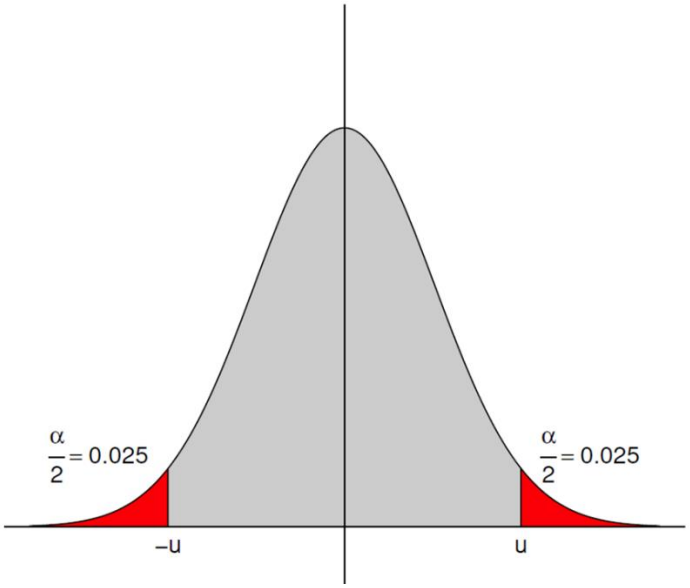


Figure 3.2: Région de rejet dans le cas du test bilatéral

Remarques :

Dans le cas d'un test unilatéral, la limite de la région de rejet est le quantile u tel que :

$$P(X < u) = 0{,}05$$

Par exemple, pour une distribution normale, on trouve u = -1,64 . Si le test était unilatéral (à droite), on aurait u = +1, 64.

Dans le cas d'un test bilatéral (et d'une distribution normale), les limites de la région de rejet sont les quantiles -u et u tels que :

$$P(-u < X < u) = 0{,}05$$

On trouve, pour une distribution normale, u = 1,96 .

La région d'acceptation est le complémentaire de la région de rejet. C'est l'ensemble des valeurs observées de la variable de décision pour lesquelles l'hypothèse H0 est acceptable.

La règle de décision des tests consiste à regarder si la valeur de la variable de décision se trouve dans la région d'acceptation ou dans la région de rejet.

4.1.4. Notion de « valeur p » d'un test

La valeur p (ou p-value) d'un test est la probabilité correspondant à la statistique Z, T, χ_p^2 ou F obtenue dans la construction du test. En prenant le cas d'une statistique qui suit une loi normale, la pvalue correspond à la probabilité à Z lue dans la table de la loi normale. Elle se définit comme la probabilité d'obtenir une statistique z qui soit supérieure à la statistique calculée Z :

$$Pvalue = P(z \geq Z)$$

$$Pvalue = 1 - P(z < Z)$$

Ainsi pour obtenir la pvalue, on lit d'abord dans la table la probabilité $P(z < Z)$. Ensuite, on calcule la pvalue.

Avec excel pour avoir $P(z < Z)$ pour la loi normale on utilise la fonction :

=LOI.NORMALE.STANDARD.N(Z;VRAI) où Z est la statistique calculée.

La pvalue fournit aussi une règle décision dans le test. En effet, lorsque la pvalue est inférieure au seuil α, on rejette H0. Mais lorsque la pvalue est supérieure au seuil α on ne peut pas rejeter H0.

4.1.5. Démarche générale d'un test

Les étapes de la mise en œuvre d'un test d'hypothèse sont les suivantes :

1. Choix du risque α
2. Choix des hypothèses H0 et H1
3. Détermination de la variable de décision i.e la variable sur laquelle porte le test (moyenne, variance, etc..)
4. Détermination de la région critique i.e la région de rejet de H0 : Choisir RC telle que $P(RC\,|H0\,) = \alpha$
5. Calcul éventuel de la puissance du test : $P(RC|H0\,) = 1 - \alpha$
6. Calcul, sur l'échantillon, de la valeur expérimentale de la variable de décision i.e la statistique du test.

7. Conclusion du test : rejet ou acceptation de H0 par comparaison de la valeur expérimentale à la valeur théorique de la statistique.

4.2. Les grandes familles de tests statistiques

En statistique, on distingue généralement trois grandes familles de tests : les tests de conformité, les tests de comparaison et des tests d'adéquation.

1. **les tests de conformité** sont des tests à 1 échantillon dans lesquels on compare la valeur d'un paramètre à une valeur théorique.

2. **les tests de comparaison** sont des tests qui comparent la valeur d'un paramètre entre 2 ou plusieurs échantillons.

3. **les tests d'adéquation** (appelés aussi tests d'ajustement) cherchent à vérifier si la distribution d'un échantillon est conforme à une loi de probabilité donnée. On les appelle tests d'homogénéïté lorsqu'on compare la distribution de deux échantillons entre eux.

4.3. Test de conformité à une valeur théorique

Les tests de conformité sont dits "tests à 1 échantillon". Ils ont pour but de vérifier si un échantillon peut être considéré comme représentatif de la population dont il est extrait.

On étudie une variable quantitative X et on cherche à établir si les observations sont en accord avec la loi théorique de cette variable.

En général, il s'agit de tester si un paramètre (tel que la moyenne, la fréquence ou la variance) calculé dans l'échantillon est conforme à sa valeur au niveau de la population. Ceci suppose que la loi théorique du paramètre est connue au niveau de la population.

Si θ est le paramètre, l'hypothèse nulle H0 est formulée de la manière suivante :

$$H_0 \ \theta = \theta_0$$

L'hypothèse alternative peut être l'une des trois formulations suivantes

$$\begin{cases} H_1 \ \theta \neq \theta_0 \\ H_1 \ \theta < \theta_0 \\ H_1 \ \theta > \theta_0 \end{cases}$$

Par exemple, dans un test sur la moyenne, on prendra la moyenne empirique comme estimateur et on posera : $H_0 \ \mu = m$.

Dans ce qui suit, on supposera que la taille d'échantillon est suffisamment importante pour que les populations soient considérées comme gaussiennes. Ce

qui signifie que le caractère observé X peut être considéré comme une variable aléatoire suivant une loi normale (aussi bien sur l'échantillon que dans la population).

Par ailleurs, on va distinguer deux principaux cas qui conduisent à des tests différents. Il s'agit notamment lorsque la variance dans la population est connue ou que la variance dans la population est inconnue.

4.3.1. Tests de conformité sur la moyenne

4.3.1.1. Exemple d'introduction

Le montant moyen des achats par client dans toutes les succursales d'une enseigne commerciale sont de 50 € avec un écart-type de 5 €. Afin de vérifier les performances d'un point de vente, le directeur prélève un échantillon des achats effectués par 20 consommateurs. Les résultats observés sont les suivants :

44.49	43.66	48.32	48.95	49.30	53.05	44.52	51.50	46.94	50.28
47.73	53.38	43.72	49.95	45.78	48.56	38.14	46.82	50.78	49.38

Les résultats fournis par cet échantillon sont-ils conformes aux valeurs globales ? En d'autres termes, la moyenne dans ce point de vente est-elle égale, supérieure ou inférieure à la moyenne générale ? La réponse à chacune de ces questions nécessite la mise en place d'un test de conformité qui peut se décliner en un test bilatéral, unilatéral (à gauche) ou unilatéral (à droite).

On suppose que les achats suivent une loi normale N(50,5).

La réalisation d'un test de conformité est toujours basée sur le calcul d'une statistique dite « statistique du test » en utilisant d'une part l'espérance et la valeur du paramètre étudié et en appliquant la loi des grandes (ou le théorème central limite).

Soit par exemple X une variable aléatoire suivant une loi normale de moyenne m et de variance σ^2. On montre d'abord que si $X \rightsquigarrow N(m, \sigma^2)$ alors : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N\left(m, \frac{\sigma^2}{n}\right)$. Ainsi avec le théorème central limite, on a

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1).$$

Cependant lorsque la variance σ^2 n'est pas connue, on utilise l'expression de la variance estimée telle que : $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Mais sachant que la variance estimée est la somme des carrés d'une loi normale, alors elle est distribuée selon une loi de χ^2 à n-1 degrés de libertés. Ainsi en appliquant le théorème central limite, on trouve :

$$T = \frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \rightsquigarrow t(n-1) \qquad (1.18)$$

Cette propriété montre que la connaissance de la variance a donc une implication importante dans la conduite des tests d'hypothèses.

Reprenons l'exemple des montants des achats par client présenté au début de la section. Le montant moyen des achats par client dans toutes les succursales d'une enseigne commerciale sont de 50 € avec un écart-type de 5 €. La moyenne calculée sur un échantillon de 20 consommateur est $\bar{X}$ = 47,7625. On souhaite mettre en œuvre trois tests (bilatéral, unilatéral à gauche et unilatéral à droite).

On verra que la démarche de mise en œuvre de chacun de ces tests dépend du fait que la variance soit connue ou pas.

4.3.1.2. Test de conformité bilatéral

Soit X une variable aléatoire suivant une loi normale de moyenne m et de variance σ^2. On souhaite tester l'hypothèse suivante :

$$\begin{cases} H_0 \ \theta = m \\ H_1 \ \theta \neq m \end{cases}$$

Ici θ est la valeur estimée à partir de l'échantillon et m est la valeur théorique

Test bilatéral lorsque la variance σ^2 est connue

Lorsque σ^2 est connue, avec le théorème central limite, on peut poser :

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0,1)$$

En fixant le seuil de première espèce α, la statistique de ce test se présente comme suit (compte tenu du caractère bilatéral) :

$$P\left(\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} < -Z^*_{1-\frac{\alpha}{2}}\right) + P\left(\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

$$P\left(Z < -Z^*_{1-\frac{\alpha}{2}}\right) + P\left(\mathrm{Z} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

Mais sachant que $P\left(Z < -Z^*_{1-\frac{\alpha}{2}}\right) = P\left(Z > Z^*_{1-\frac{\alpha}{2}}\right)$ (propriété d'une loi symétrique), on a :

$$2\,P\left(\mathrm{Z} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

Où Z est la statistique du test calculée ($\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}$) et $Z^*_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite (lue dans la table de loi normale centrée réduite).

Par ailleurs sachant que $P\left(Z < -Z^*_{1-\frac{\alpha}{2}}\right) + P\left(\mathrm{Z} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$, cela signifie que :

$$P\left(-Z^*_{1-\frac{\alpha}{2}} < Z < Z^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(|Z| < Z^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Dès lors on peut utiliser l'une des deux expressions pour prendre la décision du test : soit $2\,P\left(\mathrm{Z} > Z^*_{1-\frac{\alpha}{2}}\right) = \alpha$ qui exprime le seuil d'erreur ou $P\left(|Z| < Z^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$ qui exprime le seuil de confiance. Dans l'un ou l'autre des cas, on compare la valeur Z calculée à la valeur de $Z^*_{1-\frac{\alpha}{2}}$ lue dans la table de la loi normale. Ainsi lorsque $Z > Z^*_{1-\frac{\alpha}{2}}$, on rejette l'hypothèse H0. En revanche lorsque $Z < Z^*_{1-\alpha}$, on ne peut pas rejeter H0.

La région critique de ce test (encore appelée région de rejet de H0) se définit telle que :

$$RC = \left\{\left|\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}\right| > Z^*_{1-\frac{\alpha}{2}}\right\}\ soit\ \ \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} \notin \left[-Z^*_{1-\frac{\alpha}{2}}\,;\, Z^*_{1-\frac{\alpha}{2}}\right]$$

Ou

$$RC = \left\{|\bar{X}-m| > Z^*_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\}\ soit\ \bar{X} \notin \left]m - Z^*_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,;\, \mathrm{m} + Z^*_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\left|\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}\right| < Z^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{-Z^*_{1-\alpha} < \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} < Z^*_{1-\frac{\alpha}{2}}\right\}$$

Ou

$$RA = \left\{ \mathrm{m} - Z^*_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \bar{X} < \mathrm{m} + Z^*_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right\}$$

Ainsi connaissant la région de rejet ou la région d'acceptation, on peut donner une autre règle de décision par rapport au test. En effet après avoir calculée la moyenne $\bar{X}$ sur l'échantillon, on regarde si sa valeur appartient ou pas à la région d'acception. Ainsi si $\bar{X}$ appartient à l'intervalle RA, on ne peut pas rejeter H0. Par contre si $\bar{X}$ appartient de RC, on rejette H0.

Application : La variance est connu ($\sigma = 5$), $\alpha = 0{,}05$ et m = 50, $\bar{X} = 47{,}7625$ et n=20.

Ainsi

$$Z = \frac{47{,}7625 - 50}{\frac{5}{\sqrt{20}}} = -2{,}0013$$

On sait que dans la table de la loi normale $Z^*_{1-\frac{\alpha}{2}} = 1{,}96$.

Pour le déterminer avec Excel on peut utiliser la fonction :

=LOI.NORMALE.STANDARD.INVERSE.N(0,975)

Ainsi, comme $|Z| > Z^*_{1-\frac{\alpha}{2}}$, on rejette H0 au seuil de 5%.

Test bilatéral lorsque la variance σ^2 n'est pas connue

Lorsque la variance σ^2 n'est pas connue, on utilise la variance estimée telle que : $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Dès lors, en appliquant le théorème central limite on trouve une loi de Student qui se présente comme suit :

$$\frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \rightsquigarrow t(n-1)$$

Dans cette configuration, en fixant le seuil de première espèce α, la statistique du test se présente comme suit (compte tenu au du caractère bilatéral) :

$$P\left(\frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < -T^*_{1-\frac{\alpha}{2}} \right) + P\left(\frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > T^*_{1-\frac{\alpha}{2}} \right) = \alpha$$

$$P\left(T < -T^*_{1-\frac{\alpha}{2}} \right) + P\left(T > T^*_{1-\frac{\alpha}{2}} \right) = \alpha$$

Mais sachant que $P\left(T < -T^*_{1-\frac{\alpha}{2}}\right) = P\left(T > T^*_{1-\frac{\alpha}{2}}\right)$ (propriété d'une loi symétrique), on a :

$$2\,P\left(T > T^*_{1-\frac{\alpha}{2}}\right) = \alpha$$

Où T est la statistique du test calculée et $T^*_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite (lue dans la table de loi normale centrée réduite).

Par ailleurs sachant que $P\left(T < -T^*_{1-\frac{\alpha}{2}}\right) + P\left(T > T^*_{1-\frac{\alpha}{2}}\right) = \alpha$, cela signifie que :

$$P\left(-T^*_{1-\frac{\alpha}{2}} < T < T^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(|T| < T^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Dès lors on peut utiliser l'une des deux expressions pour prendre la décision du test : soit $2\,P\left(T > T^*_{1-\frac{\alpha}{2}}\right) = \alpha$ qui exprime le seuil d'erreur ou $P\left(|T| < T^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$ qui exprime le seuil de confiance. Dans l'un ou l'autre des cas, on compare la valeur T calculée à la valeur de $T^*_{1-\frac{\alpha}{2}}$ lue dans la table de la loi normale. Ainsi lorsque $T > T^*_{1-\frac{\alpha}{2}}$, on rejette l'hypothèse H0. En revanche lorsque $T < T^*_{1-\alpha}$, on ne peut pas rejeter H0.

La région critique de ce test (encore appelée région de rejet de H0) se définit telle que :

$$RC = \left\{\left|\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}}\right| > T^*_{1-\frac{\alpha}{2}}\right\} \; soit \quad \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \notin \left[-T^*_{1-\frac{\alpha}{2}}\,;\, T^*_{1-\frac{\alpha}{2}}\right]$$

Ou

$$RC = \left\{|\bar{X}-m| > T^*_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n-1}}\right\} \; soit \; \bar{X} \notin \left]m - T^*_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n-1}}\,;\, m + T^*_{1-\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n-1}}\right[$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\left|\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}}\right| < T^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{ -T^*_{1-\alpha} < \frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < T^*_{1-\frac{\alpha}{2}} \right\}$$

Ou

$$RA = \left\{ \mathrm{m} - T^*_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n-1}} < \bar{X} < \mathrm{m} + T^*_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n-1}} \right\}$$

4.3.1.3. Test de conformité unilatéral (à droite)

$$\begin{cases} H_0 \ \mu = m \\ H_1 \ \mu > m \end{cases}$$

Test unilatéral (à droite) lorsque la variance σ^2 est connue

Lorsque la variance est connue la statistique du test sous H_0 suit une loi normale$N(0,1)$. Ainsi connaissant le seuil d'erreur α on définit la région critique telle que :

$$P\left(\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} > Z^*_{1-\alpha} \right) = \alpha$$

$$P(Z > Z^*_{1-\alpha}) = \alpha$$

Où Z est la statistique du test calculée et $Z^*_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite (lue dans la table de loi normale centrée réduite).

Ainsi lorsque $Z > Z^*_{1-\alpha}$, on rejette l'hypothèse H0. En revanche lorsque $Z < Z^*_{1-\alpha}$, on ne peut pas rejeter H0.

La région critique de ce test (encore appelée région de rejet de H0) se définit alors comme suit :

$$RC = \left\{ \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} > Z^*_{1-\alpha} \right\}$$

Ou

$$RC = \left\{ \bar{X} > \mathrm{m} + Z^*_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\}$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} < Z_{1-\alpha}^{*}\right) = 1-\alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{ \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} < Z_{1-\alpha}^{*} \right\} \; soit \quad \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} \in]-\infty\,;\, Z_{1-\alpha}^{*}]$$

Ou

$$RA = \left\{ \bar{X} < \mathrm{m} + Z_{1-\alpha}^{*} \frac{\sigma}{\sqrt{n}} \right\} \; soit \; \bar{X} \in \left]-\infty\,;\, \mathrm{m} + Z_{1-\alpha}^{*} \frac{\sigma}{\sqrt{n}}\right]$$

Détermination de la P-value du test :

La pvalue de ce test se détermine en calculant la probabilité $1 - P(z < Z)$ où $P(z < Z)$ est la probabilité correspondant à Z dans la table de la loi normale.

Test unilatéral (à droite) lorsque la variance σ^2 n'est pas connue :

Lorsque σ^2 n'est pas connue, on a :

$$\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \rightsquigarrow T(n-1)$$

Ainsi connaissant le seuil d'erreur α on définit la région critique du test telle que :

$$P\left(\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > T_{1-\alpha}^{*}\right) = \alpha$$

$$P(T > T_{1-\alpha}^{*}) = \alpha$$

Où T est la statistique du test calculée et $T_{1-\alpha}^{*}$ le quantile d'ordre $1-\alpha$ de la loi de la loi de Student.

Ainsi lorsque $T > T_{1-\alpha}^{*}$, on rejette l'hypothèse H0. Et lorsque $T < T_{1-\alpha}^{*}$, on ne peut pas rejeter H0.

La région critique du test se définit comme suit :

$$RC = \left\{ \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > T_{1-\alpha}^{*} \right\}$$

Ou

$$RC = \left\{\bar{X} > \mathrm{m} + T^*_{1-\alpha}\frac{\hat{\sigma}}{\sqrt{n-1}}\right\}$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < T^*_{1-\alpha}\right) = 1-\alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < T^*_{1-\alpha}\right\} \; soit \quad \frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \in]-\infty\,;\, T^*_{1-\alpha}]$$

Ou

$$RA = \left\{\bar{X} < \mathrm{m} + T^*_{1-\alpha}\frac{\hat{\sigma}}{\sqrt{n-1}}\right\} \; soit \; \bar{X} \in \left]-\infty\,;\, \mathrm{m} + T^*_{1-\alpha}\frac{\hat{\sigma}}{\sqrt{n-1}}\right]$$

Détermination de la P-value du test :

La pvalue de ce test se détermine en calculant la probabilité $1 - P(t < T)$ où $P(t < T)$ est la probabilité correspondant à t dans la table de la loi normale.

On peut déterminer $P(t < T)$ avec excel , utilisant la fonction

4.3.1.4. Test de conformité unilatéral (à gauche)

$$\begin{cases} H_0 \;\; \mu = m \\ H_1 \;\; \mu < m \end{cases}$$

Test unilatéral (à gauche) lorsque la variance σ^2 est connue

Lorsque la variance est connue la statistique du test sous H_0 suit une loi normale $N(0,1)$. Ainsi connaissant le seuil d'erreur α on définit la région critique telle que :

$$P\left(\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} < -Z^*_{1-\alpha}\right) = \alpha$$

$$P(Z < -Z^*_{1-\alpha}) = \alpha$$

Où Z est la statistique du test calculée et $Z^*_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite. Ainsi lorsque $Z < Z^*_{1-\alpha}$, on rejette l'hypothèse H0. Et lorsque $Z > Z^*_{1-\alpha}$, on ne peut pas rejeter H0.

La région critique du test se définit alors comme suit :

$$RC = \left\{ \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} < -Z_{1-\alpha}^{*} \right\}$$

Ou

$$RC = \left\{ \bar{X} < \mathrm{m} - Z_{1-\alpha}^{*} \frac{\sigma}{\sqrt{n}} \right\}$$

Sachant que $\left(\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}} > -Z_{1-\alpha}^{*} \right) = 1 - \alpha$, on peut définir la région d'acceptation de H0 comme suit.

$$RA = \left\{ \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} > -Z_{1-\alpha}^{*} \right\} \; soit \quad \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \in]-Z_{1-\alpha}^{*} ; +\infty]$$

Ou

$$RA = \left\{ \bar{X} > \mathrm{m} - Z_{1-\alpha}^{*} \frac{\sigma}{\sqrt{n}} \right\} \; soit \; \bar{X} \in \left] \mathrm{m} - Z_{1-\alpha}^{*} \frac{\sigma}{\sqrt{n}} ; +\infty \right]$$

Détermination de la P-value du test :

La pvalue de ce test se détermine en calculant la probabilité $1 - P(z < Z)$ où $P(z < Z)$ est la probabilité correspondant à Z dans la table de la loi normale.

Test unilatéral (à gauche) lorsque la variance σ^2 n'est pas connue

Lorsque σ^2 n'est pas connue, sachant que $\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \rightsquigarrow T(n-1)$ et connaissant le seuil d'erreur α on définit la région critique du test telle que :

$$P\left(\frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < -T_{1-\alpha}^{*} \right) = \alpha$$

$$P(T < -T_{1-\alpha}^{*}) = \alpha$$

Où T est la statistique du test calculée et $T_{1-\alpha}^{*}$ le quantile d'ordre $1 - \alpha$ de la loi de la loi de Student.

Ainsi lorsque $T < -T_{1-\alpha}^{*}$, on rejette l'hypothèse H0. Et lorsque $T > -T_{1-\alpha}^{*}$, on ne peut pas rejeter H0.

La région critique du test se définit comme suit :

$$RC = \left\{ \frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} < -T^*_{1-\alpha} \right\}$$

Ou

$$RC = \left\{ \bar{X} < \mathrm{m} - T^*_{1-\alpha} \frac{\hat{\sigma}}{\sqrt{n-1}} \right\}$$

Connaissant donc la région critique, on peut définir la région d'acceptation de H0 sachant que $\left(\frac{\bar{X}-m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > -T^*_{1-\alpha} \right) = 1 - \alpha$. Dès lors, la région d'acceptation se définit comme suit.

$$RA = \left\{ \frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} > -T^*_{1-\alpha} \right\} \; soit \quad \frac{\bar{X} - m}{\frac{\hat{\sigma}}{\sqrt{n-1}}} \in \left]-T^*_{1-\alpha};\ +\infty\ \right[$$

Ou

$$RA = \left\{ \bar{X} > \mathrm{m} - T^*_{1-\alpha} \frac{\hat{\sigma}}{\sqrt{n-1}} \right\} \; soit \;\; \bar{X} \in \left] \mathrm{m} - T^*_{1-\alpha} \frac{\hat{\sigma}}{\sqrt{n-1}}; +\infty \right]$$

Détermination de la P-value du test :

La pvalue de ce test se détermine en calculant la probabilité $1 - P(t < T)$ où $P(t < T)$ est la probabilité correspondant à t dans la table de la loi normale.

4.3.2. Tests de conformité sur la proportion

La pvalue de ce test se détermine en calculant la probabilité $1 - P(t < T)$ où $P(t < T)$ est la probabilité correspondant à t dans la table de la loi normale.

Le test de conformité de la proportion se met en œuvre de la même façon que le test de conformité de la moyenne en supposant que l'échantillon est distribué selon une loi normale.

Par exemple, en utilisant un échantillon, on souhaite savoir si la proportion est égale à une valeur donnée. Ce test se présente sous la forme suivante

$$H_0 \;\; p = p_0$$

La proportion empirique est :

$$F_n = \frac{1}{n} \sum_{i=1}^{n_1} 1_{X_i = 1} = \frac{n_1}{n}$$

Avec $E(F_n) = p$ et $Var(F_n) = \frac{p(1-p)}{n}$ où n_1 est le nombre de succès, n la taille de l'échantillon. En utilisant la loi des grands nombres, on peut écrire :

$$\left(\frac{F_n - p_0}{\sqrt{\frac{p(1-p_0)}{n}}}\right) = \sqrt{n}\left(\frac{F_n - p_0}{\sqrt{p_0(1-p_0)}}\right) \rightsquigarrow N(0,1)$$

Ainsi, à un risque d'erreur , on peut alors calculer la statistique de test comme suit :

$$Z = \sqrt{n}\left(\frac{F_n - p_0}{\sqrt{p_0(1-p_0)}}\right)$$

- Dans le cas d'un test bilatéral, on rejette l'hypothèse nulle lorsque $|Z| > Z_{1-\frac{\alpha}{2}}$, c'est-à-dire lorsque $Z \notin]-Z_{1-\frac{\alpha}{2}}, Z_{1-\frac{\alpha}{2}}[$.
- Dans le cas d'un test unilatéral à gauche, on rejette l'hypothèse nulle lorsque $Z < -Z_{1-\frac{\alpha}{2}}$.
- Et dans le cas d'un test unilatéral à droite, on rejette l'hypothèse nulle lorsque $Z > Z_{1-\frac{\alpha}{2}}$.

Conditions de validité du test sur la proportion

Le test de proportion n'est valable que sous les conditions suivantes :

$$\begin{cases} n \geq 30 \\ np_0 \geq 5 \\ n(1-p_0) \geq 5 \end{cases}$$

Lorsque ces conditions ne sont pas vérifiées, il faut alors utiliser le Test exact ou des tests non paramétriques. Ces types de tests ne sont pas présentés ici.

4.3.3. Tests de conformité sur la variance

Le test de conformité sur la variance permet de tester la valeur de la variance $Var(X)$ d'un caractère X dans la population au vu de la variance empirique d'un échantillon. On suppose que la variable est distribuée selon une loi normale.

L'hypothèse H0 est que la variance au niveau de la population a une certaine valeur σ^2. On pose ainsi :

$$H0 : \mathrm{Var(X)} = \sigma^2$$

En notant S^2 la variance empirique (modifiée) de l'échantillon, on montre que sous l'hypothèse H0, la statistique du test est la suivante :

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

Cette statistique suit une loi de khi-deux à n-1 degrés de liberté.

Ainsi, pour déterminer la région d'acceptation (respectivement la région de rejet) de H0, il faut utiliser les quantiles de la loi du χ^2. Par exemple, dans le cas d'un test bilatéral au seuil 5% pour avoir la région de rejet, il faut trouver les bornes a et b telles que :

$$P(Y \leq a) = \frac{\alpha}{2}$$

$$P(Y \geq b) = \frac{\alpha}{2}$$

Exemple

Une société fabrique un câble en acier trempé galvanisé dont la charge de rupture est de 210 kg avec une marge de 5 kg (écart-type). Un contrôle de qualité effectué sur 10 bobines a conduit aux résultats suivants :

203.70	211.80	201.60	226.00	213.30
201.80	214.90	217.40	215.80	206.90

Cet échantillon confirme-t-il la marge annoncée ?

On calcule la moyenne et la variance modifiée de l'échantillon :

$\bar{X} = 211{:}32$ et Var(X) = S^2 =61,357

La statistique du test de variance vaut :

$$Y = \frac{n-1}{\sigma^2} S^2 = \frac{10-1}{5^2} 61{,}357^2 = 22{,}088$$

Avec un échantillon de taille n = 10, on a n -1 = 9 degrés de liberté et les tables de la loi du χ^2 donnent les valeurs suivantes pour les quantiles : a = 2,70 et b = 19,02 Car $P(Y \leq 2{,}70) = \frac{0{,}05}{2}$ et $P(Y \geq 19{,}02) = \frac{0{,}05}{2}$ (voir Annexe pour la lecture de la table de khi-deux).

La figure ci-dessous illustre la région de rejet et d'acceptation dans le cadre d'une loi de khi-deux.

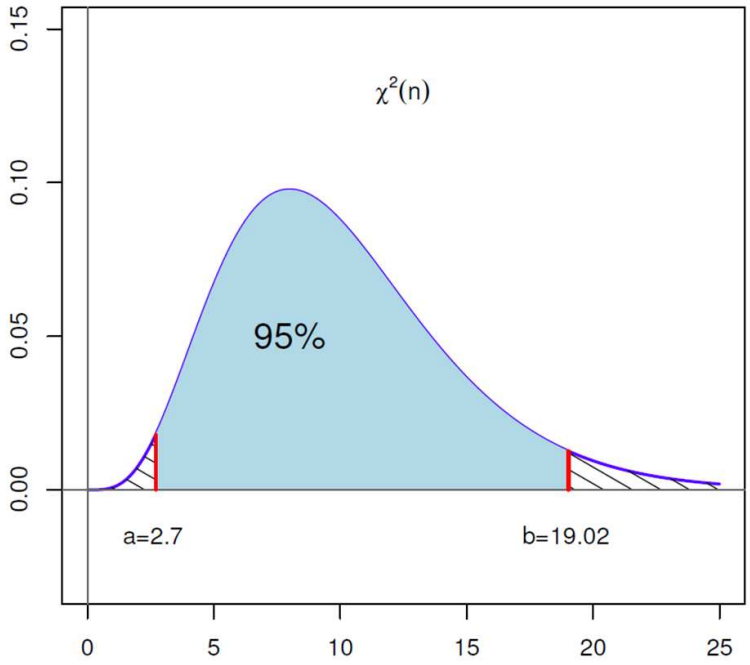


Figure 3.3: Région d'acceptation et de rejet dans le cas d'une loi de χ^2.

Cette valeur de Y se trouve dans la région de rejet, c'est-à-dire à l'extérieur de l'intervalle [2,70 ; 19,02]. On doit donc rejeter l'hypothèse et considérer, au risque 5% de se tromper, que l'échantillon étudié présente une variance qui ne correspond pas à la variance annoncée.

Remarques

1. Noter qu'ici l'intervalle n'est pas symétrique autour de l'espérance.

2. La variance utilisée dans la statistique de ce test est la variance empirique modifiée (c'est-à-dire l'estimateur sans biais de σ^2).

3. Le mode de la loi χ^2 vaut n -2 (pour n > 1). C'est l'abscisse du maximum sur le graphe ci-dessous.

Approximations de la loi du χ^2 dans le calcul de l'intervalle de confiance

Lorsque la taille de l'échantillon est grande, on peut remplacer la loi du χ^2 par des lois approchantes.

On sait qu'une loi du χ^2 à n degrés de liberté a pour espérance n et pour variance $2n$. Alors pour une variable X qui suit une loi de χ^2 (n), le théorème central limite permet d'écrire que :

$$Z = \left(\frac{X - n}{\sqrt{2n}}\right) \rightsquigarrow N(0,1)$$

On construit donc l'intervalle d'acceptation pour la variable Z avec la loi normale plutôt que pour la variable X avec la loi du χ^2.

Une autre approximation possible est fournie par le théorème suivant :

Théorème (de Fisher). Si X est une variable aléatoire suivant une loi du χ^2 à n degrés de liberté alors la quantité $\sqrt{2X} - \sqrt{2n-1}$ converge en loi vers une loi normale lorsque n tend vers l'infini.

$$\sqrt{2X} - \sqrt{2n-1} \rightsquigarrow N(0,1)$$

À partir de la statistique Y calculée dans le test de variance, on calcule la quantité $\sqrt{2Y} - \sqrt{2n-1}$ et on voit si elle est dans la région d'acceptation ou pas.

L'intérêt de l'approximation de Fisher par rapport au théorème central limite est qu'elle procure une convergence plus rapide.

4.4. Tests de comparaison d'échantillons

4.4.1. Présentation

On considère deux paramètres $\theta_1$et $\theta_2$de deux distributions X_1 et X_2 provenant de deux populations 1 et 2.

L'objectif est comparer θ_1 et θ_2 en utilisant un test statistique.

Exemple 1

Un physiologiste pense que la force musculaire est accrue par l'absorption d'un produit A. La force musculaire est mesurée par la force de préhension (pression exercée sur un certain objet; plus la valeur est élevée, plus la force musculaire est importante).

- Population 1 : personnes n'ayant pas absorbée le produit A ;Variable 1 : force de préhension X_1 ; Paramètre 1 : m_1 = moyenne de X_1.
- Population 2 : personnes ayant absorbées le produit A ;Variable 2 : force de préhension X_2 ;Paramètre 2 : m_2 = moyenne de X_2
- Question : $m_1 < m_2$?

Exemple 2

Dans une pisciculture, on étudie l'effet de deux régimes alimentaires, que l'on appellera régime 1 et régime 2, sur la croissance d'une espèce de poisson.

Pour ce faire, on alimente un lot de poissons avec le régime 1 et un autre lot avec le régime 2.

- Population 1 : poissons alimentés avec le régime 1 ; Variable 1 : longueur X_1 ; Paramètre 1 : m_1 = moyenne de X_1

- Population 2 : poissons alimentés avec le régime 2 ;Variable 2 : longueur X_2 ; Paramètre 2 : m_2 = moyenne de X_2
- Question : $m_1 = m_2$?

L'hypothèse H0 est l'hypothèse d'égalité :

H0 : $\theta_1 = \theta_2$

L'hypothèse H_1 est une des trois hypothèses suivantes :

H1 : $\theta_1 \neq \theta_2$

H1 : $\theta_1 < \theta_2$

H1 : $\theta_1 > \theta_2$

4.4.2. Test de comparaison de deux moyennes

Principe du test

On dispose de deux échantillons sur lesquels on observe les moyennes empiriques $\bar{X}_1$ et $\bar{X}_2$. On souhaite alors comparer m_1 et m_2 qui représentent respectivement les moyennes théoriques sur les deux populations.

Dans ce test, comparer m_1 et m_2 revient à comparer la différence par rapport à 0.

L'hypothèse nulle H0 : $\theta_1 = \theta_2$ peut être reformulée telle que H0 : $D = \theta_1 - \theta_2 = 0$

Hypothèses de base sur les variables

A l'instar du test de conformité d'une moyenne à une valeur théorique, il existe plusieurs procédures selon qu'on suppose connue ou non la distribution des variables X_1 et X_2, et que l'on dispose d'un grand ou petit échantillon.

Dans cet exposé nous examinons tous les cas possibles.

Nous postulons dans un premier temps deux hypothèses (qui seront par la suite levées) :

Hypothèse de normalité : Les variables X_1 et X_2 sont distribuées selon une loi normale.

Hypothèse d'égalité des variances: Les variances de X_1 et de X_2 sont égales.

Ces deux hypothèses se traduisent comme suit : $X_1 \rightsquigarrow N(m, {\sigma_1}^2)$ et $X_n \rightsquigarrow N(m, {\sigma_2}^2)$ avec ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$.

4.2.2.1. Test de comparaison sur échantillon appariés

Exemple 1

Un physiologiste veut savoir si la force musculaire est accrue par l'absorption d'un produit A. Il réalise l'expérience sur un échantillon de10 personnes choisies au hasard et mesure leurs forces de préhension **avant** et **après** absorption du produit A. De tels échantillons sont appelés « **échantillons appariés** ».

Car, on mesure la variable deux fois une même personne. On considère donc qu'on observe pour une même personne le couple (X_1,X_2).

Le tableau suivant donne les valeurs obtenues lors de l'expérience.

individu	1	2	3	4	5	6	7	8	9	10
X_1	19.60	11.20	9.00	25.10	28.40	17.90	6.50	32.00	11.60	24.00
X_2	20.20	13.40	8.50	27.40	31.50	17.60	7.30	37.50	12.00	22.90

Où X_1 représente la valeur des forces de préhension des individus avant la prise du produit et X_2 la force de préhension après la prise du produit.

On souhaite alors tester si la force de préhension s'améliore après la prise du produit. Il s'agit alors de tester si la différence $D = m_1 - m_2 = 0$ contre une hypothèse alternative à définir.

L'estimateur ponctuelle de $D = m_1 - m_2$ est $\overline{D} = \bar{X}_1 - \bar{X}_2 = 0$.

Sur l'échantillon $\bar{X}_1 = 18{,}5$ et $\bar{X}_2 = 19{,}8$.

Ainsi l'estimation ponctuelle de D est $\overline{D} = -1{,}3$

Statistique de test

Notons n la taille des échantillons appariés. Notons D_i la variable mesurant la différence entre X_1 et X_2 pour le i-ème individu de l'échantillon.

Introduisons la variance empirique modifiée S^2 associée à $D = X_1 - X_2$:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} {D_i}^2 - n\overline{D}^2\right)$$

Sachant qu'en moyenne la différence D est égal à 0 et connaissant la variance empirique de D, on peut formuler la statistique du test en appliquant le théorème central-limite telle que :

$$\sqrt{n}\left(\frac{\overline{D} - 0}{S}\right) = \sqrt{n}\left(\frac{\overline{D}}{S}\right) \rightsquigarrow T(n-1)$$

La différence centrée et réduite suit donc une loi de khi-deux à n-1 degrés de liberté.

La valeur de la statistique permet de conclure selon qu'il s'agisse d'un test bilatéral ou unilatéral.

- Dans le cas d'un test bilatéral, on rejette l'hypothèse nulle lorsque $|T| > T^*_{1-\frac{\alpha}{2}}$, c'est-à-dire lorsque $T \notin]-T^*_{1-\frac{\alpha}{2}}, T^*_{1-\frac{\alpha}{2}}[$.
- Dans le cas d'un test unilatéral à gauche, on rejette l'hypothèse nulle lorsque $T < -T^*_{1-\alpha}$.
- Et dans le cas d'un test unilatéral à droite, on rejette l'hypothèse nulle lorsque $T > T^*_{1-\alpha}$.

La règle de décision est alors la même que pour un test de conformité.

Application :

individu	1	2	3	4	5	6	7	8	9	10
X_1	19.60	11.20	9.00	25.10	28.40	17.90	6.50	32.00	11.60	24.00
X_2	20.20	13.40	8.50	27.40	31.50	17.60	7.30	37.50	12.00	22.90
D_i	-0,6	-2,2	0,5	-2,3	-3,1	0,3	-0,8	-5,5	-0,4	1,1

$S^2 = 3{,}978$, $\overline{D} = -1{,}3$

$$T = \sqrt{10}\left(\frac{-1{,}3}{1{,}994}\right) = -2{,}06$$

Sous l'hypothèse nulle H0 $m_1<m_2$, en, fixant le seuil d'erreur α à 5%, $T^*_{1-\alpha}$ à 9 degrés de liberté est à gauche est -1,833.

La région d'acceptation se présente alors comme suit :

RA=] -1,833 , +∞ [

Et comme la statistique T n'est pas située à l'intérieur de cette région, on rejette alors l'hypothèse nulle concluant ainsi que la moyenne de la force de préhension est significativement plus élevée après la prise du produit qu'avant.

4.2.2.2. Test de comparaison sur échantillon indépendants

Exemple 2

On fait suivre le régime alimentaire 1 à 9 poissons choisis au hasard et le régime 2 à 8 poissons choisis au hasard. Et on mesure la longueur des poissons à la suite de l'expérience. Les résultats obtenus sont fournis dans le tableau ci-dessous.

		1	2	3	4	5	6	7	8	9
Régime 1	X1	21.18	20.01	22.50	22.97	21.83	23.42	18.61	25.20	22.07
Régime 2	X2	22.39	21.26	22.17	25.00	22.21	20.51	22.36	24.49	

On souhaite savoir si en moyenne, la longueur des poissons est significativement différente d'un régime à l'autre. De tels échantillons sont appelés « **échantillons indépendants** ». Car contrairement appariés, les mesures ne sont pas prises sur les unités entre les deux échantillons. Ici on fait suivre chaque régime alimentaire à des lots différents de poissons.

Il s'agit alors de tester si la différence $D = m_1 - m_2 = 0$ contre une hypothèse alternative à définir (ici le cas bilatéral).

L'estimateur ponctuelle de $D = m_1 - m_2$ est $\overline{D} = \bar{X}_1 - \bar{X}_2 = 0$.

Sur l'échantillon $\bar{X}_1 = 18{,}5$ et $\bar{X}_2 = 19{,}8$.

Ainsi l'estimation ponctuelle de D est $\overline{D} = -1{,}3$

Statistique de test

Notons n_1 et n_2 les tailles respectives des échantillons de X_1 et X_2. Notons S_1^2 et S_2^2 les variances empiriques modifiées associées, respectivement X_1 et X_2.

Définissons l'estimateur suivant :

$$S_{1-2}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Où S_{1-2}^2 est la moyenne pondérée des variances empiriques modifiées des deux échantillons.

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n}(X_{i\,1} - \overline{X_1})^2$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n}(X_{i\,2} - \overline{X_2})^2$$

Lorsque les deux variables X_1 et X_2 ont la même variance σ^2, on montre que S_{1-2}^2 est un estimateur de ce σ^2 commun. On verra, par la suite, un test de comparaison des variances qui permet justement de vérifier l'hypothèse que ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$ (et qui devra toujours être effectuée au préalable). Pour l'instant, on suppose cette égalité.

Connaissant la valeur de cette variance empirique combinée, la statistique du test se présente sous H0 d'égalité comme suit :

$$\left(\frac{(\overline{X_1}-\overline{X_2})-0}{\left(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)S_{1-2}}\right)=\left(\frac{(\overline{X_1}-\overline{X_2})}{\left(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)S_{1-2}}\right)\rightsquigarrow T(n_1+n_2-2)$$

Application : $n_1 = 9$ et $n_2 = 8$; $= 0{,}05$, $\bar{d} = \overline{X_1} - \overline{X_2} = -0{,}5$, $S_1^2 = 3{,}7$, $S_2^2 = 2{,}273$, $S_{1-2}^2 = 3{,}034$.

Ainsi

T

$$T = \left(\frac{(\overline{X_1}-\overline{X_2})}{\left(\sqrt{\frac{1}{9}+\frac{1}{8}}\right)S_{1-2}}\right)=\left(\frac{(-0{,}5)}{\left(\sqrt{\frac{1}{9}+\frac{1}{8}}\right)\sqrt{3{,}034}}\right)=-0{,}591$$

Quant à la règle de décision, elle est la même que dans le cas du test de comparaison sur échantillons appariés ou dans le cas des tests de conformité.

La règle générale est la suivante :

- Dans l'hypothèse alternative où $m_1 < m_2$, l'intervalle d'acceptation de l'hypothèse nulle est :
$$IA_{1-\alpha}(T) = [-T_{1-\alpha}^*, +\infty\,[$$
- Dans l'hypothèse alternative où $m_1 > m_2$, l'intervalle d'acceptation de l'hypothèse nulle est :
$$IA_{1-\alpha}(T) = [-\infty, T_{1-\alpha}^*\,[$$
- Dans l'hypothèse alternative où $m_1 \neq m_2$, l'intervalle d'acceptation de l'hypothèse nulle est :
$$IA_{1-\frac{\alpha}{2}}(T) = [-T_{1-\frac{\alpha}{2}}^*, T_{1-\frac{\alpha}{2}}^*\,[$$

Où $T_{1-\alpha}^*$ ou $T_{1-\frac{\alpha}{2}}^*$ sont des seuils théoriques lus dans la table de Student au seuil $1-\alpha$ (test unilatéral) ou au seuil $1-\frac{\alpha}{2}$ (test bilatéral).

Dans le cas du test bilatéral, $T_{1-\frac{\alpha}{2}}^* = T_{0{,}975}^*$ à (9+8-2) degrés de liberté est 2,131.

Dès lors, l'intervalle d'acceptation de H0 est $[-2{,}131\,,\,2{,}131\,]$.

La valeur T étant comprise dans cet intervalle, on ne peut donc rejeter que la différence entre les deux régimes alimentaires est nulle. Donc deux régimes alimentaires équivalents.

NB : Dans le cas du test de comparaison d'échantillon (appariés ou indépendants), lorsque la taille de l'échantillon est suffisamment grand (tend vers l'infini), on peut approximer la loi de Student par la loi normale. Dès lors, les statistiques de test précédemment formulées selon la loi de Student peuvent être formulée directement selon une loi normale et les tests interprétés selon la statistique Z. Cela résume alors comme suit :

- **Echantillons appariés**

$$Z = \sqrt{n}\left(\frac{\overline{D}}{S}\right) \rightsquigarrow N(0,1)$$

Où S représente la variance empirique de la différence $\overline{D}$ définie telle que $\overline{D} = \bar{X}_1 - \bar{X}_2$

- **Echantillons indépendants**

$$Z = \left(\frac{(\overline{X_1} - \overline{X_2})}{\left(\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}\right)}\right) \rightsquigarrow N(0,1)$$

Où ${\sigma_1}^2$ et ${\sigma_2}^2$ représentent respectivement les variances théoriques sur les deux échantillons. Si ces deux variances ne sont pas connues, on les remplace par leur équivalent empirique S_1^2 et S_2^2.

4.4.3. Test de comparaison de deux proportions

Exemple d'introduction

Un enseignant se demande si la proportion de filles en deuxième année de Licence de Droit est différente de la proportion de filles en deuxième année de Licence d'économie et gestion. Soit p_1 la proportion de filles dans la deuxième année de Droit et p_2 la proportion de filles dans la deuxième année d'Eco. L'objectif de l'enseignant est de tester si $p_1 = p_2$.

L'hypothèse nulle du test est donc que $H_0 : p_1 = p_2$ contre l'alternative $H_1 : p_1 \neq p_2$.

On sélectionne au hasard deux échantillons : un échantillon de taille n_1=85 étudiants dans la première population et un échantillon de taille n_2=78 dans la deuxième population. On obtient les résultats suivants :

Filière	Echantillon	Filles	Garçons
L2 Droit	85	52	33
L2 Economie et Gestion	78	36	42

Soit F_1 et F_2 les fréquences empiriques de filles respectivement sur l'échantillon 1 et 2. On a :

$$F_1 = \frac{52}{85} = 0{,}612$$

$$F_2 = \frac{36}{78} = 0{,}462$$

On sait que F_1 et F_2 sont des estimations ponctuelles de p_1 et p_2.

La démarche de mise en œuvre du test de comparaison de proportion (sur échantillons indépendants) est la même que celle du test de comparaison de moyenne (sur échantillons indépendants). Elle est basée sur la différence $D = F_1 - F_2 = 0$.

Sous l'hypothèse de normalité (loi des grands nombres), la statistique du test est la suivante :

$$Z = \frac{F_1 - F_2}{\sqrt{\hat{F}(1-\hat{F})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \rightsquigarrow N(0,1)$$

Où

$$\hat{F} = \frac{n_1 F_1 + n_2 F_2}{n_1 + n_2}$$

Remarquons que $\hat{F}(1-\hat{F})$ représente la variance empirique de $\hat{F}$. Il y a donc une analogie entre le test de comparaison de moyennes sur échantillons indépendants et le test de comparaison de proportion sur échantillon indépendants.

Toutefois ce test n'est valable que lorsque les conditions suivantes sont vérifiées :

$$\begin{cases} n_1 \geq 30\,;\, n_1 p_1 \geq 5\,;\, n_1(1-p_1) \geq 5 \\ n_2 \geq 30\,;\, n_2 p_2 \geq 5\,;\, n_2(1-p_2) \geq 5 \end{cases}$$

Car on considère que l'approximation des lois de X_1 et X_2 (lois binomiales) par la loi normale est acceptable si ces conditions sont remplies.

Quant à la règle de décision, elle est la même que dans le cas du test de comparaison de moyennes.

La règle générale est la suivante :

- Dans l'hypothèse alternative où $p_1 < p_2$, l'intervalle d'acceptation de l'hypothèse nulle est
$$IA_{1-\alpha}(Z) = [-Z^*_{1-\alpha}, +\infty\,[$$
- Dans l'hypothèse alternative où $p_1 > p$, l'intervalle d'acceptation de l'hypothèse nulle est :
$$IA_{1-\alpha}(Z) = [-\infty, Z^*_{1-\alpha}\,[$$
- Dans l'hypothèse alternative où $p_1 \neq p_2$, l'intervalle d'acceptation de l'hypothèse nulle est :

$$IA_{1-\frac{\alpha}{2}}(Z) = [-Z^*_{1-\frac{\alpha}{2}}, Z^*_{1-\frac{\alpha}{2}}\,[$$

Où $Z^*_{1-\alpha}$ ou $Z^*_{1-\frac{\alpha}{2}}$ sont des seuils théoriques lus dans la table de la loi normale au seuil $1-\alpha$ (test unilatéral) ou au seuil $1-\frac{\alpha}{2}$ (test bilatéral).

Dans le cas du test bilatéral au seuil $= 0{,}05$, $Z^*_{1-\frac{\alpha}{2}} = Z^*_{0{,}975}$ =1,96. Dès lors, l'intervalle d'acceptation de H0 est $[-1{,}96\,,1{,}96\,]$.

Application

$$\hat{F} = \frac{(85 \times 0{,}612) + (78 \times 0{,}462)}{85 + 78} = 0{,}540$$

$$Z = \frac{0{,}612 - 0{,}462}{\sqrt{0{,}540(1-0{,}540)\left(\frac{1}{85}+\frac{1}{78}\right)}} = 1{,}92$$

Comme la valeur de Z est comprise dans l'intervalle d'acceptation $[-1{,}96\,,1{,}96\,]$, on ne peut donc pas rejeter l'hypothèse nulle d'égalité de proportion entre les deux échantillons. Cela signifie donc que la proportion de filles n'est pas significativement différente les deux cursus.

Dans un test unilatéral à gauche ($p_1 < p_2$), l'intervalle d'acceptation aurait été $[-1{,}64\,,+\infty]$ car $Z^*_{1-\alpha} = Z^*_{0{,}95} = 1{,}64$

Et dans un test unilatéral à droite ($p_1 > p_2$), l'intervalle d'acceptation aurait été $[-\infty\,,1{,}64]$ car $Z^*_{1-\alpha} = Z^*_{0{,}95} = 1{,}64$.

4.4.4. Test de comparaison de deux variances

Les tests de comparaison de moyennes précédemment présentés sont réalisés en supposant que la variance était la même selon les échantillons (hypothèse d'égalité de variance). Mais le non-respect de cette hypothèse rend invalide les

tests de comparaison présentés. C'est pourquoi, il est indispensable de vérifier l'égalité des variances avant de mettre en œuvre ces tests. Dans cette section, nous allons étudier les tests de comparaison de variances. Ces tests seront présentés selon qu'ils s'agisse des échantillons indépendants ou des échantillons appariés.

4.4.4.1. Cas d'échantillons indépendants

Exemple d'introduction

Dans une pisciculture, on étudie l'effet de deux régimes alimentaires, que l'on appellera "régime 1" et "régime 2", sur la croissance d'une espèce de poisson.

Pour ce faire, on alimente un lot de poissons avec le régime 1 et un autre lot avec le régime 2. On mesure la longueur X_1 de $n_1 = 9$ poissons choisis au hasard nourris avec le régime 1 et la longueur X_2 de $n_2 = 8$ poissons choisis au hasard nourris avec le régime 2. On observe les résultats suivants :

		1	2	3	4	5	6	7	8	9
Régime 1	X_1	21.18	20.01	22.50	22.97	21.83	23.42	18.61	25.20	22.07
Régime 2	X_2	22.39	21.26	22.17	25.00	22.21	20.51	22.36	24.49	

On souhaite alors tester à un seuil d'erreur de 10% si l'égalité de variance entre les deux populations de poissons est vérifiée. L'hypothèse nulle de ce test se présente alors comme suit : $H_0 : {\sigma_1}^2 = {\sigma_2}^2$ contre l'alternative $H_1 : {\sigma_1}^2 \neq {\sigma_2}^2$

Où ${\sigma_1}^2$ et ${\sigma_2}^2$ représentent respectivement les variances théoriques sur les deux échantillons.

On peut aussi proposer une reformulation de cette hypothèse. En effet,

$$ {\sigma_1}^2 = {\sigma_2}^2 \Rightarrow \frac{{\sigma_1}^2}{{\sigma_2}^2} = 1 $$

L'hypothèse nulle du test se présente alors comme suit : $H_0 : \frac{{\sigma_1}^2}{{\sigma_2}^2} = 1$ contre l'alternative $H_1 : \frac{{\sigma_1}^2}{{\sigma_2}^2} \neq 1$.

Pour réaliser ce test, on se sert des variances empiriques modifiées ${S_1}^2$ et ${S_2}^2$ (puisque les variances théoriques ne sont pas connues).

Rappels

$$ {S_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \overline{X_1})^2 \;\; ; \;\; {S_2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2i} - \overline{X_2})^2 $$

$$X_1 \rightsquigarrow N(m_1, {\sigma_1}^2)\ ;\ \frac{(n_1-1){S_1}^2}{{\sigma_1}^2} \rightsquigarrow \chi^2\,(n_1-1)$$

$$X_2 \rightsquigarrow N(m_2, {\sigma_2}^2)\ ;\ \frac{(n_2-1){S_2}^2}{{\sigma_2}^2} \rightsquigarrow \chi^2\,(n_2-1)$$

En remplaçant les variances théoriques par leur équivalent empirique, l'hypothèse nulle se présente comme suit :

$$H_0 : \frac{{\sigma_1}^2}{{\sigma_2}^2} = \frac{{S_1}^2}{{S_2}^2} = 1$$

$$\Rightarrow H_0 : \frac{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1i}-\overline{X_1})^2}{\frac{1}{n_2-1}\sum_{i=1}^{n_2}(X_{2i}-\overline{X_2})^2} = 1$$

$$\Rightarrow H_0 : \frac{\frac{\sum_{i=1}^{n_1}(X_{1i}-\overline{X_1})^2}{(n_1-1)}}{\frac{\sum_{i=1}^{n_2}(X_{2i}-\overline{X_2})^2}{(n_2-1)}} = 1$$

On remarque que cette quantité correspond à une loi de Fisher car c'est le rapport entre deux lois de khi-deux (normalisées par leurs degrés de liberté respectifs). Ainsi, on a :

$$H_0 : \frac{{\sigma_1}^2}{{\sigma_2}^2} = \frac{{S_1}^2}{{S_2}^2} \rightsquigarrow F\,(n_1, n_2)$$

Dès lors tester l'égalité de variances revient à tester que la statistique F calculée soit égale à 1 (ou presque). On accepte H0 lorsque la statistique F calculée est suffisamment proche de 1. Dans le cas contraire on rejette H0.

De façon pratique, dans un test bilatéral, on recherche la position de la valeur de la statistique $F = \frac{{S_1}^2}{{S_2}^2}$ par rapport à l'intervalle $F^*_{\frac{\alpha}{2}}\,(n_1-1, n_2-1)$ et $F^*_{1-\frac{\alpha}{2}}\,(n_1-1, n_2-1)$.

L'intervalle d'acceptation de H0 se définit comme suit :

$$IA = [F^*_{\frac{\alpha}{2}}\,(n_1-1, n_2-1), F^*_{1-\frac{\alpha}{2}}\,(n_1-1, n_2-1)$$

Lorsque F est située à l'extérieur de cet intervalle, on rejette H0 d'égalité de variance au risque α. En revanche, lorsque F est compris entre $F^*_{\frac{\alpha}{2}}\,(n_1-1, n_2-1)$ et $F^*_{1-\frac{\alpha}{2}}\,(n_1-1, n_2-1)$, on ne peut pas rejeter H0.

Propriétés : pour une loi de Fisher : on a la propriété suivante :

$$F_{\gamma}^{*}\,(a,b) = \frac{1}{F_{1-\gamma}^{*}\,(b,a)}$$

En utilisant cette propriété, on peut donc écrire que

$$F_{\frac{\alpha}{2}}^{*}\,(n_1-1,n_2-1) = \frac{1}{F_{1-\frac{\alpha}{2}}^{*}\,(n_2-1,n_1-1)}$$

Ainsi, l'intervalle d'acceptation se présente comme suit :

$$IA = [\frac{1}{F_{1-\frac{\alpha}{2}}^{*}\,(n_2-1,n_1-1)}, F_{1-\frac{\alpha}{2}}^{*}\,(n_1-1,n_2-1)$$

Application :

${S_1}^2 = 3,7\ ;\ {S_2}^2 = 2,27\ ;\ F = \frac{3,7}{2,27} = 1,63\ ;\ \alpha = 0,10$

$$F_{1-\frac{\alpha}{2}}^{*}\,(n_2-1,n_1-1) = F_{1-\frac{0,1}{2}}^{*}\,(8,9) = F_{0,95}^{*}\,(8,9) = 3,5$$

$$F_{1-\frac{\alpha}{2}}^{*}\,(n_1-1,n_2-1) = F_{1-\frac{0,1}{2}}^{*}\,(9,8) = F_{0,95}^{*}\,(9,8) = 3,73$$

Ainsi,

$$IA = [\frac{1}{3,5}, 3,73]$$

$$IA = [0,29\ , 3,73]$$

On remarque que la valeur de F est comprise dans cet intervalle, alors, on ne peut pas rejeter l'hypothèse nulle.

4.4.4.2. Cas d'échantillons appariés

Exemple d'introduction

Pour dix couples, on a observé le salaire mensuel de l'homme et celui de la femme. Les résultats (mesurés en milliers) sont rassemblés dans le tableau suivant :

	1	2	3	4	5	6	7	8	9	10
Homme	1,699	2,347	1,531	3,476	2,464	1,544	2,59	2,791	2,661	1,956
Femme	2,569	2,189	0,728	1,009	2,871	1,199	2,059	2,988	2,786	2,04

Effectuer un test de comparaison des variances des salaires des hommes et de celles des femmes (NB : Bien que les observations n'ont pas été effectuées sur les mêmes individus, ces deux échantillons peuvent être considérés comme appariés car il y a une correspondance théorique qui découle ici de la notion de couple).

Les deux échantillons ont même taille n. L'hypothèse nulle H0 est que les deux échantillons ont même variance :

$$H_0 : {\sigma_1}^2 = {\sigma_2}^2 \Rightarrow H_0 : \frac{{\sigma_1}^2}{{\sigma_2}^2} = 1$$

Pour pouvoir mettre en œuvre le test d'égalité de variance sur échantillons appariés, on s'appuie sur la propriété suivante :

Théorème : Si X_1 et X_2 sont des variables aléatoires et si on pose $U = X_1 + X_2$ et $V = X_1 - X_2$ alors on a :

$$\frac{{\sigma_1}^2}{{\sigma_2}^2} = 1 \Rightarrow r_{UV} = 0$$

Où r_{UV} est le coefficient de corrélation de Pearson (empirique) entre U et V.

$$r_{UV} = \frac{Cov(U,V)}{\sqrt{{S_U}^2 \times {S_V}^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(U_i - \bar{U})(V_i - \bar{V})}{\sqrt{\frac{1}{n}(\sum_{i=1}^{n}(U_i - \bar{U})^2 \times \sum_{i=1}^{n}(V_i - \bar{V})^2)}}$$

Dès lors, tester l'égalité de variance sur deux variables X_1 et X_2 sur échantillons appariés revient à tester la nullité du coefficient de corrélation linéaire entre leur transformée U et V. Ce test alors appelé test de **Pitman**. L'hypothèse nulle se présente comme suit :

$$H_0 : r_{UV} = 0$$

Sous l'hypothèse nulle, la statistique du test se présente comme suit :

$$\frac{\hat{r}_{UV}}{\sqrt{\frac{1 - \hat{r}_{UV}^2}{n-2}}} \rightsquigarrow T(n-2)$$

Où est le $\hat{r}_{UV}$ est le coefficient de corrélation empirique. Cette statistique suit une loi de Student à n-2 degrés de liberté. Sa valeur est comparée à la valeur théorique obtenue dans la table de Student au seuil α fixé.

Quant à la règle de décision, le principe général est le suivant :

- Dans l'hypothèse alternative où ${\sigma_1}^2 < {\sigma_2}^2$, l'intervalle d'acceptation de l'hypothèse nulle est

$$IA_{1-\alpha}(T) = [-T^*_{1-\alpha}, +\infty\,[$$

- Dans l'hypothèse alternative où ${\sigma_1}^2 > {\sigma_2}^2$, l'intervalle d'acceptation de l'hypothèse nulle est :

$$IA_{1-\alpha}(T) = [-\infty, T^*_{1-\alpha}\,[$$

- Dans l'hypothèse alternative où ${\sigma_1}^2 \neq {\sigma_2}^2$, l'intervalle d'acceptation de l'hypothèse nulle est :

$$IA_{1-\frac{\alpha}{2}}(T) = [-T^*_{1-\frac{\alpha}{2}}, T^*_{1-\frac{\alpha}{2}}\,[$$

Où $T^*_{1-\alpha}$ ou $T^*_{1-\frac{\alpha}{2}}$ sont des seuils théoriques lus dans la table de Student au seuil $1-\alpha$ (test unilatéral) ou au seuil $1-\frac{\alpha}{2}$ (test bilatéral).

Remarque

Ce résultat est valable uniquement si on peut supposer que les variables U et V sont gaussiennes (ou plus exactement que le couple (U, V) suit une loi normale bivariée). Le test de significativité est alors en réalité un test d'indépendance.

Si ce n'est pas le cas, le résultat est seulement asymptotique et n'est donc valable que si l'échantillon est suffisamment grand.

Application :

	1	2	3	4	5	6	7	8	9	10
Homme	1,699	2,347	1,531	3,476	2,464	1,544	2,59	2,791	2,661	1,956
Femme	2,569	2,189	0,728	1,009	2,871	1,199	2,059	2,988	2,786	2,04
U	4,268	4,536	2,259	4,485	5,335	2,743	4,649	5,779	5,447	3,996
V	-0,87	0,158	0,803	2,467	-0,407	0,345	0,531	-0,197	-0,125	-0,084

On calcule le coefficient de corrélation empirique entre les variables U et V. On a :

$$\hat{r}_{UV} = -0{,}263$$

On en déduit la statistique du test :

$$T = \frac{-0{,}263}{\sqrt{\frac{1-(-0{,}263)^2}{10-2}}} = -0{,}771$$

La valeur critique de la loi de Student à 8 degrés de liberté au seuil = 5% vaut $T^*_{1-\frac{\alpha}{2}}(8) = T^*_{0{,}975}(8) =$ 2,306. Ainsi l'intervalle d'acceptation est [-2,306 ;

2,306]. La valeur -0.771 se trouve dans l'intervalle d'acceptation, alors on accepte donc l'hypothèse H0.

4.5. Les tests d'adéquation

Les tests d'adéquation servent à tester si un échantillon est distribué selon une loi de probabilité donnée. Ils permettent de décider, avec un seuil d'erreur α spécifié, si les écarts présentés par l'échantillon par rapport aux valeurs théoriques attendus sont dûs au hasard ou sont au contraire significatifs.

Exemple 1

Une entreprise de BTP a dénombré, pendant une période de 100 jours ouvrés, le nombre quotidien d'accidents du travail qui se produisent sur ses chantiers :

Nb journalier d'accidents	0	1	2	3	4	5	≥ 6
Effectifs	11	27	29	20	11	1	1

Peut-on considérer que le nombre d'accidents suit une loi de poisson ?

Exemple 2

On a un échantillon de 100 notes obtenues à un examen scolaire dans une école :

6.5	9.5	5.5	15.5	10.5	5.5	11	12	11.5	8
15	10.5	6.5	0	13.5	9	9	13	12.5	11.5
12.5	12	9.5	1	11.5	9	8.5	3	7	10.5
14.5	8.5	10.5	9	3.5	7.5	7.5	9	13.5	12
8.5	8	12	11	6	6	10.5	12	8.5	12.5
10.5	6.5	10.5	4.5	14.5	17	7.5	5	11.5	8.5
18.5	9	12	9	6	10	2	15	9.5	17.5
11	6	11.5	5.5	4	10	7	9	9.5	6.5
6.5	8.5	13.5	3	11.5	10.5	13.5	8	10.5	10
7	14	13.5	12	15.5	11	4	6.5	4	7

Le directeur de l'établissement s'attend avoir une répartition normale des notes autour de la moyenne de 10 avec un écart-type de 5. Cet échantillon est-il conforme aux attentes du Directeur ?

Les tests d'adéquation servent à répondre à ce genre de questions. Ils sont utilisés dans plusieurs autres situations :

1. ils permettent de vérifier en particulier l'hypothèse de normalité faite dans beaucoup de tests ;

2. en simulation stochastique, ils permettent de tester qu'un processus qui génère des échantillons selon une loi donnée produit des valeurs acceptables ;

3. ils permettent de tester le bon fonctionnement d'un appareil ;

4. en économétrie, ils permettent de vérifier a posteriori les hypothèses probabilistes faites sur les modèles (homoscédasticité des erreurs, normalité, etc.).

On distingue deux principaux types de tests d'adéquation reposant chacun sur des approches différentes du problème : le test de khi-deux et le test de Kolmogorov-Smirnov.

1. le test de khi-deux (χ^2) convient aux données qui présentent un nombre fini de modalités (des variables discrètes finies ou des variables continues qu'on a regroupées en un nombre fini de classes). Il est souvent qualifié de test d'indépendance lorsqu'il porte exclusivement sur des variables de nature qualitative. De ce point de vue, le test d'indépendance est un cas particulier du test d'adéquation du khi-deux.

2. le test de Kolmogorov-Smirnov (beaucoup plus général) s'applique pour des distributions continues.

3. Bien entendu, il existe d'autres tests d'adéquation tels que ceux de Cramer-Von Mise sur lequel nous reviendrons plus tard.

4.5.1. Le test de khi-deux d'adéquation (χ^2)

4.5.1.1. Présentation du test

Conceptuellement, il s'agit de comparer une distribution expérimentale avec une loi de probabilité théorique.

L'hypothèse nulle est:

H0: il y a adéquation de la distribution observée avec la distribution théorique

Soient O_i les effectifs observés pour chaque classe d'évènements et C_i les effectifs calculés qui sont les effectifs théoriques qu'on obtiendrait en appliquant la loi théorique.

Théorème : Sous l'hypothèse H0 (d'adéquation), on montre que la variable aléatoire D définie par :

$$D = \sum_{i=1}^{n} \frac{(O_i - C_i)^2}{C_i} \rightsquigarrow \chi^2(\mathrm{k} - 1)$$

Où k est le nombre de valeurs prises par x (non les effectifs)

La statistique D s'appelle la distance du χ^2 entre le vecteur des valeurs observées et celui des valeurs calculées.

Pour que ce test soit valide, il faut que $C_i > 5$ pour tout i. En particulier, on ne doit pas l'appliquer si une classe a un effectif nul.

Dans le cas d'une variable continue, les données doivent être regroupées en classes et chaque classe peut être représentée par une valeur unique (par exemple, le milieu de la classe). Le nombre n représente alors le nombre de classes tandis que N représente l'effectif total de l'échantillon.

En utilisant les tables du χ^2, on détermine le quantile qui délimite la région d'acceptation au seuil α fixé. Si la distance D est supérieure au quantile, on rejette l'hypothèse H0.

Autrement dit – en termes de p-valeur – si la probabilité d'avoir une distance supérieure à D est inférieure à α, on rejette H0.

Remarque importante sur la statistique D

La formule de D présentée ci-haut reste, quelque part, incomplète car elle ne prend pas en compte le nombre de contraintes supplémentaires qui peuvent surgir dans le processus de test. Il arrive par exemple qu'on ait besoin d'abord d'estimer les paramètres de la loi de référence avant d'appliquer sa formule aux données pour calculer la distribution théorique. Par exemple on peut énoncer dans l'exercice de tester l'adéquation des données à une loi de poisson sans nécessairement fournir les paramètres de cette loi. Dès lors pour calculer la distribution théorique, il faut estimer les paramètres de la loi de poisson à partir de l'échantillon et utiliser cette valeur dans le calcul. La statistique D doit donc tenir compte de cette situation. Pour cela, il faut simplement diminuer le nombre de degrés de liberté en soustrayant le nombre de paramètres estimés. Ainsi de façon générale, la formule se présente alors comme suit :

$$D = \sum_{i=1}^{n} \frac{(O_i - C_i)^2}{C_i} \rightsquigarrow \chi^2(\,\mathrm{k} - 1 - \mathrm{r})$$

Où k est le nombre de valeurs prises par x et r est le nombre de paramètres estimé en amont avant de calculer D.

4.5.1.2. Quelques exemples d'application du test de khi-2 d'adéquation

Cas 1 : Test d'adéquation à une loi de poisson

Reprenons l'exemple introductif 1 sur le nombre journalier d'accidents.

D'abord, les deux dernières classes (k = 5 et k = 6) ayant un effectif inférieur à 5, on va les regrouper avec la classe k = 4. On obtient le tableau suivant :

Nb journalier d'accidents	0	1	2	3	≥ 4
Effectifs	11	27	29	20	13

A présent, nous avons un tableau à 5 classes, k = 5. Attention, à ce niveau à ne pas confondre N qui est l'effectif total (100) et k qui est le nombre de classes (5).

L'énoncé ne précise pas le paramètre λ de la loi de Poisson. On sait que pour une variable X qui suit une loi de Poisson P(λ), on a E(X) = λ. On va donc estimer λ au moyen de la moyenne empirique obtenue à partir de l'échantillon (NB : En règle générale, l'énoncé fournit les paramètres de la loi théorique comme par exemple dans l'exemple 2 où l'on donne les paramètres de la loi normale. Mais lorsque les paramètres ne sont pas donnés, il faut calculer leur équivalent empirique à partir de l'échantillon. Dans le cas d'un test de khi-deux ce calcul doit être fait après le regroupement des classes dont les effectifs sont inférieurs à 5).

En utilisant le nouveau tableau, la valeur estimée du paramètre λ est

$$\hat{\lambda} = \frac{1}{100}(11 \times 0 + 27 \times 1 + 29 \times 2 + 20 \times 3 + 13 \times 4) = 1{,}97 \cong 2$$

Calculons maintenant les effectifs théoriques en utilisant une loi de Poisson de paramètre λ =2. On sait que pour la loi de poisson la probabilité est définie par pour tout x comme suit :

$$P(X = k) = e^{-\lambda}\left(\frac{\lambda^k}{k!}\right)$$

Ici k=0, 1 , 2 , 3, et ≥ 4. Et remplaçant ces valeurs dans la formule, on obtient les probabilités suivantes :

$P(X = 0) = 0{,}135$; $P(X = 1) = 0{,}271$; $P(X = 2) = 0{,}271$; $P(X = 3) = 0{,}180$

La distribution de la loi de poisson n'étant pas bornée en théorie, pour calculer la probabilité de la classe $k \geq 4$ il faut procéder par soustraction étant donné que la somme des densités de probabilités est égale à 1.

Ainsi on a :

$$P(X \geq 4) = 1 - \big(P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)\big) = 0{,}143$$

En utilisant ainsi ces probabilités (proportions) comme des clés de répartition pour répartir l'échantillon, on a :

k	0	1	2	3	≥ 4
O_i	11	27	29	20	13
P(X=k)	0,135	0,271	0,271	0,180	0,143
C_i	13,5	27,1	27,1	18	14,3

On peut maintenant calculer la statistique $D = \sum_{i=1}^{4} \frac{(O_i - C_i)^2}{C_i}$. Et on trouve

$$D = 0{,}937$$

Cette valeur doit être comparée à la statistique de khi-deux lue dans la table au seuil de 5%. Etant donné qu'on a estimé le paramètre λ à partir de l'échantillon, il y a donc une contrainte supplémentaire qui diminue le nombre de degré de liberté. Ainsi le nombre de degré de liberté égale à (n-1-r) avec r=1. D'où ddl= 5-1-1 =3. (On a $\chi^2{}_{1-\alpha}(n-1-1) = \chi^2{}_{0,95}(3) = 7{,}81$.

Comme la statistique calculée est nettement en dessous de cette valeur, on ne peut pas rejeter l'hypothèse H0.

Remarque

Plutôt que de travailler sur les effectifs, on pouvait aussi réaliser le test à partir des proportions. Pour cela, on calculait d'abord les fréquences observés telles que :

$$f_i = \frac{O_i}{N}$$

(où N est l'effectif total de l'échantillon) et on note p_i les probabilités théoriques.

$$p_i = P(X = i) = e^{-\lambda}\left(\frac{\lambda^i}{i!}\right)$$

La statistique du χ^2 devient :

$$D = \sum_{i=1}^{n} \frac{(O_i - C_i)^2}{C_i} = \sum_{i=1}^{n} \frac{(f_i N - p_i N)^2}{p_i N} = \sum_{i=1}^{n} \frac{(f_i - p_i)^2}{p_i} = \rightsquigarrow \chi^2(\mathrm{n}-1)$$

Cas 2 : Test d'adéquation à une loi normale

Exemple 1 : Cas où les paramètres sont connus

Reprenons l'exemple introductif 2 sur les notes de classe des 100 élèves

D'abord, on construit le tableau d'effectifs en réunissant en des classes.

Classes	[0; 5]	[5; 8[	[8; 12[	[12; 15[	[15; 20]
Effectifs	10	21	43	19	7

Pour calculer les effectifs théoriques, on doit calculer la probabilité de chaque classe pour la loi normale N(10; 5) en centrant et en réduisant à chaque fois pour obtenir une loi normale (0,1).

En effet, on sait que :

$$X \rightsquigarrow N(m, \sigma^2) \Rightarrow Z = \left(\frac{X - m}{\sigma}\right) \rightsquigarrow N(0,1)$$

Par exemple :

$$P(0 < X < 5) = P(X < 5) - P(X < 0)$$

En centrant et en réduisant par la moyenne et l'écart-type, on a :

$$P\left(\frac{0-10}{5} < \frac{X-10}{5} < \frac{5-10}{5}\right) = P\left(\frac{X-10}{5} < \frac{5-10}{5}\right) - P\left(\frac{X-10}{5} < \frac{0-10}{5}\right)$$

$$P(-2 < Z < -1) = P(Z < -1) - P(Z < -2)$$

On lit alors simple $P(Z < -1)$ et $P(Z < -2)$ à partir de la table de la loi normale (0,1). On peut effectuer rapidement le calcul sous excel avec la fonction : =LOI.NORMALE.STANDARD.N(z;VRAI) où Z est la valeur dont on cherche la probabilité.

On a alors :

$$P(Z < -1) = 0{,}159$$

$$P(Z < -2) = 0{,}022$$

Alors $P(0 < X < 5) = P(-2 < Z < -1) = P(Z < -1) - P(Z < -2) = 0{,}159 - 0{,}022 = 0{,}136$

De même pour $P(5 < X < 8)$, on a :

$$P(5 < X < 8) = P(Z < -0{,}4) - P(Z < -1) = 0{,}345 - \ 0{,}1590{,}186$$

En suivant la même démarche pour les autres classes et en déduisant la probabilité de la dernière classe par soustraction de 1, on obtient finalement le tableau théorique suivant (Noter qu'on pouvait aussi effectuer le calcul de probabilité en utilisant uniquement le centre de classe plutôt que d'utiliser le calcul par encadrement. Mais le calcul par encadrement est beaucoup plus précis car il n'y a pas de perte d'information).

i	1	2	3	4	5
O_i	10	21	43	19	7
p_i	0.136	0.186	0.311	0.186	0.136
C_i	13.6	18.6	31.1	18.6	13.6

Ainsi, en calculant la statistique D, on obtient $D = 11{,}095$.

La table du khi-2 avec 5 - 1 = 4 degrés de liberté fournit la valeur critique 9.488. Comme 11,095 > 9,488, on rejette l'hypothèse d'adéquation. On conclut, avec un risque de 5% de se tromper, que les valeurs observées ne sont pas distribuées selon une loi normale N(10; 5). Si on calcule la moyenne et l'écart-type empiriques, on obtient $\bar{X}$= 9,455 et s = 3,61.

Exemple 2 : Cas où les paramètres sont connus

Le caractère « taille » a été mesurée sur un échantillon d'individus. Les résultats en classes sont consignés dans le tableau suivant.

x_i (Taille en cm)	<155	[155-165 [	[165-175 [	[175-185 [	>185
n_i (Nombre d'individus)	6	70	500	379	50

Peut-on considérer que la taille suit une loi normale ?

Dans cet exercice les paramètres de la loi suivie ne sont pas fournis. Il faut alors les estimer.

On sait que $X \rightsquigarrow N(m, \sigma^2)$ et que $\bar{X}$ et S^{*2} sont des estimateurs sans biais respectivement de m et de σ^2.

$$\bar{X} = \frac{1}{k}\sum_{i=1}^{k} n_i x_i$$

$$S^{*2} = \frac{1}{k-1}\sum_{i=1}^{k} n_i (x_i - \bar{X})^2$$

La variable X étant présentée sous forme de classes, pour calculer la moyenne empirique et la variance empirique, on considère les centres de classes (pour les classes intermédiaires) et on réajuste les valeurs sur les extrémités en tenant compte de la longueur entre les classes. (On pouvait aussi simplement laisser comme telles les bornes pour les valeurs aux extrémités). Ainsi, on a :

x_i (Taille en cm)	155	160	170	180	185
n_i (Nombre d'individus)	6	70	500	379	50

En calculant à partir de ces valeurs, on obtient : $\bar{X} = 174$ et $S^{*2} = 51{,}4$.

Connaissant maintenant les deux paramètres de la loi normale, on peut reprendre la démarche du test en représentation en classes. On va apporter une légère modification à la présentation des classes afin de faciliter le calcul des probabilités. D'abord, pour la dernière classe (>185), on peut calculer sa probabilité par soustraction des autres probabilités de 1. Mais pour la première classe (<155), il faut fixer une borne inférieure « fictive » en cohérence avec les autres classes mais surtout avec la réalité. On choisit par exemple [145 ; 155]. On a alors :

x_i (Taille en cm)	[145 ; 155]	[155-165 [	[165-175 [	[175-185 [	>185
n_i (Nombre d'individus)	6	70	500	379	50

On sait que :

$$X \rightsquigarrow N(m, \sigma^2) \Rightarrow \left(\frac{X-m}{\sigma}\right) \rightsquigarrow N(0,1)$$

Mais en utilisant $\bar{X}$ à la place de m et S^* à la place de σ, on a : $\left(\frac{X-\bar{X}}{S^*}\right) \rightsquigarrow T(n-1)$

Mais on sait aussi que lorsque la taille de l'échantillon est très importante, on peut approximer la loi de Student par la loi normale. Dès lors on peut écrire :

$$\left(\frac{X-\bar{X}}{S^*}\right) = \rightsquigarrow N(0,1)$$

On peut alors calculer les probabilités comme suit :

$$P(145 < \mathrm{X} < 155) = P(\mathrm{X} < 155) - P(\mathrm{X} < 145)$$

En centrant et en réduisant, on alors le tableau suivant

$$P\left(\frac{145-\bar{X}}{S^*} < \frac{\mathrm{X}-\bar{X}}{S^*} < \frac{155-\bar{X}}{S^*}\right) = P\left(\frac{\mathrm{X}-\bar{X}}{S^*} < \frac{155-\bar{X}}{S^*}\right) - P\left(\frac{\mathrm{X}-\bar{X}}{S^*} < \frac{145-\bar{X}}{S^*}\right)$$

$$P(-4{,}04 < Z < -2{,}64) = P(Z < -2{,}64) - P(Z < -4{,}04)$$

Le tableau suivant résume les résultats du calcul.

X	$Z = \frac{X - \bar{X}}{S^*}$	$P(\mathrm{Z} < \mathrm{z})$
145	-4,04	2,7E-05
155	-2,64	0,004
165	-1,25	0,11
175	0,15	0,56
185	1,54	0,94

$$P(145 < \mathrm{X} < 155) = 0{,}004 - 2{,}710^{-05} = 0{,}00408$$

$$P(155 < \mathrm{X} < 165) = 0{,}11 - 0{,}004 = 0{,}1018$$

$$P(165 < \mathrm{X} < 175) = 0{,}56 - 0{,}11 = 0{,}45$$

$$P(175 < \mathrm{X} < 185) = 0{,}94 - 0{,}56 = 0{,}38$$

$$P(\mathrm{X} > 185) = 1 - (0{,}00408 + 0{,}1018 + 0{,}45 + 0{,}38) = 0{,}062$$

N'oublions pas que s'il y avait des classes dont l'effectif est inférieur à 5, on allait les regrouper avec les classes les plus proches avant de calculer les probabilités théoriques.

On obtient alors le tableau suivant :

x_i	[145 ; 155]	[155-165 [	[165-175 [	[175-185 [	>185
$n_i = O_i$	6	70	500	379	50
p_i	0,00408	0,1018	0,45	0,38	0,062
$C_i = p_i \times N = 1005 \times p_i$	4,10	102,37	454,49	382,05	61,99
$\frac{(O_i - C_i)^2}{C_i}$	0,88	10,23	4,56	0,02	2,32

$$D = \sum_{i=1}^{n} \frac{(O_i - C_i)^2}{C_i} = 18{,}01$$

Cette statistique suit une loi de khi-deux à k-1-r degrés de liberté. Ici r=2 car nous avons deux paramètres estimés : la moyenne et la variance. Donc ddl=5-1-2=2. Au seuil $\alpha = 0{,}05$, on a $=\chi^2_{0,95}(2) = 5{,}99$.

La statistique calculée étant supérieure à la valeur tabulée, on rejette l'hypothèse nulle d'adéquation à une loi normale N(174 ; 51,41).

Cas 3 : Test d'adéquation à une loi binomiale

Exemple 1 : Cas où le paramètre p est connu

On souhaite tester si la distribution du nombre de filles dans une fratrie de 5 enfants suit-elle une loi binomiale de paramètres 5 et 0,5 c'est-à-dire $B(5, 0{,}5)$. On dispose pour cela le tableau de répartition sur 320 fratries de 5 enfants.

x_i (Nombre filles)	0	1	2	3	4	5
n_i (Nombre fratries)	18	56	110	88	40	8

Dans cet exemple, il s'agit de tester l'adéquation des données à une loi binomiale dont les paramètres sont connus. Ici $p = 0{,}5$. Il faut alors commencer par élaborer le tableau de fréquence théorique en utilisant la fonction de densité de la loi binomiale. On a :

$$P(x_i) = P(x_i = k) = C_n^k p^k (1-p)^{n-k}$$

Ici k=0, 1 , 2 , 3, 4 et 5. Et remplaçant ces valeurs dans la formule, on obtient les probabilités suivantes :

$P(x_i = 0) = 0{,}031$; $P(x_i = 1) = 0{,}156$; $P(x_i = 2) = 0{,}313$; $P(x_i = 3) = 0{,}313$; $P(x_i = 4) = 0{,}156$; $P(x_i = 1) = 0{,}031$

On peut rapidement effectuer ces calculs en utilisant la fonction excel :

=LOI.BINOMIALE.N(k;5;0,5;FAUX) où k représente les différentes valeurs.

On a alors le tableau de fréquences théoriques avec lequel on peut calculer les effectifs théoriques. On a :

x_i	0	1	2	3	4	5
$n_i = O_i$	18	56	110	88	40	8
p_i	0,031	0,156	0,313	0,313	0,156	0,031
$C_i = p_i \times N = 320p_i$	10	50	100	100	50	10
$\frac{(O_i - C_i)^2}{C_i}$	6,4	0,72	1	1,44	2	0,4

$$D = \sum_{i=1}^{n} \frac{(O_i - C_i)^2}{C_i} = 11{,}96$$

Cette statistique suit une loi de khi-deux à (n-1-r) degrés de liberté. Ici k=6 et r=0 car aucun paramètre de la loi binomiale n'a été estimé. Donc ddl=5. Avec $\alpha = 0{,}05$ la valeur $\chi^2(5) = 11{,}07$.

La valeur calculée étant supérieure à la valeur lue, on rejette l'hypothèse nulle d'adéquation à une loi binomiale $B(5, 0{,}5)$ du nombre de filles dans une fratrie de 5 enfants.

Exemple 2 : Cas où le paramètre p n'est pas donné

Reprenons l'exemple précédent en modifiant légèrement l'énoncé où la probabilité p n'est pas donnée. On souhaite simplement tester si la distribution du nombre de filles dans une fratrie de 5 enfants suit une loi binomiale.

Dans cette configuration, puisque p n'est pas connue, il faut alors l'estimer à partir des données. Reprenons le tableau initial :

x_i (Nombre filles)	0	1	2	3	4	5
n_i (Nombre fratries)	18	56	110	88	40	8

Calculons la proportion empirique de filles dans un échantillon de 320 fratries.

Attention : Il ne s'agit pas de la proportion de fratries mais de la proportion de filles parmi tous les enfants sachant que chaque fratrie compte 5 enfants.

On peut alors voir que le nombre total d'enfants est égal au nombre de fratries multiplié par 5. Le tableau suivant résume le calcul :

x_i (Nombre filles)	0	1	2	3	4	5	Total
n_i (Nombre fratries)	18	56	110	88	40	8	**320**
Nombre total d'enfants	90	280	550	440	200	40	**1600**
Nombre de filles	0	56	220	264	160	40	**740**

On peut alors calculer la proportion empirique de filles telle que :

$$\hat{p} = \frac{740}{1600} = 0{,}4625$$

D'une manière générale la proportion estimée pour une loi binomiale B(n, $\hat{p}$) se présente comme suit :

$$\hat{p} = \frac{\sum_{i=1}^{k} n_i x_i}{n \sum_{i=1}^{k} n_i x_i}$$

La valeur estimée de p est alors $\hat{p} = 0{,}4625$. Il revient finalement de tester si le nombre de filles suit une B(n, $\hat{p}$). On suit alors étapes vues dans le premier exemple. Attention à, toutefois, prendre en compte dans le calcul du nombre de degrés de liberté la contrainte supplémentaire car un paramètre a été estimé donc r=1. Le nombre de degré de liberté est donc égale à 6-1-1=4.

x_i	0	1	2	3	4	5
$n_i = O_i$	18	56	110	88	40	8
p_i	0,045	0,193	0,3322	0,2858	0,123	0,021
$C_i = p_i \times N = 320p_i$	14,36	61,77	106,29	91,46	39,35	6,77
$\frac{(O_i - C_i)^2}{C_i}$	0,92	0,54	0,13	0,13	0,01	0,22

$$D = \sum_{i=1}^{n} \frac{(O_i - C_i)^2}{C_i} = 1{,}96$$

Avec $\alpha = 0{,}05$ la valeur $\chi^2(4) = 9{,}487$.

La valeur calculée étant largement inférieure à la valeur lue, on ne peut pas rejeter l'hypothèse nulle selon laquelle lu nombre de filles dans une fratrie de 5 enfants suit une loi binomiale $B(5, 0{,}46)$.

Ce résultat montre que la conclusion du test d'adéquation peut bien être différente selon que les paramètres soient connus ou pas.

4.5.2. Test du Khi-deux χ^2 d'indépendance

Le test d'indépendance du Khi-deux est une forme particulière du test d'adéquation du test de khi-2. Il concerne deux variables aléatoires discrètes ou continues avec un nombre fini de classes et opère sur la table de contingence qui donne les effectifs croisés des deux variables. Il permet de tester si les deux variables peuvent être considérées comme indépendantes.

Sa mise en œuvre est basée sur la statistique de test suivant :

$$D = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}} \rightsquigarrow \chi^2[(\mathrm{k}-1)(l-1)]$$

Où D est la distance de khi-deux (mesurant la distance entre le tableau théorique et le tableau observé) ; n_{ij} est l'effectif de la cellule du tableau de contingence « observé » situé au croisement de la ligne i et de la colonne j ; $\hat{n}_{ij}$ est l'effectif de la cellule du tableau de contingence « théorique » situé au croisement de la ligne i et de la colonne j (voir formule ci-dessous) ; k correspond au nombre de lignes du tableau de contingence (théorique ou observé). Il correspond au nombre de modalités de la variable en ligne ; l représente le nombre de colonne du tableau de contingence (théorique ou observé). Il correspond au nombre de modalités de la variable en colonne.

La statistique D est distribuée selon une loi de khi-2 à $(k-1)(l-1)$ degrés de libertés. $(k-1)(l-1) \geq 1$. NB : Lorsque $k = 1$, on retient $k(l-1)$ et lorsque $l = 1$, on a $l(k-1)$.

Dans un tableau de contingence, on note traditionnellement par $n_{i\bullet}$ la somme des effectifs sur une ligne i et par $n_{\bullet j}$ la somme des effectifs sur une colonne j. On montre que, s'il y a indépendance entre les deux variables, les effectifs théoriques dans chaque case du tableau de contingence se repartiraient comme suit :

$$\hat{n}_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

Plus clairement, il s'agit, pour chaque cellule du tableau, de multiplier le total-ligne et le total-colonne correspondant à cette cellule et en rapportant cette valeur à l'effectif total.

Dans le test d'indépendance, l'hypothèse nulle est toujours l'absence d'association entre les deux phénomènes étudiés. Ainsi, pour deux variables X_1 et X_2, lorsque

la statistique D calculée est supérieure la valeur du khi-deux lue dans la tableau à un seuil α donné et aux degrés de liberté $(k-1)(l-1)$, alors on rejette H0. On conclut alors que les deux variables ne sont pas indépendantes.

Exemples d'application

Exemple 1

La table suivante représente les résultats d'une enquête portant sur 300 étudiants à qui il a été demandé s'ils avaient une activité sportive régulière (S=Oui Sport/NS=Non Sport) et s'ils fumaient (F= Fumeur/NF= Non-fumeur).

	F	**NF**	Total
S	60	76	136
NS	56	108	164
Total	116	184	300

On peut par exemple soupçonner que le taux de fumeur est significativement faible parmi ceux qui pratiquent une activité sportive régulière, ou encore que la pratique du sport est relativement faible parmi les fumeurs relativement à des non-fumeurs. Dans les deux cas, on soupçonne une association entre les deux faits. Mais pour trancher la question, on va alors utiliser le test d'indépendance du khi-deux.

Ici, l'hypothèse H0 est qu'il y a indépendance entre le fait de fumer et le fait de pratiquer régulièrement le sport. On va alors calculer, sous l'hypothèse H0, les valeurs théoriques du tableau de contingence.

En appliquant la formule de répartition $\hat{n}_{ij}$ au tableau de contingence ci-dessus, on obtient le tableau d'effectifs théoriques suivants :

$$\hat{n}_{S,F} = \frac{136 \times 116}{300}$$

$$\hat{n}_{S,NF} = \frac{136 \times 184}{300}$$

$$\hat{n}_{NS,F} = \frac{164 \times 116}{300}$$

$$\hat{n}_{NS,NF} = \frac{164 \times 184}{300}$$

Tableau d'effectifs théoriques

	F	**NF**	Total
S	52.59	83.41	136
NS	63.41	100.59	164
Total	116	184	300

Il faut remarquer les totaux des lignes et des colonnes du tableau théorique sont exactement les mêmes que ceux du tableau d'effectifs observés.

Connaissant les effectifs théoriques, on peut maintenant calculer la statistique D.

$$D = \frac{(60-52{,}59)^2}{52{,}59} + \frac{(56-63{,}41)^2}{63{,}41} + \frac{(76-83{,}41)^2}{83{,}41} + \frac{(108-100{,}59)^2}{100{,}59}$$

$$D = 3{,}117$$

Le nombre de degrés de liberté est ici (p-1)(q-1) = (2-1)(2-1) = 1. La table du khi-2, indique une valeur critique égale à 3.841 au seuil 5%. Comme 3,117 < 3,841, on ne peut pas rejeter l'hypothèse d'indépendance.

Exemple 2

On interroge 1505 ménages choisis au hasard et on les classe suivant la catégorie socio-professionnelle du chef de famille et le nombre d'enfants :

Catégorie socio-professionnelle	0 ou 1enfant	2 ou 3 enfants	+ de 3 enfants
A	203	150	6
B	266	112	1
C	258	126	2
D	196	168	17

Tester l'indépendance des deux critères de classification.

<u>Corrigé</u> :

C'est un test de liaison entre deux variables qualitatives : X est la variable représentant la catégorie socio-professionnelle et Y celle représentant le nombre d'enfants. On appelle n_{ij} le nombre de ménages présentant les caractères x_i et y_j, $n_{i.}$ le nombre de ménages présentant le caractère x_i et $n_{.j}$ le nombre de ceux présentant le caractère y_j.

L'hypothèse H_0 d'indépendance se traduit par : $n\, n_{ij} = n_{i.}\, n_{.j}$.

$$D = n\left(\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right) \qquad D = 1505\left(\frac{203^2}{359\times 923} + \frac{266^2}{379\times 923} + \ldots + \frac{17^2}{26\times 381} - 1\right) = 53{,}67$$

La statistique d^2 suit un $\chi^2(p-1)(q-1)$. $\chi^2_{0,95}(6) = 12{,}592$. Les deux critères ne sont pas indépendants.

Exemple 3

Dans une étude préalable, avant la mise sur le marché, d'un médicament prévenant le rhume, on a obtenu les résultats suivants ; Ce médicament peut-il être considéré comme efficace ?

	Patients traités	Patients non traités
Un rhume	20	80
Pas de rhume	60	40

<u>Corrigé</u> :

Soit p_1 la proportion exacte de sujets atteints de rhume chez les patients traités et p_2 cette proportion chez les patients non traités. On formalise les hypothèses :

H_0 $p_1 = p_2$ H_1 $p_1 < p_2$.

Test du χ^2 : $D = 200\dfrac{(20\times 40 - 80\times 60)^2}{80\times 120\times 100\times 100} = 33{,}33$. Or, $\chi_{0,05}^2(1) = 3{,}84$ On rejette H_0.

Test gaussien : $f_1 = 0{,}25$, $f_2 = 0{,}66$, et $p^* = 0{,}5$. $u = \dfrac{f_1 - f_2}{\sqrt{p^*(1-p^*)(\frac{1}{n_1}+\frac{1}{n_2})}} = -5{,}77$

On rejette H_0. On peut remarquer que $(-5{,}77)^2 = 33{,}30\ldots$

Mesure du degré d'association à la suite d'un test d'indépendance

A la suite d'un test d'indépendance, lorsqu'on rejette l'hypothèse nulle, c'est-à-dire lorsqu'on conclut à une association entre les phénomènes, on peut alors utiliser le coefficient V de Cramer pour mesurer le degré de liaison entre les deux variables. En effet, dans le cas des variables qualitatives, le coefficient de corrélation linéaire n'est pas valable pour mesurer le degré de liaison. Il est utilisé uniquement dans le cas de variables continues. Lorsque les variables sont qualitatives, on peut utiliser le V de Cramer pour évaluer leur degré de liaison. Bien entendu il existe plusieurs autres indicateurs pour mesurer le degré de liaison entre variables qualitatives telle que le coefficient de contingence, le coefficient phi de Pearson ou le pourcentage de l'écart maximum. Mais ceux-ci ne seront pas présentés ici.

Le coefficient V de Cramer se définit comme suit :

$$V_{Cramer} = \sqrt{\frac{D}{N(r-1)}}$$

Où D est la statistique de khi-deux calculée, r est la valeur minimum entre k et l : $r = min\ (k, l)$ et N l'effectif total de l'échantillon.

Les règles de décision par rapport à V sont les suivantes :

- V ≤0,1 : relation très faible
- V ∈ [0,1; 0,3] : relation faible
- V ∈ [0,3; 0,5] : relation modérée
- V ∈ [0, 5; 0,7] : relation forte
- V > 0,7 : relation très forte

4.5.3. Test d'adéquation de Kolmogorov-Smirnov

Le test d'adéquation de Kolmogorov-Smirnov est un test non paramétrique d'ajustement à une distribution continue dont la fonction de répartition est F (x).

Soit F_n^* la fonction de répartition empirique d'un échantillon de taille n de X. La variable de décision est la variable aléatoire :

$$D_n = Sup_{x \in R} |F_n^*(x) - F(x)|$$

Soit K_n une fonction de répartition définie telle que :

$$K_n = \sqrt{n} D_n$$

Les travaux de Kolmogorov et Glivenko ont montré que lorsque $n \to +\infty$ la fonction K_n converge vers :

$$\begin{cases} k(y) = \sum_{k=-\infty}^{+\infty} (-1)^k\, e^{-2k^2y^2} \ si\ y > 0 \\ k(y) = 0 \qquad\qquad\qquad\qquad si\ y \leq 0 \end{cases}$$

En fonction de la valeur de D_n la règle de décision est la suivante.

On rejette l'hypothèse H0 lorsque la valeur de la statistique D_n est supérieure à la valeur d_n au seuil $1 - \alpha$ lue dans la table du test de Kolmogorov. Dans le cas contraire, on garde H0 et on considère que la distribution proposée est acceptable.

La distribution de D_n et de K_n ont été tabulées par Kolmogorov et Kolmogorov et Glivenko.

D'une manière générale pour D_n lorsque $n \to +\infty$, (en pratique $n \geq 40$), on peut directement calculer quelques valeurs de référence en fonction du seuil d'erreur α :

$$\alpha = 0{,}01\ ;\ d_{n,1-\alpha} = d_{n;\,0{,}99} = \frac{1{,}63}{\sqrt{n}}$$

$$\alpha = 0{,}05\ ;\ d_{n,1-\alpha} = d_{n;\,0{,}95} = \frac{1{,}36}{\sqrt{n}}$$

$$\alpha = 0{,}1\ ;\ d_{n,1-\alpha} = d_{n;\,0{,}90} = \frac{1{,}22}{\sqrt{n}}$$

Remarque :

Le test de Kolmogorov est préférable à celui du χ^2 pour des variables continues car la variable de décision utilise l'échantillon tel qu'il est fourni au départ sans regrouper les données en classes.

Exemples d'application

Exemple 1

On dispose de n matériels identiques et on note les durées de vie x_i en heures de chacun.

133	169	8	122	58

Si la durée de vie est supposée être une variable exponentielle, tester la conformité des données à cette loi.

Ici, on dispose d'un échantillon de taille n=5 car tous les x_i répètent qu'un seule fois.

On sait que la fonction de répartition d'une loi exponentielle est :

$$F(x) = P(X < x) = 1 - e^{-\lambda x}$$

Où λ représente le paramètre de la loi. Dans cet exercice la valeur du paramètre n'est pas fournit. On doit donc l'estimer à partir de l'échantillon (cela doit toujours être le cas lorsque le paramètre n'est pas connu !)

On sait que pour une loi exponentielle

$$\hat{\lambda} = {}^{1}/_{\bar{X}}$$

Où $\bar{X}$ est la moyenne des X sur l'échantillon. Ici on a :

$$\hat{\lambda} = {}^{1}/_{98}$$

On a donc

$$F(x) = P(X < x) = 1 - e^{-\frac{x}{98}}$$

En calculant les $F(x)$ en utilisant les valeurs fournie dans le tableau (après avoir ordonné les valeurs), on obtient le nouveau tableau suivant :

x_i	8	58	122	133	169
F(x_i)	0,079	0,447	0,711	0,743	0,821

Pour calculer la fonction de répartition empirique, on calcule d'abord les fréquences relatives (fonction de densité), on a le tableau suivant :

x_i	8	58	122	133	169
$f^*(x_i)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

En utilisant ces densités, on calcule la fonction de répartition empirique $F^*(x_i)$ avec $F^*(x_i) = P^*(X < x)$. Ainsi on obtient le tableau de répartition empirique :

x_i	8	58	122	133	169
$f^*(x_i)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
$F^*(x_i) = P^*(X < x)$	0	0,2	0,4	0,6	0,8

Dès lors on construit de comparaison suivant

x_i	8	58	122	133	169
F(x_i)	0,079	0,447	0,711	0,743	0,821
$F^*(x_i)$	0	0,2	0,4	0,6	0,8
$\lvert F^*(x_i) - F(x)\rvert$	0,079	0,247	**0,311**	0,143	0,021

On calcule la statistique D_n

$$D_5 = Sup_{x \in R} |F_5^*(x) - F(x)|$$

On a $D_n = \mathbf{0,311}$ correspondant à la valeur x_i =122.

En se fixant un seuil $\alpha = 0{,}05$, on trouve dans la table de Kolmogorov la valeur $d_n(1-\alpha) = d_5(0{,}95) = 0{,}563$

Etant donné que $D_5 < d_5(0{,}95)$, on ne rejette pas l'hypothèse nulle d'adéquation. C'est-à-dire que la durée de vie des ampoules suit une loi exponentielle au seuil de confiance de 95%.

Exemple 2

La directrice des ressources humaines d'une entreprise a pour mission d'étudier l'absentéisme des salariés de cette entreprise.

1) Elle effectue le relevé des absences durant une période de 200 jours :

Nombre de personnes absentes	0	1	2	3	4	5	>5
Nombre de jours	15	30	48	46	34	22	5

Elle conclue que le nombre des personnes absentes suit une loi de Poisson avec un taux moyen d'absentéisme de 3 personnes par jour. Est-ce vraisemblable au seuil de signification de 5%? (utiliser le test de Kolmogorov et le test de khi-deux pour comparer)

2) Cette directrice a aussi noté le nombre d'absences pour chaque jour de la semaine :

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi
Nombre de personnes absentes	126	96	90	98	130

Tester l'hypothèse selon laquelle le nombre de personnes absentes est uniformément distribué au cours des jours de la semaine. On effectuera aussi les tests du χ^2 et de Kolmogorov (α= 5%).

<u>Corrigé</u> :

1) On désigne par F* la fonction de répartition empirique et par F la fonction de répartition théorique.

En utilisant la table de la loi de Poisson de paramètre 3, on peut donc dresser le tableau suivant :

(NB : au lieu de travailler avec les probabilités, on peut travailler avec les fréquences en multipliant les probabilités par l'effectif total ; la fonction de répartition est toujours un cumul où la dernière classe correspond à l'effectif total ou à 1 s'il s'agit des probabilités. Faire attention alors pour savoir si c'est P(X<x) ou P(<=x)).

N	0	1	2	3	4	5	>5
f*	15	30	48	46	34	22	5
200 F*(n)	15	45	93	139	173	195	200
200 F(n)	9,96	39,82	84,64	129,44	163,06	183,22	200
f	9,96	29,86	44,82	44,8	33,62	20,16	16,78
$200\lvert F - F^* \rvert$	5,04	5,18	8,36	9,56	9,94	**12,78**	0

$D_n = Sup_n|F - F^*| = 200/12{,}78 = 0{,}064$. $D_n = \sqrt{200}d_n = 0{,}90$. Les tables de Kolmogorov pour D_n et la table de Kolmogorov et Glivenko pour K_n nous donnent, au risque α = 0,05, le seuil k = 1,358. L'hypothèse de la loi de Poisson de paramètre 3 est donc gardée.

Sous l'hypothèse H_0 : « Loi de Poisson de paramètre 3 » on fait le test du χ^2 :

$$d^2 = \sum_i \frac{(f - f^*)^2}{f}$$

Le calcul donne d^2 = 11,25. Sous H_0, d^2 suit un χ^2 à 6 degrés de liberté. Par les tables : $\chi^2_{0,95}(6) = 12{,}592$. Là encore on ne rejette pas l'hypothèse H_0 de loi de Poisson.

2) On suppose une hypothèse de répartition uniforme dans la semaine :

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi
f*	126	96	90	98	130
f	108	108	108	108	108
540F*(n)	126	222	312	410	540
540F(n)	108	216	324	432	540

- Test du χ^2 d^2 = 12,74. d^2 suit un χ^2 à 4 degrés de liberté. Le seuil est k = 9,488.

On rejette l'hypothèse d'équi-répartition des absences entre les jours.

4.5.4. Le test d'adéquation de Cramer et de Von-Mises

Soit F(x) une distribution théorique à laquelle on souhaite tester l'adéquation des données sur une variable X et F_n^* la fonction de répartition empirique X sur échantillon de taille n.

Le test se présente comme suit :

$$\begin{cases} H_0 : F_n^*(x) - F(x) \\ H_1 : F_n^*(x) \neq F(x) \end{cases}$$

On définit ainsi une variable aléatoire telle que :

$$V = nW_n^2 = \int_{-\infty}^{+\infty} |F_n^*(x) - F(x)|^2 \, dF(x)$$

La variable V c'est une variable aléatoire dont la loi, indépendante de F, sert de variable de test. Elle représente la mesure de l'écart entre une répartition théorique et une répartition empirique.

On démontre que :

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

Où x_i sont les valeurs ordonnées croissantes de l'échantillon.

Règle de décision :

On rejette l'hypothèse H0 dès que la valeur de la statistique V est supérieure à une valeur n'ayant qu'une probabilité a d'être dépassée.

Au seuil $\alpha = 0{,}05$, la valeur critique est 0,46136. On rejette alors H0 dès que V > 0,46136 .

Chapitre 5 : Techniques de sondage

5.1. Généralités sur le sondage

But

Ramené à sa dimension la plus élémentaire, le sondage vise à estimer le total, la moyenne, la proportion ou le ratio d'une variable d'étude y sur une population U finie de taille N. Par exemple, le nombre de votants pour un parti aux prochaines échéances électorales ; Etablir une prévision du chiffre d'affaire annuel dans un secteur d'activité à partir de l'historique des ventes des entreprises qui y opèrent.

Si l'on pouvait mesurer y sur chaque individu de la population, alors on ne ferait pas un sondage mais plutôt un recensement et il n'y aurait plus de problème d'estimation. Seulement il est souvent très coûteux, peu réaliste, voire impossible de mesurer la variable d'étude sur toute la population et on doit donc se contenter de l'observation de y sur un échantillon aléatoire.

En résumé, un sondage est un mécanisme probabiliste qui permet d'observer une variable y sur un échantillon s de la population U dont on veut estimer une caractéristique (par exemple la moyenne de y sur U). La méthode d'estimation de la caractéristique doit fournir :

– un estimateur de la caractéristique,

– la variance de cet estimateur,

– des estimations basées sur s de ces deux quantités.

Notons y_k la valeur de la variable d'étude y pour l'individu ou unité k de cette population. On note respectivement par t_{yU} et $\bar{y}_U$ le total et la moyenne de y, on a :

$$t_{yU} = \sum_{k=1}^{N} y_k$$

$$\bar{y}_U = \frac{1}{N}\sum_{k=1}^{N} y_k = \frac{1}{N} t_{yU}$$

Désormais, pour simplifier la notation, écrira

$$\sum_{k=1}^{N} y_k = \sum_{U} y_k$$

Ainsi, on a

$$\bar{y}_U = \frac{1}{N}\sum_U y_k$$

On appelle paramètre d'intérêt, la fonction des y_k, $k \in U$ qu'on veut estimer, par exemple t_{yU} (pour le total) ou $\bar{y}_U$ (pour la moyenne de la population).

On est souvent amené à estimer d'autres paramètres que le total d'une variable. Par exemple un revenu par tête R dans une région est un rapport de totaux ou ratio :

$$R = \frac{\sum_U y_k}{\sum_U z_k}$$

Où y_k et z_k désignent respectivement le revenu et la taille du ménage k de la population U des ménages de la région. Si la taille $\sum_U z_k$, de la population n'est pas connue, l'estimation de $\bar{y}_U$ revient à l'estimation d'un ratio.

Plan de sondage

Un plan de sondage sur une population U est un mécanisme probabiliste qui permet d'obtenir un échantillon aléatoire s, d'éléments de U. Un plan de sondage décrit la démarche de tirage des individus ou unités de U qui formeront l'échantillon aléatoire. L'aléatoire en sondage provient d'abord de la variabilité de l'échantillon tiré dans la population finie fixée. A chaque application d'un plan de sondage sur une population on devrait obtenir un échantillon différent.

La taille d'un plan de sondage est la taille des échantillons qu'il génère. Elle peut être constante, on parle alors de plan de taille fixe, ou bien aléatoire pour des plans que nous étudierons un peu plus tard dans le document.

L'échantillon étant obtenu suivant un plan de sondage défini, on peut, à partir de cet échantillon, obtenir les éléments suivants :

(1) une estimation du paramètre d'intérêt,

(2) une estimation de la variance de l'estimateur du paramètre d'intérêt.

C'est la démarche probabiliste qui permet d'obtenir une mesure de précision de l'estimation. L'aspect aléatoire est donc crucial. Un sondage qui se limite à fournir une estimation de total ou de moyenne, sans donner une estimation de l'écart-type de cette estimation reste incomplet.

Vocabulaire des sondages

Unité d'observation : Objet sur lequel on fait une mesure. C'est l'unité de base observée.

Population cible ou champ d'une enquête : Collection complète des unités d'observations qu'on veut étudier. Il faut la définir soigneusement pour chaque étude.

Population échantillonnée : La liste de toutes les unités d'observation qui pourraient être choisies pour former un échantillon. Elle ne coïncide pas toujours avec la population cible.

Échantillon : Un sous-ensemble de la population échantillonnée.

Unité d'échantillonnage : Les unités susceptibles d'être tirées.

Base de sondage : une liste des unités d'échantillonnage. Par exemple, un annuaire par nom, une carte où sont situées des exploitations agricoles, peuvent être des bases de sondage. Il arrive qu'on ait plusieurs bases de sondage pour un même problème. Il arrive aussi qu'on n'ait pas de base de sondage pour une population, cas par exemple d'une population d'animaux sauvages.

Défaut de couverture : Le fait qu'il existe des individus de la population cible qui ne sont pas dans la base de sondage.

Biais de sélection : le biais qui survient quand une partie de la population cible n'est pas dans la population échantillonnée. Par exemple, si on veut étudier les revenus des ménages d'une commune et qu'on oublie les travailleurs migrants, on va trouver des revenus plus élevés qu'ils ne le sont en vérité. Causes classiques de ce biais : Non-réponse, recours au volontariat pour obtenir des réponses ...

Biais de mesure : il survient quand l'instrument de mesure à tendance à donner une valeur qui s'écarte de la vraie mesure dans une direction particulière. Par exemple, dans des sondages sur la végétation, on découpe la surface en parcelles et on choisit un échantillon de parcelles. On compte le nombre de plantes dans chaque parcelle. Que faire des plantes en bordure de parcelle ? Si un observateur a tendance à les compter toutes, il fournira une estimation du nombre total de plantes supérieur à la réalité. Autre exemple : les gens peuvent ne pas dire la réalité (sous déclaration de revenus, d'âge), une question peut être mal comprise.

Notion d'information auxiliaire en sondage

Pour l'instant on a évoqué une population U pour laquelle on veut estimer le total $t_{yU} = \sum_U y_k$

Or, souvent on dispose d'une information auxiliaire sur U. Cette information se ramène le plus souvent à la connaissance d'une variable x pour chaque individu de U, liée à y. L'aspect important est que x est connue sans coût pour chaque individu ou du moins, à un moindre coût que y. On comprend qu'il est très important d'avoir des méthodes de sondage qui exploitent une telle information.

Par exemple, on doit estimer le nombre moyen de fois qu'une personne d'âge compris entre 15 ans et 30 ans, habitant dans une certaine région, va au cinéma chaque mois. Si on admet que les habitants de zones urbaines vont plus souvent au cinéma que les habitants de zones rurales. On peut tenir compte du lieu de résidence (rural/Urbain) dans le processus de tirage. Par exemple, on peut décider de faire un plan de sondage différenciant les deux types de zones. Ici x est le statut urbain/rural du lieu de résidence d'un individu de la population étudiée.

Dans cet exemple l'information auxiliaire permet d'améliorer l'estimation du total ou de la moyenne de la variable d'étude.

Erreurs non dues à l'échantillonnage

L'aléa dans le tirage introduit l'erreur d'échantillonnage. C'est une erreur attendue et qu'on sait quantifier si l'on a fait un échantillonnage probabiliste. Mais il peut exister d'autres erreurs dans un sondage : erreurs non dues à l'échantillonnage. Nous en donnons ici une courte description.

Erreurs de couverture : une erreur de couverture survient lorsqu'il y a une omission, une répétition ou un ajout erroné des unités dans la population ou l'échantillon. Les omissions sont appelées sous-dénombrement, tandis que les répétitions et les ajouts erronés sont appelés sur-dénombrement. Ces erreurs surviennent quand la base de sondage utilisée ne recouvre pas la population à étudier.

Erreurs de réponse : elles surviennent quand les réponses finalement enregistrées ne correspondent pas aux réponses réelles. Elles peuvent survenir à cause d'une mauvaise rédaction des questions, du comportement de l'interviewer. Par exemple un interviewer peut modifier la formulation d'une question en fonction de la personne en face. Ou encore un répondant peut vouloir donner une certaine réponse pour être agréable à l'enquêteur, politiquement correct...

Erreurs de non-réponse : Elles surviennent quand le répondant ne répond pas à suffisamment de questions de l'enquête. La non-réponse peut être partielle ou complète.

– **Erreurs de non-réponse complète** : Ces erreurs peuvent se produire lorsque l'enquête ne mesure pas certaines unités de l'échantillon sélectionné. Les causes de ce type d'erreur peuvent être : (1) que le répondant n'est pas disponible ou est temporairement absent, (2) qu'il est incapable de participer à l'enquête ou qu'il refuse.

Si un nombre important de personnes ne répondent pas à une enquête, alors les résultats peuvent être biaisés, étant donné que les caractéristiques des non-répondants peuvent différer des caractéristiques de ceux qui ont participé.

– Erreurs de non-réponse partielle. Ce type d'erreur se produit lorsque l'information obtenue du répondant est incomplète. Par exemple, certaines questions peuvent être difficiles à comprendre pour certaines personnes. Afin de réduire cette forme de biais, il faut porter une attention particulière à la conception et à la mise à l'essai du questionnaire. Il faut le tester longuement, le re-rédiger tant que des imprécisions, des malentendus sur le sens des questions, des incompréhensions de questions, demeurent.

Le problème de la non-réponse sera étudié plus loin.

Le(s) questionnaire(s)

Dans une enquête par sondage on ne peut espérer avoir de bonnes données sans un bon questionnaire bien administré. Un questionnaire bien conçu permet de recueillir des données en toute efficacité et sans grand risque d'erreur. Il facilite le codage et la saisie des données et permet généralement de réduire les frais et les délais de collecte et de traitement des données. La grande difficulté de l'élaboration d'un questionnaire est d'arriver à traduire les objectifs de la collecte de données en un cadre cohérent d'un point de vue conceptuel et méthodologique.

Avant de mobiliser de grands moyens dans la conception d'un questionnaire on devrait se poser les questions suivantes pour définir clairement les objectifs du projet :

– Faut-il faire une enquête ou bien définir un plan d'expérience ?

– Que veut-on apprendre ?

– Comment l'information sera-t-elle utilisée ? En particulier, quel traitement fera-t-on des réponses à chaque question ?

Une fois qu'on a des réponses claires à ces questions, on peut envisager la conception du questionnaire, sa réalisation, son administration et l'analyse de ses résultats.

5.2. Notations et rappels sur les paramètres d'intérêt

Nous reprenons et complétons les notations qui seront utilisées dans le reste du document.

Soit y la variable d'intérêt relevée sur l'unité d'observation k, les N observations constituent la population U. On adopte les notations suivantes pour la suite de la présentation :

- **Le total** sur la population : t_{yU} avec

$$t_{yU} = \sum_{k=1}^{N} y_k = \sum_{U} y_k$$

Soit A un sous ensemble tel que $A \subset U$ alors, on notera

$$\sum_{A} y_k \equiv \sum_{k \in A} y_k$$

- **La moyenne** sur la population : $\bar{y}_U$ avec

$$\bar{y}_U = \frac{1}{N}\sum_{k=1}^{N} y_k = \frac{1}{N} t_{yU} = \frac{1}{N}\sum_{U} y_k$$

- **La variance** sur la population

$$S_{yU}^2 = \frac{1}{N-1}\sum_{k=1}^{N} (y_k - \bar{y}_U)^2 = \frac{1}{N-1}\sum_{k \in U}^{N} (y_k - \bar{y}_U)^2$$

- **La covariance** de deux variables y et z sur U,

$$S_{yzU} = \frac{1}{N-1}\sum_{k \in U}^{N} (y_k - \bar{y}_U)(z_k - \bar{z}_U)$$

La somme sur tous les couples $(k; l)$; k; k; $l \in U$ d'une certaine quantité $g_{k,l}$définie sur chacun de ces couples sera notée :

$$\sum_{k=1}^{N}\sum_{l=1}^{N} g_{k,l} = \sum^{N}\sum_{U}^{N} g_{k,l}$$

5.3. Les plans de tirages simples à un degré

5.3.1. Généralités

Dans ce chapitre nous étudions plusieurs types de plans de tirages aléatoires simples. Il s'agit notamment du plan simple à probabilités égales sans remise, du plan simple à probabilités égales avec remise ; les plans simples à probabilités inégales sans remise, les plans simples à probabilités inégales avec remise et le plan systématique. Nous présentons également le plan de tirage stratifié.

Pour chaque plan de tirage étudié, nous présentons les démarches de calcul des paramètres d'intérêt du sondage : la moyenne, le total, la proportion et le ratio ainsi que leurs caractériels (espérance et variances).

5.3.2. Définitions des plans simples

On appelle plan simple, un plan de sondage dans lequel on accède aux individus qui formeront l'échantillon par une seule opération aléatoire. Dans le cas contraire on parle de plan complexe. Par exemple, supposons qu'on s'intéresse à tous les enfants scolarisés dans les écoles primaires d'une région. Il est clair qu'on ne peut accéder à ces enfants que par l'intermédiaire de l'école qu'ils fréquentent. Un plan de sondage sur ces enfants comportera au moins une étape de sélection d'écoles, puis peut-être de sélection de classes dans l'école et enfin d'enfants dans la classe. C'est un plan complexe, précisément un plan à plusieurs degrés étudié plus loin.

5.4. Le plan de tirage simple à probabilités égales

Nous commençons par étudier le plan de tirage simple à probabilités égales sans remise que nous noterons par ici par plan SI (Ce plan est aussi souvent noté PESR).

On se fixe comme paramètre une taille n d'échantillon et on tire n individus, sans ordre et sans remise dans la population des N individus. Dans le plan **SI** il y a C_n^N échantillons possibles et équiprobables. Ceci est une application directe des techniques de dénombrement. On notera ce plan $SI(N,n)$.

La loi de probabilité sur les échantillons possibles se présente comme suit :

$$\begin{cases} p(s) = \dfrac{1}{C_n^N} \ si \ \ cards(s) = n \\ p(s) = 0 \ \ si \ \ cards(s) \neq n \end{cases}$$

5.4.1. Le taux de sondage

Le taux de sondage se définit par l'expression suivante :

$$f = \frac{n}{N}$$

5.4.2. La probabilité d'inclusion

La probabilité d'inclusion d'un individu k, c'est la probabilité que se réalise un échantillon qui contient k.

Un certain individu k étant choisi, pour compléter un échantillon à n, il y a C_{n-1}^{N-1} possibilités, ou encore il y a C_{n-1}^{N-1} échantillons qui contiennent un individu fixé.

La probabilité d'inclusion de k dans un échantillon est la somme des probabilités de tous les échantillons qui contiennent k :

$$\pi_k = \sum_{S \ni k}^{N} p(s)$$

Dans le cas du plan SI(N,n), la probabilité d'inclusion se présente comme suit :

$$\pi_k = \frac{C_{n-1}^{N-1}}{C_n^N} = \frac{n}{N}$$

La probabilité d'inclusion se ramène simplement au taux de sondage dans le cadre du plan SI.

Noter que k est donné et que c'est s qui varie, π_k est appelée une probabilité d'inclusion du premier ordre. Pour le plan SI elle ne dépend pas de k. On définit de même la probabilité d'inclusion du deuxième ordre de deux éléments k et l, $k \neq l$:

$$\pi_{k,l} = \sum_{S \ni k \,\&\, l}^{N} p(s) = \frac{Nombre\ d'échantillons\ contenant\ k\ et\ l}{Nombre\ total\ d'échantillons\ possibles}$$

Dans le cas du plan SI, on a :

$$\pi_{k,l} = \frac{C_{n-2}^{N-2}}{C_n^N} = \frac{n(n-1)}{N(N-1)}$$

5.4.3. Indicatrices d'inclusion

L'objectif maintenant est d'avoir une méthode plus simple que l'utilisation de la loi de probabilité des échantillons pour calculer les caractéristiques de certains estimateurs en sondage.

On a défini et calculé les probabilités d'inclusion d'ordre 1 et 2, les π_k et $\pi_{k,l}$ pour le plan SI. Associons à la probabilité d'inclusion d'ordre1, l'indicatrice d'inclusion de k dans l'échantillon s :

$$1_k\ (s) = \begin{cases} 1\ si\ l'échantillon\ qui\ se\ réalise\ contient\ k \\ \\ 0\ sinon \end{cases}$$

Pour le plan SI(N,n) , on a :

$$E\big(1_k\ (s)\big) = P(1_k\ (S) = 1) = \frac{n}{N}$$

où l'espérance mathématique est à comprendre au sens du plan de sondage.

Notons par $\Delta_{k,l}$ la covariance entre deux observations k et l, on a :

$$\Delta_{k,l} = \mathrm{cov}\big(1_k\ (S), 1_l\ (S)\big) = \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N} = -\frac{f(1-f)}{N-1} \qquad \forall\ k \neq l$$

Et lorsque $k = l$, on obtient la variance de $1_k\ (S)$ telle que :

$$\Delta_{k,l} = \mathrm{cov}\big(1_k\ (S), 1_k\ (S)\big) = \mathrm{var}\big(1_k\ (S)\big) = \frac{n}{N}\left(1 - \frac{n}{N}\right) = f(1-f)$$

Comme on peut le remarquer la covariance $\Delta_{k,l}\ \forall\ k \neq l$ est négative car le plan étant de taille fixe, si on sait que k ∈ s, les chances d'avoir l ∈ s diminuent.

5.4.4. Estimation dans le cadre d'un plan SI

Dans un sondage, le but ultime est l'estimation des quantités d'intérêts. Dès, après la sélection de l'échantillon, il faut alors proposer les estimateurs de ces paramètres d'intérêt. D'une manière générale, dans un sondage, on cherche à estimer la moyenne, le total, le ratio ou la proportion.

5.4.4.1. Estimation de la moyenne dans le plan SI

On a tiré dans U un échantillon s d'individus suivant le plan SI(N,n). On veut construire un estimateur de $\bar{y}_U$ à l'aide des y_k ; $k \in S$.

Considérons $\bar{y}_s$, la moyenne sur l'échantillon obtenu. L'espérance mathématique de cette moyenne par rapport au plan de sondage est :

$$E(\bar{y}_S) = \sum_{s\in S}^{N} p(s)\bar{y}_s = \frac{1}{C_n^N}\sum_{s\in S}^{N}\frac{1}{n}\sum_{k=1\,|s}^{n} y_k$$

où S désigne l'élément aléatoire dont s est une réalisation et s parcourt l'ensemble S des C_n^N échantillons possibles. On a vu que chaque y_k apparaît dans C_{n-1}^{N-1} termes. L'espérance mathématique de la variable aléatoire $\bar{y}_S$ est donc

$$E(\bar{y}_S) = \sum_{s\in S}^{N} p(s)\bar{y}_s = \left(\frac{C_{n-1}^{N-1}}{C_n^N}\right)\frac{1}{n}\sum_{U} y_k = \bar{y}_U$$

Ainsi, dans le plan SI, la moyenne sur l'échantillon est un estimateur sans biais de la moyenne sur la population.

On peut montrer également que :

$$var(\bar{y}_S) = \left(1-\frac{n}{N}\right)\frac{S_{yU}^2}{n} = (1-f)\frac{S_{yU}^2}{n}$$

On montrera également que : $S_{ys}^2 = \frac{1}{n-1}\sum_s (y_k - \bar{y}_s)^2$ est un estimateur sans biais de $S_{yU}^2 = \frac{1}{N-1}\sum_U (y_k - \bar{y}_U)^2$.

Finalement un estimateur sans biais de la variance de l'estimateur $\bar{y}_s$ de la moyenne $\bar{y}_U$ est :

$$\widehat{var}(\bar{y}_S) = \left(1-\frac{n}{N}\right)\frac{S_{ys}^2}{n} = (1-f)\frac{S_{ys}^2}{n}$$

Indicatrices d'inclusion et calcul de l'espérance et de la variance de $\bar{y}_s$

D'abord on peut observer que :

$$\bar{y}_s = \frac{1}{n}\sum_s y_k = \frac{1}{n}\sum_U 1_k\,(s)$$

Sachant que $E\left(1_k\ (s)\right) = P(1_k\ (S) = 1) = \frac{n}{N}$, on peut écrire alors :

$$E(\bar{y}_S) = \frac{1}{n}\sum_U E(1_k\ (s)) = \frac{1}{n}\sum_U y_k \frac{n}{N} = \bar{y}_U \rightarrow sans\ biais$$

Notons que dans ce calcul d'espérance, grâce aux indicatrices d'inclusion on a remplacé l'écriture d'un nombre aléatoire de termes $1_k\ (s)$ par celle d'un nombre certain de termes $\frac{n}{N}$, qui ne pose pas de problème.

Pour la variance, on a :

$$\mathrm{var}(\bar{y}_S) = \frac{1}{n^2}\mathrm{var}\left(\sum_U y_k\ 1_k\ (s)\right) = \frac{1}{n^2}\sum\sum_U y_k y_l \Delta_{k,l}$$

NB : pour une variable Z,

$$\sum\sum_U z_{k,l} = \sum_{k\in U}\sum_{l\in U} z_{k,l}$$

Ainsi, on a :

$$\sum_U\sum_U y_k y_l \Delta_{k,l} = -\frac{f(1-f)}{N-1}\sum\sum_{U,k\neq l} y_k y_l + f(1-f)\sum_U (y_k)^2$$

$$= -\frac{f(1-f)}{N-1}\sum\sum_{U,k\neq l} y_k y_l + \frac{f(1-f)}{N-1}\sum (y_k)^2 + f(1-f)\sum_U (y_k)^2$$

On a utilisé le fait que dans le plan SI, $\Delta_{k,l}$ ne prend que deux valeurs selon que $k \neq l$ ou $k = l$. Une identité élémentaire de la statistique descriptive nous donne :

$$\sum_U {y_k}^2 = \sum_U (y_k - \bar{y}_U)^2 + \frac{{t_{yU}}^2}{N}$$

(Penser à la formule développée de la variance !)

Ainsi, en remplaçant $\sum_U {y_k}^2$ par sa valeur dans l'expression de $\sum_U \sum_U y_k y_l \Delta_{k,l}$, on obtient :

$$\mathrm{var}(\bar{y}_S) = \left(1 - \frac{n}{N}\right)\frac{S_{yU}^2}{n} = (1-f)\frac{S_{yU}^2}{n}$$

dont la valeur estimée sur l'échantillon est

$$\widehat{var}(\bar{y}_S) = \left(1 - \frac{n}{N}\right)\frac{S_{ys}^2}{n} = (1-f)\frac{S_{ys}^2}{n}$$

5.4.4.2. Estimation du total dans le plan SI

On sait que :

$$t_{yU} = N\bar{y}_U$$

Mais $\bar{y}_U$ n'étant pas connu, on utilise son estimateur (à partir de l'échantillon) c'est-à-dire $\bar{y}_s$ qui est un estimateur sans biais de $\bar{y}_U$. Ainsi, on a :

$$\hat{t}_{yU} = N\bar{y}_s = N\left(\frac{1}{n}\sum_s y_k\right)$$

$$\hat{t}_{yU} = \left(\frac{N}{n}\sum_s y_k\right)$$

$\hat{t}_{yU}$ est un estimateur non biaisé du total t_{yU} car :

$$E(\hat{t}_{yU}) = E\left(\sum_s y_k\right) = \frac{N}{n}\sum_s E(y_k)$$

$$= \frac{N}{n}\sum_s \bar{y}_U = \frac{Nn\bar{y}_U}{n} = N\bar{y}_U = t_{yU}$$

$$E(\hat{t}_{yU}) = t_{yU} \rightarrow sans\ biais$$

On peut considérer que chaque élément de l'échantillon représente N/n éléments de la population, ou encore y_k est dilatée par le facteur $\frac{N}{n}$ pour construire l'estimateur du total. La variance de $\hat{t}_{yU}$ est :

$$\widehat{var}(\hat{t}_{yU}) = N^2 var(\bar{y}_S) = N^2(1-f)\frac{S_{yU}^2}{n}$$

Elle est estimée sans biais par :

$$\widehat{var}(\hat{t}_{yU}) = N^2\widehat{var}(\bar{y}_S) = N^2(1-f)\frac{S_{ys}^2}{n}$$

5.4.4.3. Estimation d'un total et d'une moyenne sur une sous-population (domaine)

Exemple. On fait un sondage auprès des ménages d'une région pour savoir combien d'heures en moyenne ils consacrent par mois à s'occuper d'une personne âgée dépendante. Il est clair que cette moyenne ne concerne que les ménages hébergeant une personne dépendante. Comme on ne dispose pas de la liste de tels ménages, on va tirer un échantillon de ménages auxquels on demandera s'ils hébergent une personne âgée dépendante et combien de temps ils y consacrent. On tire un échantillon dans une population qui contient la population qui nous intéresse.

Situation : On tire s, échantillon sur U suivant un plan SI(N; n), mais on est intéressé par le total ou la moyenne de la variable d'étude sur U_d, sous-population de U, de taille N_d. On note $s_d = s \cap U_d$

On estime la moyenne $\bar{y}_{U_d}$ par son équivalent empirique $\widehat{\bar{y}_{U_d}}$ telle que :

$$\widehat{\bar{y}_{U_d}} = \frac{1}{n_d} \sum_{s_d} y_k = \bar{y}_{s_d}$$

Exercice d'application

On considère une population de N = 5 individus, pour lesquels on connaît les valeurs de la variable y : $y_1 = 3$, $y_2 = 1$, $y_3 = 0$, $y_4 = 1$, $y_5 = 5$. On choisit un plan SI avec une taille d'échantillon n = 3.

1. Donner les valeurs de la moyenne, de la médiane et de la variance de la variable y dans la population.

Lister tous les échantillons possibles de taille n = 3. Quelle est la probabilité de sélection de chaque échantillon ?

2. Pour un échantillon donné, on estime la moyenne (respectivement la médiane) de la population.

Calculer les valeurs de ces estimateurs pour chaque échantillon et en déduire que l'estimateur de la moyenne est sans biais alors que l'estimateur de la médiane est biaisé.

3. Pour chaque échantillon, calculer l'estimateur S_{ys}^2 de S_{yU}^2 et en déduire que cet estimateur est sans biais.

Indications :

1. Il y a 10 échantillons possibles de taille 3 et puisque le plan est un plan SI, ces échantillons sont équiprobables.

2. Calculer les moyennes arithmétiques des estimateurs de la moyenne d'une part et de la médiane d'autre part. Comparer avec les vraies valeurs calculées à la question précédente.

3. Calculer les S_{ys}^2 (un par échantillon), en faire la moyenne arithmétique et comparer à la vraie valeur S_{yU}^2.

5.4.4.4. Estimation d'une proportion dans un plan SI

On peut dénombrer de nombreux cas où il s'agit d'estimer une proportion plutôt qu'une moyenne ou un total après le tirage de l'échantillon.

Exemples

1. Estimer la proportion de familles hébergeant une personne âgée dépendante dans une certaine ville.

2. Estimer la proportion de clients d'une banque susceptible d'acheter un nouveau produit de la banque.

Soit une variable binaire (indicatrice) y, avec $y_k = 1$ si l'individu k (famille ou banque dans les exemples) a la caractéristique recherchée, $y_k =$ si l'individu k n'a pas la caractéristique. Le nombre total d'individus ayant la caractéristique dans U est évidemment $t_{yU} = \sum_U y_k$: et la proportion d'individus ayant la caractéristique dans la population est $p = \frac{t_{yU}}{N} = \bar{y}_U$. Une proportion est donc la moyenne d'une variable indicatrice et les résultats obtenus pour une moyenne s'appliquent immédiatement. On les rassemble maintenant dans le cas d'un plan SI.

On veut estimer

$$p = \frac{t_{yU}}{N} = \bar{y}_U$$

Comme la variable y est une indicatrice, on a :

$$y_k^2 = y_k \Rightarrow \sum_U y_k^2 = \sum_U y_k = Np$$

Donc

$$S_{yU}^2 = \frac{1}{N-1}\sum_U (y_k - \bar{y}_U)^2 = \frac{1}{N-1}(Np - Np^2) = \frac{N}{N-1}p(1-p)$$

Et lorsque N tend vers l'infini, $\frac{N}{N-1}$ tend vers 1 alors

$$S_{yU}^2 \approx p(1-p)$$

Ainsi, pour une variable binaire (indicatrice), la moyenne et la variance de la population s'expriment comme suit :

$$\bar{y}_U = p$$

$$S_{yU}^2 = p(1-p)$$

Soit s un échantillon sur U, obtenu par un plan SI de taille n. L'estimateur de la proportion p par les valeurs dilatées est :

$$\hat{p}_s = \frac{1}{n}\sum_s y_k$$

Sa variance est

$$Var(\hat{p}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}^2 = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{N}{N-1}p(1-p)$$

Et si $\frac{N}{N-1} \approx 1$, alors un estimateur approximativement sans biais de cette variance est :

$$\widehat{Var}(\hat{p}_s) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{n}{n-1}\hat{p}_s(1-\hat{p}_s) = \frac{1-f}{n-1}\hat{p}_s(1-\hat{p}_s)$$

Si on peut négliger la correction de population finie, on aura finalement :

$$\widehat{Var}(\hat{p}_s) \approx \frac{1}{n-1}\hat{p}_s(1-\hat{p}_s)$$

5.4.4.5. Estimation d'un ratio dans le cadre du plan SI

Considérons d'abord deux exemples.

– Exemple 1. Supposons une population U de ménages, $y_k$le revenu du ménage k et z_k le nombre de personnes composant le ménage. Le revenu moyen par tête dans cette population est :

$$R = \frac{\sum_U y_k}{\sum_U z_k} = \frac{\bar{y}_U}{\bar{z}_U}$$

R est ce qu'on appelle un ratio, c'est-à-dire le rapport sur une même population de deux quantités non nécessairement liées (contrairement à la proportion où le numérateur est un sous ensemble du dénominateur).

– Exemple 2. La proportion d'électeurs qui, dans une élection présidentielle, choisissent un candidat particulier est le rapport : Nombre de votants qui choisissent le candidat / Nombre de suffrages exprimés.

Cette proportion doit être estimée comme un ratio car la taille de la population, c'est-à-dire le nombre d'électeurs qui votent n'est pas connue.

Estimation

En tirant dans une population U de taille N un échantillon s suivant un plan SI(N; n) , on estime le ratio, R, par le quotient des estimateurs des moyennes comme suit :

$$\hat{R} = \frac{\bar{y}_s}{\bar{z}_s}$$

Où $\bar{y}_s$ et $\bar{z}_s$ représentent les moyennes obtenues sur l'échantillon tiré.

On admettra provisoirement que la variance de cet estimateur s'estime par :

$$\widehat{var}(\hat{R}) = \frac{1}{\bar{z}^2{}_s}(1-f)S^2_{y-\hat{R}z,s}$$

5.4.4.6. Estimation par ratio

Attention : l'estimation d'un ratio n'est pas la même chose que l'estimation par ratio. Dans la première, il s'agit d'estimer un ratio. Dans la seconde, il s'agit d'utiliser un ratio pour estimer un total ou une moyenne sur une variable en utilisant une information auxiliaire sur une autre variable.

Exemples

Décrivons brièvement deux exemples d'estimation d'un total par ratio.

Exemple 1

Au début du 19e siècle, il n'existe pas en France de recensement, mais un registre des naissances est tenu dans chaque commune. Partant de cette situation, pour estimer la population de la France, Laplace considère un échantillon de communes, fait le recensement de leur population et mesure le rapport R = population totale de ces communes/ nombre de naissances de ces communes. Considérant que ce rapport doit être à peu près stable sur les communes il en déduit une estimation de la population totale :

R × nombre total de naissances en France

Le nombre de naissances est une information auxiliaire : elle est connue pour toutes les communes de France et elle est corrélée avec la population.

Exemple 2

Un chalutier doit estimer le poids des poissons de taille supérieure à une certaine longueur dans un chalut pour décider s'il décharge le chalut à bord ou s'il le rejette à la mer. Pour faire cette estimation, on peut évidemment mesurer le poids de tels poissons dans un échantillon. La taille de la population des poissons dans le chalut n'est pas connue. Mais il est facile de peser le chalut et l'échantillon.

Notons U la population des poissons dans le chalut, x_k le poids du poisson k et y_k tel que y_k= le poids du poisson k s'il est de taille supérieure à 25cm, y_k = 0 sinon. Comme il est facile de peser le chalut, on peut également estimer $= \frac{\sum_U y_k}{\sum_U z_k}$. Enfin, on peut faire l'hypothèse que, sur un échantillon s tiré dans le chalut suivant un plan SI, on doit avoir $\frac{\sum_s y_k}{\sum_s x_k} \approx R$. L'estimation de ce rapport multiplié par $\sum_u x_k$ fournit une estimation de $\sum_u y_k$.

Propriétés de l'estimateur par ratio dans le plan SI

Nous donnons maintenant les propriétés de l'estimateur par ratio quand l'échantillon est obtenu par plan SI.

Par un plan SI on tire un échantillon s de taille n dans une population U de taille N. On observe y_k et x_k; k∈ s et on connaît x_k ∀ k∈ U. On doit estimer $t_{yU} \equiv t_y$. Ainsi, on peut écrire :

$$t_{yU} = t_{xU}\frac{t_{yU}}{t_{xU}} = t_{xU}R$$

L'estimateur par ratio de t_{yU} est donc:

$$\hat{t}_{yU} = t_{xU}\hat{R}$$

Nous admettrons provisoirement qu'une estimation de la variance de $\hat{t}_{yU}$ est donnée par :

$$\widehat{var}(\hat{t}_{yU}) = \frac{\bar{x}^2{}_U}{\bar{x}^2{}_s}N^2\left(\frac{1}{n}-\frac{1}{N}\right)\left(S_{ys}^2 - 2\hat{R}S_{yx,s} + \hat{R}^2S_{xs}^2\right) = \frac{\bar{x}^2{}_U}{\bar{x}^2{}_s}N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_{y-\hat{R}x,s}^2$$

Il est important d'observer qu'au facteur $\frac{\bar{x}^2{}_U}{\bar{x}^2{}_s}$ près, souvent proche de 1, l'approximation de la variance ci-dessus est la variance du total des résidus $y_k - \hat{R}x_k$

NB : En plus de l'estimateur par ratio, il existe d'autres méthodes d'estimation telles que l'estimation par différence ou l'estimation par régression. Ces méthodes ne sont pas présentées ici.

5.5. Plan simple à probabilités égales avec remise

Dans le plan PEAR, l'échantillon s est obtenu par m tirages simples avec remise dans U de taille N. Un même individu peut donc apparaître plusieurs fois dans les m valeurs. Soit y_{k_i}la valeur obtenue au tirage i c'est à dire la i-ème observation du tirage et Y_i la variable aléatoire associée. La v.a. Y_i peut prendre n'importe laquelle des valeurs y_1;…; y_N avec la même probabilité. La loi de probabilité de Y_i est donc :

$$Pr(Y_i = y_k) = \frac{1}{N}, k = 1, \dots, N$$

5.5.1. Probabilité d'inclusion de premier ordre

A chaque tirage, un élément particulier k de la population est tiré avec la probabilité $\frac{1}{N}$. Ainsi, la probabilité d'inclusion dans l'échantillon est :

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m$$

Remarquons que $\left(1 - \frac{1}{N}\right)^m$ représente la probabilité que l'élément k apparaît 0 fois parmi les m objets tirés. Dès lors π_k représente la probabilité qu'il soit tiré au moins une fois. On a $\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m \approx \frac{m}{N}$ (si $\frac{1}{N}$ est faible i.e l'échantillon est de petite taille par rapport à la population finie ; propriétés de développement limité).

5.5.2. Estimateur de la moyenne

L'espérance et la variance de Y_i s'expriment comme suit :

$$E(Y_i) = \frac{1}{N}\sum_U y_k = \bar{y}_U$$

$$var(Y_i) = \frac{1}{N}\sum_s (y_k - \bar{y}_U)^2 = \frac{N-1}{N} S_{yU}^2 \cong S_{yU}^2 \quad si\ N \to +\infty$$

On obtient alors pour l'estimateur de la moyenne de la population, on a:

$$\bar{y}_S = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

Son espérance est :

$$E(\bar{y}_S) = E\left(\frac{1}{m}\sum_{i=1}^{m} Y_i\right) = \bar{y}_U$$

Et les tirages étant indépendants, on a :

$$var(\bar{y}_S) = var\left(\frac{1}{m}\sum_{i=1}^{m} Y_i\right) = \frac{1}{m}S_{yU}^2$$

5.5.3. Comparaison de la variance avec celle du plan SI

Comparons les variances des estimateurs de la moyenne dans les deux plans avec et sans remise.

Notion de correction de population finie

Nous avons montré que la variance de la moyenne dans un plan SI est

$$\widehat{var}(\bar{y}_S) = \left(1 - \frac{n}{N}\right)\frac{S_{ys}^2}{n} = (1-f)\frac{S_{ys}^2}{n}$$

Et celle dans un plan PEAR est :

$$var(\bar{y}_S) = \frac{1}{m}S_{yU}^2$$

On constate que l'expression de la variance dans le plan SI incorpore un facteur $(1-f) = 1 - \frac{n}{N}$. On l'appelle la correction de population finie. Ce facteur tend vers 1 lorsque n tend vers N.

Si le taux de sondage f = n/N est faible, c'est-à-dire si l'échantillon est de petite taille par rapport à la population finie, on a :

$$\text{var}_{SI}\left(\hat{t}_{yU\pi}\right) \approx N^2\frac{S_{yU}^2}{n}$$

Cette expression n'incorpore plus de correction de population finie à la différence de la première expression de la variance du total (présenté un peu plus haut) notamment qui est que :

$$\widehat{var}\left(\hat{t}_{yU}\right) = N^2\widehat{var}(\bar{y}_S) = N^2(1-f)\frac{S_{ys}^2}{n}$$

5.6. Plan simples à probabilités inégales sans remise

Nous voulons définir un estimateur du total t_{yU} à partir d'une population U où les probabilités π_k d'inclusion dans l'échantillon ne sont pas nécessairement égales. Nous commençons d'abord par exprimer les probabilités d'inclusion dans l'échantillon et nous calculons leurs moments d'ordre 2. Ensuite, nous présentons l'estimateur de t_{yU}. La démarche générale d'estimation reste très proche de celle pour le plan SI.

5.6.1. Indicatrices d'inclusion

Soit $p()$ un plan de tirage à probabilité inégales, notons S la variable aléatoire (v.a.) associée à un échantillon observé s. Soit k un élément de U. Nous rappelons et précisons les notions de probabilités d'inclusion et d'indicatrice d'inclusion.

5.6.2. Probabilités d'inclusion du premier ordre

La probabilité d'inclusion de k dans un échantillon quelconque est la probabilité que se réalise un échantillon qui contient k. Elle est exprimée par:

$$\pi_k = \sum_{s \ni k} p(s)$$

Où π_k est la probabilité d'inclusion de l'observation k et $p(s)$ la probabilité de tirage de l'échantillon s.

5.6.3. Probabilité d'inclusion du deuxième ordre

La probabilité d'inclusion du deuxième ordre des éléments k et l est :

$$\pi_{k,l} = P(1_k\ (S) = 1\ et\ \ 1_l\ (S) = 1\ ) = \sum_{s \ni k\&l} p(s)$$

Par convention $\pi_{k,k}$=π_k

Espérance et variance des indicatrices d'inclusion

On a

$$E(1_k\ (S)\ ) = \pi_k$$

On note par $\Delta_{k,l}$ la covariance entre $1_k\ (S)$ et $1_l\ (S)$ on a :

$$\Delta_{k,l} = cov\big(1_k\ (S), 1_l\ (S)\big) = \pi_{k,l} - \pi_k\pi_l$$

Avec

$$\Delta_{k,k} = var\left(1_k\ (S)\right) = \pi_k(1-\pi_k)$$

On constate alors que ces expressions sont plus générales et incluent les tirages à probabilité égales comme un cas particulier.

Pour rappel dans le plan SI, on avait obtenu :

$$var\left(1_k\ (S)\right) = \pi_k(1-\pi_k) = f(1-f)$$

Avec $f = \frac{n}{N}$

$$cov\left(1_k\ (S), 1_l\ (S)\right) = -\frac{f(1-f)}{N-1}$$

5.6.4. Exemple de plan à probabilités inégales sans remise : le Plan Bernoulli (plan BE)

Pour tirer un échantillon suivant le plan Bernoulli dans une population de taille N, on se fixe comme paramètre un nombre π_0 tel que $0 < \pi_0 < 1$. Ensuite, on utilise ce paramètre en tirant N nombres aléatoires $v_k \quad k = 1, \dots, N$; N indépendants suivant une loi uniforme U(0,1). Si $v_k < \pi_0$ alors on inclut l'élément k dans l'échantillon. Et si $v_k > \pi_0$ l'élément k n'est pas inclu dans l'échantillon. On voit alors que le plan BE donne des échantillons s de taille aléatoire n_s. On ne sait donc pas à l'avance quelle sera la taille de l'échantillon tiré. Elle varie de 0 à N. On notera ce plan : BE(π_0).

NB : π_0 peut être arbitrairement choisi ou tiré aléatoirement.

On vérifie sans difficulté que la loi de n_s dans ce plan est binomiale de paramètres N et n_s.

Admettons par exemple que la taille n de l'échantillon soit connue ; ainsi l'échantillon doit être choisi parmi tous ceux de taille n. Tous les individus ont les mêmes chances π_0 d'être sélectionnés. On peut montrer alors que conditionnellement à la taille, la loi de probabilité des échantillons est celle du plan SI.

Probabilités d'inclusion dans le plan BE(π_0)

Par définition, la probabilité d'inclusion d'un élément k du plan BE est :

$$\pi_k = \pi_0$$

Et comme l'appartenance d'un élément à l'échantillon ne dépend pas de l'appartenance des autres alors on a $\pi_{k,l} = {\pi_0}^2$ si $k \neq l$.

Remarque générale sur les plans simples à probabilités inégales sans remise

On peut se demander pourquoi on fait des plans de sondage à probabilités inégales. En réalité, de tels plans n'ont aucun intérêt quand on ne connaît rien de la population. Mais on dispose très souvent d'information auxiliaire sur la population. Cette information peut servir à organiser le tirage en étapes (plans de sondage complexes - à plusieurs degrés, en plusieurs phases...), à choisir avec une plus grande probabilité certains individus. Les plans de sondage complexes sont constitués de plans élémentaires qui sont très souvent des plans SI. On verra ces questions un peu plus tard.

5.7. Estimation par les valeurs pondérées : méthode de Horvitz-Thompson

5.7.1. Estimation du total par les valeurs pondérées

Soit un plan de sondage quelconque dont les probabilités d'inclusion de premier et de second ordre sont des π_k et $\pi_{k,l}$a donné un échantillon s et on a observé y_k. L'échantillon étant « représentative » de la population, dès lors chaque élément de s doit représenter plusieurs éléments de U. On estimera donc le total de t_{yU} par une somme des valeurs observées pondérées (ou encore somme des valeurs dilatées). Ainsi, chaque élément y_k sélectionné est multiplié par un poids ω_k avec $\omega_k \geq 1$.

La valeur de ces poids dépend du plan de sondage ou des probabilités d'inclusion correspondantes. Les poids peuvent dépendre également de caractéristiques de la population. Cela veut donc dire que les pondérations ne sont pas nécessaire l'inverse des probabilités d'inclusions.

L'estimateur du total par les valeurs pondérées est un estimateur linéaire qui présente comme suit :

$$\hat{t}_{yU} = \sum_{s} \omega_k y_k$$

Pour que cet estimateur soit sans biais, il faut que la condition suivante soit vérifiée :

$$\omega_k = \frac{1}{\pi_k}$$

Où π_k est la probabilité d'inclusion.

En effet, calculons l'espérance de $\hat{t}_{yU}$. On a

$$E(\hat{t}_{yU}) = E\left(\sum_{s} \omega_k y_k\right) = E\left(\sum_{U} \omega_k y_k 1_k(s)\right) = \sum_{U} E(\omega_k y_k 1_k(s))$$

Ainsi, pour que $E(\hat{t}_{yU}) = t_{yU}$ il faut que

$$E(\omega_k y_k 1_k(s) = y_k$$

En d'autres termes, il faut que :

$$\omega_k = \frac{1}{\pi_k}$$

et l'estimateur ainsi défini est :

$$\hat{t}_{yU} = \sum_{s} \frac{y_k}{\pi_k}$$

$\omega_k y_k$ représentent les valeurs pondérées de y_k alors que $\frac{y_k}{\pi_k}$ représentent les valeurs dilatées de y_k. Dès lors $\hat{t}_{yU_{\omega_k}} = \hat{t}_{yU_{\pi_k}}$ est appelée **estimateur par valeur dilatées** de y ou encore estimateur de **Horvitz-Thompson**.

Remarque : dans le cas du plan SI et du plan BE, tous les éléments ont la même probabilité d'inclusion. Par conséquent l'estimateur s'écrit comme suit.

$$\hat{t}_{yU_{SI}} = \frac{1}{\pi} \sum_{s} y_k$$

Avec $\pi = \frac{n}{N}$

$$\hat{t}_{yU_{BE}} = \frac{1}{\pi} \sum_{s} y_k$$

Avec $\pi = \pi_0$

Calcul de la variance dans l'estimation par valeur dilatées

En notant par $\check{y}_k$ la valeur dilatée de y_k telle que :

$$\check{y}_k = \frac{y_k}{\pi_k} = \left(\frac{1}{\pi_k}\right) y_k = \omega_k y_k$$

La variance du total $var(\hat{t}_{yU})$ s'exprime comme suit :

$$var(\hat{t}_{yU}) = var\left(\sum_s \check{y}_k\right) = var\left(\sum_U \check{y}_k 1_k(s)\right) = \sum\sum_U \Delta_{k,l}\check{y}_k\check{y}_l$$

NB : Pour une variable z, on a

$$\sum\sum_U z_{k,l} = \sum_{k\in U}\sum_{l\in U} z_{k,l}$$

En observant que cette variance n'est autre qu'une somme sur $U \times U$ avec des probabilités d'inclusion $\pi_{k,l}$, on peut montrer que l'estimateur (sans biais) par les valeurs dilatées de cette somme est :

$$\widehat{\text{var}}(\hat{t}_{yU}) = \sum\sum_s \check{\Delta}_{k,l}\check{y}_k\check{y}_l$$

Où

$$\check{\Delta}_{k,l} = \frac{\Delta_{k,l}}{\pi_{k,l}}$$

Mais cet estimateur sans biais présente quelques inconvénients : il peut prendre des valeurs négatives, de plus les $\pi_{k,l}$ interviennent en dénominateur et si certains sont nuls, cette formule n'est plus applicable. Enfin, la somme double peut être difficile à calculer. C'est pourquoi il existe des formules d'approximation que nous verrons plus loin.

Variance dans le cadre d'un plan de taille fixe

Pour un plan de taille fixe, la variance l'estimateur du total s'écrit :

$$var(\hat{t}_{yU}) = -\frac{1}{2}\sum\sum_U \Delta_{k,l_k}(\check{y}_l - \check{y}_l)^2$$

Et un estimateur sans biais de cette variance à partir de l'échantillon est :

$$\widehat{var}(\hat{t}_{yU}) = -\frac{1}{2}\sum\sum_s \Delta_{k,l_k}(\check{y}_l - \check{y}_l)^2$$

Ce dernier estimateur est appelée estimateur de **Sen-Yates-Grundy**.

5.7.2. Estimation de la moyenne par les valeurs pondérées

Etant donné un plan de sondage dont les probabilités d'inclusion de premier et de second ordre sont π_k et $\pi_{k,l}$, , l'estimation de la moyenne par les valeurs pondérées de Horwitz-Thompson est égale à l'estimateur Horwitz-Thompson du total divisée par la taille de la population N. Elle se présente alors comme suit :

$$\bar{y}_U = \frac{1}{N} t_{yU}$$

Sa valeur estimée se présente alors comme suit :

$$\widehat{\bar{y}_U} = \frac{1}{N} \hat{t}_{yU}$$

avec

$$\hat{t}_{yU} = \sum_s \frac{y_k}{\pi_k} = \sum_s \omega_k y_k$$

On peut aisément montrer que cet estimateur est sans biais car

$$E(\widehat{\bar{y}_U}) = \bar{y}_U$$

Sa variance se présente comme suit :

$$\text{var}(\widehat{\bar{y}_U}) = \frac{1}{N^2} \sum \sum_U \Delta_{k,l} \check{y}_k \check{y}_l$$

Sur l'échantillon cette variance se présente :

$$\text{var}(\widehat{\bar{y}_U}) = \frac{1}{N^2} \sum \sum_s \Delta_{k,l} \check{y}_k \check{y}_l$$

Remarque

Il arrive que la taille de la population N ne soit pas connue avec précision. Dans ce cas pour estimer la moyenne on utilise l'estimateur de **Hajek** qui se présente comme suit :

$$\widehat{\bar{y}_U} = \frac{\hat{t}_{yU}}{\widehat{N}}$$

Avec

$$\hat{t}_{yU} = \sum_s \frac{y_k}{\pi_k}$$

$$\widehat{N} = \sum_s \frac{1}{\pi_k} = \sum_s \omega_k$$

$$\widehat{\bar{y}_U} = \frac{\sum_s \frac{y_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}} = \frac{\sum_s \breve{y}_k}{\sum_s \omega_k}$$

NB : Lorsque les probabilités d'inclusion sont connues, la taille de la population se détermine comme la somme des inverses des probabilités d'inclusions (i.e somme des pondérations).

5.8. Le plan de tirage simple à probabilités inégales avec remise

On a déjà étudié le cas d'un tirage avec remise à probabilités égales où l'on tire m objets dans une population U de N, avec remise entre deux tirages. A chaque tirage, un élément particulier de la population est tiré avec la probabilité $\frac{1}{N}$. On a montré que la probabilité d'inclusion dans l'échantillon est :

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m$$

Remarquons que $\left(1 - \frac{1}{N}\right)^m$ représente la probabilité que l'élément k apparaît donc 0 fois parmi les m objets tirés. Dès lors π_k représente la probabilité qu'il soit tiré au moins une fois. On a : $\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m \approx \frac{m}{N}$ (si $\frac{1}{N}$ est faible i.e l'échantillon est de petite taille par rapport à la population finie ; propriétés de développement limité).

Supposons maintenant que les probabilités de tirage soient inégales c'est-à-dire $\neq \frac{1}{N}$. Supposons alors que chaque élément k a maintenant une probabilité de tirage p_k telle que $p_k > 0$ et $\sum_U p_k = 1$. Dès lors la probabilité d'inclusion de k se définit comme suit :

$$\pi_k = 1 - (1 - p_k)^m$$

Si p_k est faible (i.e p_k est faible), on peut écrire :

$$\pi_k = mp_k$$

En utilisant ces propriétés, l'estimateur du total se présente comme suit :

$$\hat{t}_{yU} = \frac{1}{m}\sum_{k=1}^{m}\frac{y_k}{p_k}$$

Cet estimateur est appelé estimateur de **Hansen-Hurwitz** (cas du plan de tirage simple à probabilité inégales avec remise).

Cet estimateur est sans biais car son espérance $E(\hat{t}_{yU}) = t_{yU}$.

Sa variance est :

$$var(\hat{t}_{yU}) = \frac{1}{m}\left[\sum_{k=1}^{N} p_k \left(\frac{y_k}{p_k} - \hat{t}_{yU}\right)^2\right]$$

La valeur estimée de cette variance sur l'échantillon est :

$$\widehat{var}(\hat{t}_{yU}) = \frac{1}{m(m-1)}\sum_{k=1}^{m}\left(\frac{y_k}{p_k} - \hat{t}_{yU}\right)^2$$

5.9. Le plan de tirage systématique

On peut distinguer plusieurs types de tirage systématique : le tirage simple, le tirage répété et le tirage proportionnel à la taille.

5.9.1. Le plan de tirage systématique simple

Exemple introductif

Supposons qu'un ensemble de dossiers d'épaisseur à peu près constante soit stocké dans 12 étagères de 60 cm de longueur chacune et qu'on veuille tirer un échantillon de 100 dossiers. Si le rangement n'est pas lié à la variable d'étude, on peut procéder ainsi : il y a 60 × 12 = 720 cm de dossiers. On peut prendre un dossier tous les r = 720/100 = 7,2cm. Pour amorcer le tirage, on tire un nombre aléatoire entre 0, 7.2 en utilisant si possible la loi uniforme. Ensuite on démarre à partir du dossier le plus proche du résultat. On n'a pas de liste des dossiers mais une organisation qui permet d'y accéder.

Définition

Soit une population qui se présente dans un certain ordre. Soit N la taille de la population et n la taille de l'échantillon à tirer. Supposons pour simplifier l'écriture que : N = na. On appelle a le pas d'échantillonnage ou de tirage. Pour

tirer un échantillon de taille n dans cette population, selon un plan systématique, on suit les étapes suivantes :

– on tire un premier élément r *uniformément* sur les a premiers éléments de la population,

– on prend comme échantillon, sr formé des éléments r; $r + a$; $r + 2a$; $\dots r + (n-1)a$ selon une progression arithmétique jusqu'à obtenir la taille d'échantillon souhaitée.

On peut donc remarquer que le tirage systématique commence comme un plan SI, puisque le premier élément est tiré sans aucune contrainte particulière. Le tirage doit être simplement aléatoire.

5.9.1.1. Estimation du total et de la moyenne dans un plan systématique

Le total de la population se calcul de la même manière que dans un plan SI car tous les individus ont la même probabilité d'inclusion (fixé à partir de la probabilité du premier tirage). On a alors :

$$\hat{t}_{yU} = N\bar{y}_s$$

Où $\bar{y}_s$ est la moyenne sur l'échantillon obtenue. On a :

$$\bar{y}_s = \frac{1}{n}\sum_s y_k$$

Ainsi, on a

$$\hat{t}_{yU} = N\bar{y}_s = N\left(\frac{1}{n}\sum_s y_k\right)$$

On peut alors aisément deviner l'expression de l'estimateur de la moyenne de la population. En effet, on a :

$$\widehat{\bar{y}_U} = \bar{y}_s = \sum_s y_k$$

5.9.1.2. Bonnes pratiques lors des tirages systématiques simples

Si l'on dispose d'une information auxiliaire, par exemple une variable x, connue sur U telle que $y_k = \beta_0 + \beta_1 x$ il est recommandé de trier la population suivant x avant de faire le tirage. Cela augmente l'hétérogénéité des échantillons.

Mais si on n'a pas d'autres informations sur la population, avant d'y faire le tirage, il est fortement recommandé de faire d'abord un tri aléatoire des données. Pour faire un tri aléatoire des données, la démarche est la suivante :

1- Générer une variable aléatoire supplémentaire suivant par exemple une loi uniforme.
2- Ensuite effectuer un tri croissant (ou décroissant) selon les valeurs de cette variable
3- Numéroter maintenant les observations. Ce sont ces numéros qui serviront à tirer le premier élément.

5.9.1.3. Autres particularités du tirage systématique simple

Pour un ordonnancement particulier de la population, il n'y a que a échantillons possibles. Chacun de ces a échantillons forme en quelque sorte une « grappe » au sens de la théorie des sondages : une grappe est un sous ensemble de la population tel que dès que le sous ensemble est tiré, on tire tous les individus qu'il contient. (Le tirage en grappe proprement dit sera abordé plus loin). Par comparaison, dans un plan SI, il y a C_c^N échantillons possibles et tous les échantillons ont les mêmes chances. En effet :

- la probabilité d'inclusion de premier ordre d'un élément k est :

$$\pi_k = \frac{1}{a}$$

-la probabilité d'inclusion de second ordre est pour deux élément k et l est :

$$\pi_k = \begin{cases} \frac{1}{a} \ si\ k\ et\ l\ appartient\ au\ même\ échantillon \\ 0\ \ sinon \end{cases}$$

Notons qu'on peut effectuer plusieurs tirages systématiques indépendants sur la même population pour constituer l'échantillon. Ainsi, pour chaque tirage, lorsque celui-ci est effectué sans remise de l'élément (aléatoirement tiré), alors chaque élément appartiendra à un et un seul échantillon. Un cas particulier de mode de tirage est le tirage systématique répété (présenté ci-après).

5.9.2. Le tirage systématique répété

On peut répéter des tirages systématiques avec différents points de départ. Etant donné la population de taille N, on doit tirer un échantillon systématique de taille n. On peut soit le tirer en une fois avec un pas de tirage $a = \frac{N}{n}$. C'est le cas classique. Ou bien on peut réaliser systématique répété c'est-à-dire on tire m échantillons systématiques de taille de chacun est $n' = \frac{n}{m}$. En d'autres termes, on subdivise l'échantillon final en m sous-échantillons. Chaque sous-échantillon sera alors tiré en utilisant un tirage systématique avec un pas $a' = m\frac{N}{n}$ et son propre point de départ (tiré aléatoirement indépendamment des autres). Ce tirage peut être avec remise ou sans remise.

On obtient alors différentes estimations de la même quantité (total ou moyenne) et on peut ensuite déduire une estimation de la variance.

Chaque échantillon s_h ; h = 1,…,m donne une estimation sans biais de la moyenne telle que :

$$\bar{y}_{s_h} = \frac{1}{n'}\sum_{s_h} y_k$$

La variance de y_k sur chaque échantillon reste bien :

$$S_{ys_h}{}^2 = \frac{1}{n'-1}\sum_{k=1}^{m}\left(y_k - \bar{y}_{s_h}\right)^2$$

On déduit immédiatement une estimation sans biais de $\bar{y}_U$

$$\widehat{\bar{y}_U} = \frac{1}{m}\sum_{h=1}^{m}\bar{y}_{s_h}$$

Pour l'estimation de la variance de cet estimateur, on va distinguer deux selon que les tirages des m premiers points soient des tirages avec remise ou sans remise.

Dans le cas du tirage avec remise on a :

$$\widehat{var}(\widehat{\bar{y}_U}) = \frac{1}{m(m-1)}\sum_{h=1}^{m}\left(\bar{y}_{s_h} - \widehat{\bar{y}_U}\right)^2$$

Et dans le cas du tirage sans remise, on a :

$$\widehat{var}(\widehat{\bar{y}_U}) = \frac{(1-\frac{m}{a})}{m(m-1)}\sum_{h=1}^{m}\left(\bar{y}_{s_h} - \widehat{\bar{y}_U}\right)^2$$

5.9.3. Tirage systématique proportionnel à la taille

Supposons que le gouvernement veuille réaliser une prévision des recettes de la fiscalité des entreprises (PME et grandes entreprises). Il décide alors réaliser un sondage sur le niveau d'activité des entreprises afin d'avoir une prévision du volume total des activités mesuré par le chiffre d'affaire. Il est clair que dans ce contexte certaines entreprises ont une activité importante et d'autres une activité moindre. Dès lors faire un tirage systématique des entreprises dans une liste alphabétique, n'est pas efficace car cette technique sélectionnera indifféremment des sociétés sans aucune distinction sur le volume d'activité.

En supposant qu'on dispose des informations sur le volume d'activité de l'année précédente. On peut alors faire un tirage systématique en exploitant cette information. On met alors en œuvre un tirage systématique proportionnel au poids. La variable de poids est le chiffre d'affaire de l'année précédente. C'est une **information auxiliaire**.

Notons x_k la mesure de taille de l'unité k, connue quel que soit k et par t_{xU} le total sur la population telle que $t_{xU} = \sum_U x_k$ et par n la taille de l'échantillon désirée, les étapes de ce tirage sont les suivantes :

1. On calcule la quantité $\frac{t_{xU}}{n}$

Ainsi, on met d'office dans l'échantillon tous individus pour lesquels on a $x_k > \frac{t_{xU}}{n}$, on les retire donc de la population. Soit n_1 le nombre d'individus de ce type

2. Posons par

$$p_q = \frac{x_q}{t_{xU}}, q = 1, \dots, N - n_1$$

et par

$$\pi_q = np_q$$

3. On forme la quantité V_k telle que :

$$V_k = \sum_{q=1}^{k} p_q$$

En clair la somme de la probabilité de toutes les observations dont le rang d'apparition sur la liste de la population est inférieur au rang d'apparition de k. On suppose ici que la liste est triée sur les valeurs croissantes de x_k. Bien entendu, les n_1 éléments étant d'office sélectionnés, on prend $V_k = 0$ pour eux.

4. On génère un nombre aléatoire u en utilisant une loi uniforme $U(0, 1)$.

5. L'échantillon sera alors formé des unités telles que

$$V_{k_1-1} < u \leq V_{k_1}$$

$$V_{k_2-1} < u \leq V_{k_2}$$

…

$$V_{k_n-1} < u \leq V_{k_n}$$

On voit que la probabilité que l'unité k soit dans l'échantillon est la longueur de l'intervalle $[V_{k-1}, V_k]$ c'est-à-dire la quantité π_q. Dans cette méthode beaucoup de probabilité d'inclusion d'ordre 2 sont nulles.

5.10. Le plan de tirage stratifié

Exemples introductifs

Exemple 1

1 Une région contient un certain nombre d'écoles primaires. On doit en constituer un échantillon. Si l'on fait un tirage simple d'écoles dans la liste des écoles de la région, que peut-il se passer ? On peut obtenir par hasard :

1 surtout des écoles de faible effectif ce qui biaiserait les résultats si l'on s'intéressait à une variable liée à la taille de l'école, comme la dépense annuelle en électricité par école,

2 seulement des écoles rurales, ce qui biaiserait les résultats si la caractéristique étudiée dépend du caractère rural/urbain de l'école

3 des écoles réparties dans toute la région, sans qu'elles soient pour autant très différentes, ce qui occasionnerait des coûts élevés de collecte des données.

On voit sur cet exemple qu'on doit choisir un plan d'échantillonnage qui tient compte autant que possible, des différences entre niveaux moyens de la variable d'étude et de la répartition géographique de la population, dans différentes sous-populations qu'on appelle strates.

Exemple 2

On doit estimer le chiffre d'affaire total des entreprises d'un certain secteur à partir d'un sondage. Or les entreprises sont d'effectifs très variables et le chiffre d'affaire est lié à la taille de l'entreprise. On voit que si l'on prélève l'échantillon par un plan simple, on aura une grande variabilité de l'estimateur avec par exemple un échantillon essentiellement formé de petites entreprises et une forte sous-estimation. On a donc intérêt à mesurer la variable chiffre d'affaire sur des entreprises de différentes tailles, c'est-à-dire à découper l'ensemble des entreprises en strates définies à partir de la taille et à échantillonner dans les différentes strates. Tenant compte de notre précédente observation sur la variabilité, on voudrait échantillonner proportionnellement plus d'entreprises de grande taille que de petite taille.

Dans chacun de ces deux exemples, le plan de tirage stratifié peut s'avérer utile car c'est une technique simple qui peut grandement améliorer l'efficacité.

5.10.1. Définition

Le plan stratifié est un plan dans lequel :

1 la population étudiée est partitionnée en strates,

2 un plan de sondage est défini pour chaque strate,

3 on tire dans chaque strate un échantillon, indépendamment des échantillons tirés dans les autres strates.

NB : La possibilité de définir des strates (ou une stratification) correspond à l'existence d'une variable auxiliaire dans la base de données. Dans l'exemple des écoles, cette variable auxiliaire peut être la région où est située l'école, le caractère urbain/rural de sa commune. Dans l'exemple des entreprises, la variable auxiliaire peut être l'effectif salarié, discrétisé pour donner des classes.

Soit U la population partitionnée par l'intermédiaire d'une variable auxiliaire en H sous-groupes appelés strates: $U_1, \dots, U_H$. Et y_k la variable d'intérêt observé pour l'élément k de la population. On peut définir les caractéristiques statistiques suivantes :

	Sur la population	Pour la strate h
Effectif	$N = card(U)$	$N_h = card(U_h)$
Total	t_{yU}	t_{yU_h}
Moyenne	$\bar{y}_U$	$\bar{y}_{U_h}$
Variance	$S^2{}_{yU}$	$S^2{}_{yU_h}$

On a les relations suivantes :

$$N = \sum_{h=1}^{H} N_h$$

$$\bar{y}_{U_h} = \frac{1}{N_h} \sum_{k \in h} y_k$$

$$t_{yU} = \sum_{h=1}^{H} t_{yU_h}$$

$$\bar{y}_U = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_{U_h}$$

Il reste maintenant à calculer t_{yU_h} et $\bar{y}_{U_h}$.

En effet, il faut savoir que l'une des particularités importantes du plan stratifié c'est qu'il n'est pas nécessaire d'adopter le même plan de tirage pour toutes les strates. Pour chaque strate U_h on peut choisir **un plan de sondage propre** avec une probabilité d'inclusion $\pi_{h,k}$ **indépendant** des plans des autres strates et on tire un échantillon s_h. L'échantillon total sera alors $s = \cup_1^H s_h$.

Ainsi, en tenant compte de la probabilité d'inclusion, l'estimateur du total sur chaque strate est :

$$\hat{t}_{yU_h} = \sum_{s_h} \frac{y_k}{\pi_{h,k}}$$

Cet estimateur est bien évidemment un estimateur de Horwitz-Thompson (H-T)

La somme des estimateurs des totaux sur les strates forme alors l'estimateur du total de la population. Il se présente comme suit :

$$\hat{t}_{yU} = \sum_{h=1}^{H} \hat{t}_{yU_h}$$

Comme les différents plans sont indépendants, la variance de $\hat{t}_{yU}$est la somme des variances des strates :

$$\widehat{var}(\hat{t}_{yU}) = \sum_{h=1}^{H} \widehat{var}(\hat{t}_{yU_h})$$

5.10.2. Affectation de l'échantillon entre les strates

Dans le plan de tirage stratifié comme le tirage dans chaque strate est indépendant du tirage dans les autres mais aussi comme la taille de l'échantillon est fixée d'avance. Cela pose alors la question de l'affectation de l'échantillon entre les strates. En d'autres termes combien d'observations faut-il tirer dans chaque strate pour constituer l'échantillon final.

Il existe, en effet plusieurs méthodes d'affectation de la taille de l'échantillon. On dénote notamment :

- **la méthode d'affectation proportionnelle à la taille des strates** : dans cette méthode, on calcule le nombre d'observations de la strate en proportion du poids de la strate dans la population. (i.e la proportion du nombre d'observations initiale dans la strate par rapport à la population N $\frac{N_h}{N}$. Par exemple, si la strate représente 10% de la population, on tirera dans cette strate un nombre d'observation égal à 10% de n (la taille de l'échantillon final à tirer). La formule de répartition est la suivante :

$$n_h = n\left(\frac{N_h}{N}\right) = n\left(\frac{N_h}{\sum_{h=1}^{H} N_h}\right)$$

Où n est la taille d'échantillon final désiré et n_h la part de l'échantillon qui sera tiré dans la strate h en adoptant un plan spécifique.

Cette affectation est optimale quand les écart-types dans les strates sont égaux. On l'emploie parfois quand on ignore tout des dispersions dans les strates.

-**la méthode d'affectation proportionnelle au total dans la strate** : Dans cette méthode, il s'agit de calculer les proportions non pas en fonction du nombre

d'éléments mais en fonction du poids de la somme des y_k dans la strate par rapport à la somme de la population. Sa formule est la suivante :

$$n_h = n\left(\frac{t_{yU_h}}{t_{yU}}\right)$$

Ce mode d'affectation est optimal quand les coefficients de variation sont égaux.

NB : Notons aussi qu'on peut utiliser le total de la variable auxiliaire x_k à la place de la variable d'intérêt y_k pour réaliser l'affectation.

Il existe aussi d'autres méthodes d'affection dite «**méthodes d'affectation optimales** » calculée soit en fonction de y_k (on dit alors *y-optimale*) ou calculée en fonction de x_k (auquel cas on dit *x-optimale*). Même si ces méthodes ne sont pas présentées ici, elles restent extrêmement utiles. Elles incluent les deux méthodes présentées ci-dessus comme des particuliers.

5.10.3. La post-stratification

D'une manière sommaire, la post-stratification consiste à estimer les paramètres d'intérêt en stratifiant l'échantillon après son tirage.

Par exemple, supposons qu'on veuille mener une étude sur le revenu sur un échantillon constitué d'hommes et de femmes. Mais lors du tirage, nous n'avons pas exploité cette information auxiliaire pour stratifier l'échantillon. La post-stratification permet d'utiliser l'information auxiliaire pour calculer les paramètres comme si l'échantillon était stratifié.

Le cas le plus courant est la post-stratification après un plan SI.

Supposons une population U partitionnée en H sous-populations comme précédemment et le niveau moyen de la variable y est a priori différent d'une strate à l'autre. On tire un échantillon s dans U de n éléments par plan SI. Pour tout élément k on observe sa valeur y_k et sa strate h_k. On note $s_h = s \cap U_h$ le sous échantillon observé dans la strate h, n_h la taille de s_h et $\bar{y}_{s_h} = \frac{1}{n_h}\sum_{s_h} y_k$.

5.10.3.1. Estimateur post-stratifié du total

L'estimateur post-stratifié du total sur la population se définit alors comme :

$$\hat{t}_{yU} = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$$

La différence essentielle de l'estimateur post-stratifié par rapport à l'estimateur stratifié est que maintenant la moyenne $\bar{y}_{s_h}$est un quotient de 2 v.a.(par rapport au mécanisme de sondage) alors que dans le premier, N_h n'est pas aléatoire.

L'estimateur post-stratifié du total est un estimateur sans biais, on a :

$$E(\hat{t}_{yU}) = t_{yU}$$

Sa variance s'exprime comme suit :

$$Var(\hat{t}_{yU}) = \sum_{h=1}^{H} {N_h}^2 {S^2}_{y_{U_h}} E\left(\frac{1}{n_h} - \frac{1}{N_h}\right)$$

5.10.3.2. Estimateur post-stratifié de la moyenne

L'estimateur post-stratifié de la moyenne sur la population se définit alors comme :

$$\widehat{\bar{y}_U} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{s_h} = \sum_{h=1}^{H} W_h \bar{y}_{s_h}$$

avec

$$W_h = \frac{N_h}{N}$$

et

$$\bar{y}_{s_h} = \frac{1}{n_h} \sum_{s_h} y_k$$

Cet estimateur est sans biais et sa variance est :

$$var\ (\widehat{\bar{y}_U}) = \sum_{h=1}^{H} {W_h}^2 {S^2}_{y_{U_h}} E\left(\frac{1}{n_h} - \frac{1}{N_h}\right)$$

5.11. Les plans de sondage à deux degrés

5.11.1. Généralités

Les plans de sondage à deux ou plusieurs degrés visent à réponde aux insuffisances des plan à un degré notamment pour améliorer l'efficacité de la sélection de l'échantillon mais aussi pour améliorer la qualité des estimateurs. Par exemple, le sondage aléatoire simple (plan SI) a plusieurs limitations. D'abord, il nécessite de disposer d'une base de la population des unités à enquêter. Même si une telle base existe, la population peut être disséminée sur un vaste territoire et un plan SI peut fournir des individus très dispersés et donc entraîner des coûts de collecte très variables. Le sondage à plusieurs degrés est une réponse à ces limitations.

Le principe général d'un plan de sondage à deux degrés est le suivant.

– D'abord, on partitionne la population U en plusieurs sous-groupes (sous-populations), appelés unités primaires (UP).

– Ensuite, on tire aléatoirement un échantillon d'unités primaires puis dans chaque unité primaire sélectionnée on effectue un tirage des unités secondaires (US). Ce tirage peut-être partiel (donc aléatoire) ou exhaustif où on sélectionne toutes les US présentes dans l'UP sélectionnée (sondage par grappes).

Notons que dans ce mécanisme on n'a pas besoin d'avoir une base de sondage de la population des unités secondaires comme dans le plan SI. On a simplement besoin de la base des UP. La plupart du temps ces bases sont disponibles auprès de l'administration. Par exemple, liste des régions, commune, liste des quartiers, etc…

Dans cette section, nous étudions d'abord le cas particulier du sondage par grappes avant de généraliser les résultats.

5.11.2. Etude d'un plan particulier : le sondage par grappes

Le sondage par grappes est un sondage à deux degrés particulier où au deuxième degré on observe tous les individus, et pas seulement un échantillon, de chaque UP sélectionnée au premier degré. Ainsi, sur l'UP sélectionnée g, on connaît $t_g = \sum_{U_g} y_k$.

On peut simplement noter que le sondage systématique étudié dans les sections précédentes est en quelque sorte un sondage en grappes particulier où on tire une seule grappe. Pour le comprendre, il faut réaliser plusieurs tirages systématiques

de suite où les nombres tirés en premier sont faits avec aléatoirement sans remise.

Pour revenir au tirage par grappe proprement dit, adoptons les notations suivantes. Soit :

- π_{Ig} la probabilité d'inclusion de l'UP g dans l'échantillon des UP ;
- $\pi_{Igg'}$ la probabilité d'inclusion de second ordre de deux UP g et g' $g \neq g'$;
- $\Delta_{Igg'}$ la covariance des indicatrices d'inclusion des UP g et g'.
- s_I l'échantillon d'UP obtenu au premier degré et
- s_g la taille totale de la sous-population g

Le sondage en grappes correspond au cas où au deuxième degré on prend toutes les US des UP sélectionnées à la première étape.

Estimation du total en sondage par grappes

Le total sur la population s'écrit alors comme suit :

$$t_{yU} = \sum_{U} y_k = \sum_{U_I} \sum_{U_g} y_k = \sum_{U_I} t_{yg}$$

En posant :

$$\check{t}_{yg} = \frac{t_{yg}}{\pi_{Ig}}$$

On peut calculer l'estimateur par valeur dilatées du total, sa variance et une estimation de sa variance par une application directe du résultat général.
En effet, l'estimateur du total t_{yU} par les valeurs dilatées dans le sondage en grappes est :

$$\hat{t}_{yU} = \sum_{s_I} \frac{t_{yg}}{\pi_{Ig}} = \sum_{s_I} \check{t}_{yg}$$

Sa variance est :

$$var(\hat{t}_{yU}) = \sum \sum_{U_I} \Delta_{Igg'} \check{t}_{yg} \check{t}_{yg'}$$

Elle est estimée sans biais sur l'échantillon par :

$$\widehat{var}(\hat{t}_{yU}) = \sum \sum_{U_I} \frac{\Delta_{Igg'}}{\pi_{Igg'}} \check{t}_{yg} \check{t}_{yg'}$$

Considérons le cas où le plan de degré 1 est de taille fixe : $n_I = card(s_I)$ est constant. Les expressions qu'on a obtenues quand la taille est fixe, s'appliquent ici aussi. On obtient :

$$var(\hat{t}_{yU}) = -\frac{1}{2}\sum\sum_{U_I} \Delta_{Igg'}(\check{t}_{yg} - \check{t}_{yg'})^2$$

dont un estimateur sans biais est :

$$\widehat{var}(\hat{t}_{yU}) = -\frac{1}{2}\sum\sum_{s_I} \check{\Delta}_{Igg'}(\check{t}_{yg} - \check{t}_{yg'})^2$$

Remarque :

On observe sur cette formule que si l'on peut avoir des $t_{yg} = 0$, c'est-à-dire $\pi_{Ig} = bt_{yg}$, la variance est nulle. Evidemment, on ne connaît pas les t_{yg} mais on peut quand même exploiter cette remarque.

– Si on dispose d'une variable x dont on connaît le total t_{xg} par grappe et qu'on sait de plus que : $t_{xg} \approx ct_{yg}$, on choisira :

$$\pi_{Ig} = \frac{n_I}{\sum_{U_I} t_{xg}} t_{xg}$$

Si l'on sait que les $\bar{y}_{U_g}$sont à peu près constants, on choisira π_{Ig} proportionnel à N_g.

5.11.3. Etude du plan de tirage à deux degrés généralisé

Nous abordons maintenant le cas où au deuxième degré, on n'observe pas toutes les unités secondaires (US) des UP tirées au premier degré mais seulement un échantillon (contrairement au tirage par grappe).

Notation : soit

- s_g l'échantillon d'US tirées dans l'UP g.
- $\pi_{k|g}$ la probabilité d'inclusion du premier ordre de l'US k de l'UP g dans s_g.
- $\pi_{k,l|g}$ la probabilité d'inclusion du deuxième ordre des US k et l de l'UP g
- $\Delta_{k,l|g}$ la covariance des indicatrices d'inclusion de k et l de U_g dans s_g.
- $s = \cup_{g \in s_I}$ l'échantillon de toutes les US tirées.

Nous supposons que les plans sur les UP sont indépendants.

La probabilité d'inclusion de $k \in \cup_g$ dans s est :

$$\pi_k = P\big(g \in s_I \,\&k \in\ s_g\big) = P(g \in s_I\ ) \times P\big(k \in\ \cup_g\ |_{g \in s_I}\big) = \pi_{Ig}\pi_{k|g}$$

Estimateur du total

On en tire l'estimateur du total par les valeurs dilatées dans le plan à deux degrés :

$$\hat{t}_{yU} = \sum_{g \in s_I} \sum_{k \in\ s_g} \frac{y_k}{\pi_{Ig}\pi_{k|g}} = \sum_{g \in s_I} \frac{1}{\pi_{Ig}} \sum_{k \in\ s_g} \frac{y_k}{\pi_{k|g}}$$

Mais on reconnait que

$$\sum_{k \in\ s_g} \frac{y_k}{\pi_{k|g}}$$

est l'estimateur Horvitz–Thompson du total t_{yg} sur U_g sachant que $g \in s_I$ c'est-à-dire que l'UP g est sortie au premier tirage. On note

$$\hat{t}_{yg|g \in s_I} = \sum_{k \in\ s_g} \frac{y_k}{\pi_{k|g}}$$

D'où finalement :

$$\hat{t}_{yU} = \sum_{s_I} \frac{1}{\pi_{Ig}} \hat{t}_{yg|g \in s_I}$$

5.12. Les mesures de précision d'un plan de sondage

Un résultat de sondage doit toujours s'accompagner d'une mesure de précision des estimateurs obtenus. Le premier indicateur d'un plan de sondage est son efficience. En effet, puisqu'un plan est choisi parmi tant d'autres, il s'avère donc utile d'évaluer ce que ce plan apporte en termes de précision par rapport à un plan de référence. Pour de telle évaluation, on calcule l'« effet plan » ou *design effect* en anglais.

Mais en plus de l'effet, on distingue plusieurs autres indicateurs de précision. Il s'agit notamment du coefficient de variation qui est une mesure de variabilité relative mais aussi les marges d'erreur (absolue et relative) liées à la largeur d'un intervalle de confiance de l'estimateur.

5.12.1. Effet plan

Les plans de sondage simples que nous avons présentés sont rarement utilisés seuls. Ce sont le plus souvent les éléments d'un plan de sondage complexe qu'un statisticien d'enquête est amené à construire. Schématiquement, le budget dont on dispose permet d'interroger un certain nombre d'individus et on se pose la question du choix du plan : faut-il bâtir un plan complexe ou bien peut-on se contenter du plan SI ? On est donc amené à comparer la précision d'un plan quelconque à celle d'un plan SI de même taille par le rapport des variances des estimateurs.

On appelle, effet plan d'un certain plan P, fournissant un estimateur sans biais de t_y, le rapport le rapport de la variance de l'estimateur du total dans ce plan à la variance de l'estimateur du total dans le plan SI, pour une même taille d'échantillon. Il est exprimé comme suit :

$$\rho_{\hat{t}_{yU}} = \frac{var(\hat{t}_{yU})_P}{var(\hat{t}_{yU})_{SI}}$$

Une valeur $\rho = 3$ s'interprète comme suit : la variance est 3 fois plus grande lorsque l'échantillon est sélectionné par le plan P plutôt que par le plan SI. Une interprétation alternative est : 1/3 de la taille de l'échantillon tiré par le plan P suffisent au plan SI pour obtenir le même estimateur. Cela veut donc dire que le plan SI est plus efficient.

Cette dernière interprétation est très intuitive car elle permet définir la taille d'échantillon effective. Celle-ci se présente comme suit :

$$n_{eff} = \frac{n_P}{\rho_{\hat{t}_{yU}}}$$

Où n_{eff} représente la taille d'échantillon effective et n_P la taille d'échantillon tirée dans le plan P.

5.12.2. Coefficient de variation

Pour une population finie U et une variable d'intérêt $y \geq 0$ le coefficient de variation est par définition :

$$CV_{yU} = \frac{S_{yU}}{\bar{y}_U}$$

Un coefficient de variation est équivalent à une erreur relative en physique. Notons qu'il est défini pour une quantité $\bar{y}_U \geq 0$. C'est la variabilité des y rapportée à leur moyenne. Il est sans dimension, il permet donc de comparer des grandeurs exprimées dans des unités différentes.

Et pour tout autre estimateur θ, dont l'estimateur sans biais $\hat{\theta}$, le coefficient de variation se définit comme suit :

$$CV(\hat{\theta}) = \frac{\sqrt{\text{var}(\hat{\theta})}}{\theta}$$

Une estimation par substitution habituellement employée est :

$$\widehat{CV}(\hat{\theta}) = \frac{\sqrt{\widehat{\text{var}}(\hat{\theta})}}{\hat{\theta}}$$

Dans un plan SI par exemple, le coefficient de variation de $\hat{t}_{yU}$ est :

$$\widehat{CV}(\hat{t}_{yU}) = \frac{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}}}{\bar{y}_U}$$

Notons que $\bar{y}_s$ et $\hat{t}_{yU}$ ont le même coefficient de variation.

5.12.3. Intervalle de confiance et marges d'erreur

Rappel :

Soit $\hat{\theta}$ l'estimateur sans biais d'un paramètre θ calculé sur un échantillon de grande taille. On a généralement la propriété suivante :

$$\hat{\theta} \rightsquigarrow N\left(\theta, var(\hat{\theta})\right)$$

Où $var(\hat{\theta}) \to 0$ quand $n \to +\infty$.

Cette propriété asymptotique permet de construire des intervalles de confiance (IC) approchés pour θ en se fixant un seuil d'erreur α.

Ainsi, pour tout paramètre θ, l'intervalle de confiance se présente comme suit :

$$IC_\theta = \left[\hat{\theta} - Z_{1-\frac{\alpha}{2}} \times \sqrt{\widehat{\text{var}}(\hat{\theta})}\,;\ \hat{\theta} + Z_{1-\frac{\alpha}{2}} \times \sqrt{\widehat{\text{var}}(\hat{\theta})}\right]$$

où $Z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite et où on a remplacé la variance par une estimation de celle-ci. Par exemple avec $\alpha = 0{,}05$ on a $Z_{1-\frac{\alpha}{2}} = Z_{0{,}975} = 1{,}96$

Marge d'erreur absolue (MEA)

On appelle **marge d'erreur absolue** la demi-longueur de l'IC :

$$MEA = Z_{1-\frac{\alpha}{2}} \times \sqrt{\widehat{\text{var}}(\hat{\theta})}.$$

Marge d'erreur relative (MER)

Le rapport entre la MEA et la moyenne donne la marge d'erreur relative.

$$MER = \varepsilon = \frac{Z_{1-\frac{\alpha}{2}} \times \sqrt{\widehat{\text{var}}(\hat{\theta})}}{\hat{\theta}} = Z_{1-\frac{\alpha}{2}} \times \widehat{CV}(\hat{\theta})$$

La marge d'erreur relative est un indicateur utile dans le calcul de la taille de l'échantillon.

En appliquant la formule de l'intervalle de confiance au cas du total, de la moyenne, de la proportion et du ratio on a :

Intervalle de confiance pour la moyenne

$$IC_{\bar{y}_U} = \left[\bar{y}_s - Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}} \; ; \; \bar{y}_s + Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_{yU}} \right]$$

En remplaçant S_{yU} par son estimateur sur l'échantillon, on a finalement :

$$IC_{\bar{y}_U} = \left[\bar{y}_s - Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_{ys}} \; ; \; \bar{y}_s + Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_{ys}} \right]$$

Intervalle de confiance pour le total

$$IC_{t_{yU}} = \left[\hat{t}_{yU} - Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) N S_{yU}} \; ; \; \hat{t}_{yU} + Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) N S_{yU}} \right]$$

En remplaçant S_{yU} par son estimateur sur l'échantillon, on a finalement :

$$IC_{t_{yU}} = \left[\hat{t}_{yU} - Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) N S_{ys}} \; ; \; \hat{t}_{yU} + Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) N S_{ys}} \right]$$

Intervalle de confiance pour une proportion

$$IC_p = \left[\hat{p} - Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} \sqrt{\left(\frac{N}{N-1}\right) p(1-p)} \; ; \; \hat{p} + Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} \sqrt{\left(\frac{N}{N-1}\right) p(1-p)} \right]$$

En utilisant les estimateurs de la variance sur l'échantillon, on a :

$$IC_p = \left[\hat{p} - Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} \sqrt{\left(\frac{n}{n-1}\right) \hat{p}(1-\hat{p})} \; ; \; \hat{p} - Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} \sqrt{\left(\frac{n}{n-1}\right) \hat{p}(1-\hat{p})} \right]$$

5.13. Niveau précision et détermination de la taille de l'échantillon

En pratique, dans la démarche de calcul de la taille de l'échantillon, on choisit d'abord la précision (marge d'erreur absolue ou relative) et on en déduit la taille de l'échantillon à tirer pour atteindre cette précision. Si l'on n'a pas d'ordre de grandeur pour S_{yU} on peut faire un premier sondage dont les résultats permettront d'avoir un ordre de grandeur de S_{yU}.

5.13.1. Utilisation de la marge d'erreur relative

Taille d'échantillon pour estimer une moyenne

Supposons donc un niveau de confiance $1-\alpha$ fixé. $CV(\hat{t}_{yU})$ est propre à la population U et on ne peut donc pas le choisir. On doit agir sur n pour diminuer la marge d'erreur. Ainsi, si on veut une marge d'erreur relative de ε, on doit choisir n tel que :

$$\varepsilon \geq Z_{1-\frac{\alpha}{2}} \times CV_{yU}$$

Si la taille de la population est grande, $\frac{1}{N}$ est négligeable et la condition sur n devient :

$$n \geq \frac{Z^2{}_{1-\frac{\alpha}{2}} \times CV^2{}_{yU}}{\varepsilon^2}$$

Mais comme on ne connaît pas CV_{yU}, on doit l'estimer sur un premier échantillon de petite taille.

Taille d'échantillon pour estimer une proportion

Pour N suffisamment grand, on doit choisir n tel que :

$$\varepsilon \geq Z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{1}{n}}\sqrt{\frac{(1-p)}{p}}$$

c'est-à- dire :

$$n \geq \frac{Z^2{}_{1-\frac{\alpha}{2}} \times (1-p)}{\varepsilon^2 p}$$

On peut vérifier que (1 - p)/p décroit de $+\infty$ à 0 quand p croit de 0 à 1. Si l'on sait que p est supérieur à une certaine valeur p_0, on pourra choisir :

$$n \geq \frac{Z^2{}_{1-\frac{\alpha}{2}} \times (1-p)}{\varepsilon^2 p_0}$$

5.13.2. Utilisation de la marge d'erreur absolue

Au lieu de s'intéresser à la marge d'erreur relative, on peut s'intéresser à la marge d'erreur absolue, la demi-longueur de l'intervalle de confiance

Taille d'échantillon pour estimer une moyenne

Partant à la fois de l'expression de l'intervalle de confiance de la moyenne

$$IC_{\bar{y}_U} = \left[\bar{y}_s - Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_{ys}}\ ;\ \bar{y}_s + Z_{1-\frac{\alpha}{2}} \times \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_{ys}}\right]$$

et de la taille d'échantillon précédemment calculée

$$n \geq \frac{Z^2{}_{1-\frac{\alpha}{2}} \times CV^2{}_{yU}}{\varepsilon^2}$$

Ainsi étant donné un niveau de confiance fixé et une marge d'erreur choisie, l et si on suppose N grand on voit qu'on doit prendre un échantillon de taille n vérifiant

$$n \geq \frac{Z^2{}_{1-\frac{\alpha}{2}} \times S^2{}_{yU}}{l^2}$$

Taille d'échantillon pour estimer une proportion

Partant de l'expression de l'intervalle de confiance de la proportion, et étant donné un niveau de confiance fixé et une marge d'erreur choisie, l sur la proportion à estimer, on voit qu'on doit prendre un échantillon de taille n vérifiant

$$n \geq \frac{Z^2{}_{1-\frac{\alpha}{2}} \times Np(1-p)}{p(1-p) + \frac{l^2}{Z^2{}_{1-\frac{\alpha}{2}}}(N-1)}$$

Si on néglige la correction de population finie, la condition devient :

$$n \geq \frac{Z^2{}_{1-\frac{\alpha}{2}} \times p(1-p)}{l^2(N-1)}$$

Taille d'échantillon pour estimer un total

Il est immédiat de transposer ce qu'on a obtenu pour une moyenne à un total. Si on doit estimer un total avec une marge d'erreur l, et si on suppose N grand, on voit à partir de l'expression de l'intervalle de confiance que la condition devient :

$$n \geq \frac{{Z^2}_{1-\frac{\alpha}{2}} \times N^2 {S^2}_{yU}}{l^2}$$

Exemple d'application

Un société souhaite estimer le montant moyen des commandes d'un produit sur une population de 1800 clients. Sachant par le passé, que la moyenne et l'écart-type des commandes est 6 € et 4 €. On choisit un niveau de confiance de 95%. Quelle taille d'échantillon faut-il prendre pour estimer (1) le montant moyen des commandes avec une marge d'erreur relative de 7% ? (2) le montant total des commandes avec une erreur absolue de 400.

Réponses :

- Pour (1) on utilise $\geq \frac{{Z^2}_{1-\frac{\alpha}{2}} \times {CV^2}_{yU}}{\varepsilon^2}$, on trouve $n \geq 349$.
- -Pour (2), on utilise $\geq \frac{{Z^2}_{1-\frac{\alpha}{2}} \times N^2 {S^2}_{yU}}{l^2}$, on trouve $n \geq 77{,}79$; en prenant un échantillon de 78 clients on estimera le montant total des commandes avec une marge d'erreur de moins de 400, à 95%.

NB : Les méthodes de calcul présentées ici sont faites en fonction de la marge d'erreur relative ou la marge d'erreur absolue. Il peut arriver qu'on détermine la taille de l'échantillon en se basant sur une amplitude maximale souhaitée pour l'intervalle de confiance ΔIC_n. Dans ce cas, il faut faire la différence entre la borne supérieure et la borne inférieure et égaliser cette différence à la valeur de l'amplitude souhaitée. Et ainsi, tirer la valeur n pour avoir la taille minimale de d'échantillon.

Bibliographie

Ardilly P.; (2006), Techniques de sondages, Technip, 1st Editions, 675p

Cochran W. G., (1977), Sampling techniques, Wiley, 3rd Edition, 448p

Dagnelie P., (2011), Statistique théorique et appliquée. Tome 2. Inférence statistique à une et à deux dimensions, De Boeck, 736 p.

Dagnelie P. (2013), Statistique théorique et appliquée. Tome 1. Statistique descriptive et bases de l'inférence statistique, De Boeck, 517 p.

Lecoutre J.-P., (2002), Statistiques et probabilités, manuel et exercices corrigés, Dunod, 3e édition, 296p

Lohr S. L., (1999), Sampling : Design and Analysis, Duxbury Press; 2nd edition, 608p.

Saporta G. (1990), Probabilités, analyse de données et statistique, Technip, 3e édition, 656p.

Särndal K.E., Swenson B., and Wretmann J. 1992, Model Assisted Survey Sampling ; Springer-Verlag New York, 2nd Edition 1,695p.

Thompson S.K.,(1992), Sampling, Wiley, 3nd Edition, 472 pages..

Tillé Y., (2001), Théorie des sondages – Échantillonnage et estimation en populations finies Cours et exercices avec solutions, Collection: Sciences Sup, Dunod, 296p.

Annexe

Les règles d'utilisation des tables statistiques usuelles

Utilisation de la table de la loi normale centrée réduite

La table de la loi normale centrée réduite présente sur la première ligne et la première colonne les valeurs des fractiles Z (encore appelés quantiles). Dans la première colonne, on lit la valeur du quantile Z à un décimal près. Et sur la première ligne, on lit le nombre de décimaux restants à 10-2 près. Ainsi c'est en faisant la somme d'un élément en colonne et d'un élément en ligne qu'on obtient la valeur de Z. Par exemple $Z = 1.96$ s'obtient en faisant 1,9+0,06 où 1,9 provient de la première colonne alors que 0,06 provient de la première ligne. Ainsi pour trouver la valeur de n'importe quel fractile, on procède à cette décomposition. Par exemple $Z = 3.37$ se lit en décomposant 3.3 (en colonne) et 0,07 (en ligne).

En plus de la première colonne extérieure et la première ligne extérieure (qui permettent de connaitre la valeur de Z), on se réfère aux cellules intérieures pour lire les probabilités associées aux fractiles. La probabilité associée à une fractile correspond à la valeur contenue dans la cellule qui se trouve au croisement des deux membres qui forment la valeur de la fractile. Par exemple, sachant que 2,47 est formée par 2,4(en ligne) et 0,07 (en colonne), alors, la probabilité correspondant à 2,47 se trouve au croisement entre 2,4 et 0,07. Cette valeur est égale à 0,993.

Cette compréhension de la structure de la table de loi normale est extrêmement importante car elle servira à déterminer les fractiles (lorsque l'on connait les probabilités) ou à l'inverse déterminer les probabilités (lorsque l'on connait les fractiles).

Lecture des fractiles connaissant les probabilités α

Dans une optique de détermination de la statistique d'un test suivant une loi normale et dont le seuil d'erreur est α, on lit le fractile correspondant à α. Pour cela, on calcule d'abord$(1 - \alpha) + \frac{\alpha}{2}$ c'est-à-dire $1 - \frac{\alpha}{2}$.. Ensuite, on recherche cette valeur dans les cellules intérieures de la table. Une fois cette valeur identifiée, on fait la somme des deux cellules extérieures (en ligne et en colonne) dont le croisement correspond à cette valeur $1 - \frac{\alpha}{2}$ lue dans la table. Par exemple, pour le trouver la statistique (le fractile) correspondant à α =5 %, on calcule d'abord $1 - \frac{\alpha}{2}$ (soit 0,975). Ensuite, en recherchant 0,975 dans les cellules intérieures de la table, on constate que cette valeur se trouve au croisement entre 1,9 et 0.06. Par conséquent le fractile correspondant à 5% est 1,96.

Notons aussi que cette valeur peut être obtenue avec la plus part des logiciels statistiques et économétriques plus ou moins spécialisés. Par exemple, certaines fonctions de Microsoft™ Excel® fournissent les valeurs contenues dans les tables statistiques usuelles. Pour obtenir le fractile correspondant au seuil α, on utilise la formule suivante :

$$= loi.normale.standard.inverse(1 - \frac{\alpha}{2})$$

Où α représente la probabilité (et correspond généralement au seuil d'erreur).

Remarque :

Puisque la loi normale est une loi symétrique, si l'on veut déterminer la valeur opposée du fractile (en vue par exemple de la détermination d'un intervalle de confiance (ou autre), on considère juste l'opposé de ce fractile pour trouver la borne inférieure de l'encadrement.

Lecture des probabilités α connaissant les fractiles

Lorsque l'on connait le fractile, pour déterminer la probabilité correspondante par lecture d'une table de la loi normale centrée et réduite, on décompose d'abord ce fractile en deux éléments (selon la méthode précédemment discutée). Ensuite, on recherche la cellule intérieure de la table se trouvant au croisement (de la ligne et de la colonne extérieure) des deux valeurs. Ainsi, après avoir déterminé cette valeur notée P, on calcule α telle que $1 - \frac{\alpha}{2} = P$ soit $\alpha = 2(1 - P)$. Cette valeur α correspond donc à la probabilité recherchée. Par exemple, pour trouver la probabilité correspondant à 1,53, on décompose d'abord cette valeur entre 1,5 et 0,03. En recherchant dans les cellules intérieures de la table la probabilité se trouvant au croisement de ces deux valeurs, on trouve 0,937. Ainsi puisque cette valeur équivaut à $1 - \frac{\alpha}{2}$, on peut alors en déduire α comme $\alpha = 2(1 - 0{,}937)$. Soit $\alpha = 0{,}126 = 12{,}6\%$

Notons aussi que valeur de la probabilité peut s'obtenir en utilisant également les fonctions de Microsoft Excel. Pour cela, on peut utiliser la formule suivante :

$$= loi.normale.standard.n(q; VRAI)$$

Où q représente la valeur du fractile dont on cherche à déterminer la probabilité.

Remarque :

Dans une démarche de test, la valeur de la probabilité ainsi calculée correspond généralement à la p.value lorsque la détermination de la probabilité porte sur

une statistique de test. En effet dans un test, on connaît à priori le seuil théorique α. Ce seuil est utilisé pour lire la statistique théorique (ou seuil critique) du test. On a alors le couple (α ; S^*). Ensuite en construisant le test et en calculant la statistique du test sous l'hypothèse nulle, on obtient S. Ce qui manque alors c'est la probabilité associée à cette statistique calculée. Elle est dénommée la p.value p_0. Dès lors, pour déterminer la p.value afin de former le couple (p_0 ; S), on lit la probabilité correspondant à S dans la table. C'est donc après avoir définie cette probabilité qu'on forme la règle de décision du test :

- Si $S > S^* \Rightarrow p_0 < \alpha$ alors on rejette H0.
- Si $S < S^* \Rightarrow p_0 > \alpha$ alors on ne peut pas rejeter H0.

Lecture de la table de la loi normale dans le cas d'un encadrement de la fractile

- **Cas d'un encadrement de type** : $P\ (-b < Z < b)$

Lorsque le fractile Z est une valeur encadrée par une borne supérieure b et une borne inférieure -b, pour lire la probabilité pour que Z soit compris entre -b et b, on procède d'abord à un développement comme suit :

$$P\ (-b < Z < b) = P(Z < b) - P(Z < -b)$$

Or, sachant les propriété d'une loi symétrique, on a: $P(Z < -b) = P(Z > b)$. Mais on sait aussi que $P(Z > b) = 1 - P(Z < b)$. Dès lors, on a : $P\ (-b < Z < b) = P(Z < b) - [\,1 - P(Z < b)]$. Au final, après développement, on trouve : $P\ (-b < Z < b) = 2P(Z < b) - 1$.

Cela montre donc que dans une loi symétrique, pour trouver la probabilité d'un encadrement symétrique (qui correspond en général au seuil de confiance), il faut simplement multiplier par 2 la probabilité obtenue en considérant uniquement la borne supérieure (en suivant les méthodes de lectures précédemment présentées). Ensuite retrancher 1 pour trouver la probabilité de l'encadrement (au seuil de confiance). Par exemple, quand on demande de calculer la probabilité pour que Z soit comprise entre -2,72 et 2,72. On lit d'abord la probabilité associée à 2,72 (soit 0,9967). Ensuite, on multiplie cette valeur par 2 et on retranche 1. On trouve alors 0,9934. Ainsi le seuil d'erreur α s'obtient simplement comme est 1-0,9934 soit 0,66%. Il faut noter que dans un encadrement α n'est pas calculée telle que $1 - \frac{\alpha}{2} = P$ mais comme $1 - \alpha = P$.

- **Cas d'un encadrement de type** $P\ (Z < b)$ ou de type $P\ (Z > b)$

Lorsqu'il s'agit d'un encadrement de type $P\ (Z < b)$, on garde sans aucune transformation la valeur lue dans la table (ou obtenue par la fonction : $= loi.normale.standard.n(q; VRAI)$. Ainsi, le seuil d'erreur α s'obtient en utilisant la relation $1 - \alpha = P$.

Mais quand il s'agit d'un encadrement de type $P(U > b)$, on lit d'abord $P(U < b)$, ensuite on calcule $P(U > b) = 1 - P(U < b)$. Ainsi, le seuil d'erreur α s'obtient en utilisant la relation $1 - \alpha = P$.

Utilisation de la table de Student

Lecture des fractiles connaissant les probabilités α

En général la table de Student se présente de telle sorte que les lignes correspondent aux degrés de liberté et les colonnes correspondent aux valeurs des probabilités. Pour utiliser une table se présentant sous ce format, on retrouve d'abord le degré de liberté, puis on lit sur la ligne correspondante (de gauche à droite) jusqu'à trouver la première valeur de t* supérieure au t calculé. Et on retient en haut de la colonne la valeur P correspondante à cette valeur.

NB : Néanmoins, il faut noter que dans la table de Student, le quantile $1 - \frac{\alpha}{2}$ se lit dans la colonne $P = \alpha$ alors que le quantile $1 - \alpha$ se lit dans la colonne $P = 2\alpha$. Cette distinction est importante car, elle permet de différencier la lecture de la table selon qu'il s'agit d'un test bilatéral $(1 - \frac{\alpha}{2})$ ou d'un test unilatéral $(1 - \alpha)$.

Notons aussi que pour déterminer le fractile d'ordre $1 - \frac{\alpha}{2}$ ou $1 - \alpha$ de la loi Student, on peut aussi utiliser la fonction Excel :

$= loi.student.inverse(\alpha\ ; ddl)$ pour le cas d'un test bilatéral et
$= loi.student.inverse(2\alpha\ ; ddl)$ pour le cas d'un test unilatéral

Aussi lorsque l'on veut déterminer la valeur symétrique (opposée) d'un fractile en vue, par exemple, de la détermination d'un intervalle de confiance, etc, on prend juste l'opposé du fractile calculée puisque la loi de Student est une loi symétrique.

Par ailleurs, il faut aussi noter que lorsque n est grand (n>30), on peut approximer la loi de Student par la loi normale. Dès lors, on peut utiliser la table de la loi normale comme décrite précédemment.

Lecture des probabilités connaissant les fractiles

Pour déterminer la probabilité α dans une table de la loi de Student, on se sert uniquement du fractile et du nombre de degré de liberté en prenant le chemin inverse qui conduit à la détermination du fractile. Dans la table de Student, on se place sur la ligne correspondant au nombre de degrés de liberté et on se déplace de gauche vers la droite en essayant d'identifier la valeur la plus proche possible du fractile recherché. Une fois la valeur du fractile identifiée, on retrouve la valeur de la probabilité en lisant dans le libellé de la colonne correspondant en haut de la table.
Cette procédure peut aussi être mise en œuvre en utilisant les fonctions d'Excel spécifiée comme suit :

$$= loi.student(q; ddl; 2)$$

Où q représente le fractile dont on cherche la probabilité ; ddl représente le nombre de degrés de liberté. L'option 2 signifie que le logiciel doit fournir directement la valeur α. En effet, en mettant l'option 1, on obtient $\frac{\alpha}{2}$ qu'il va falloir ensuite multiplier par 2 pour retrouver α.

Utilisation de la table de khi-deux

Lecture des fractiles connaissant les probabilités α

La lecture d'une table de khi-deux se fait de la même manière que la lecture de la table de Student discutée précédemment notamment pour ce qui concerne la recherche la recherche du fractile correspondant à un seuil donné.

Cependant la procédure diffère significativement lorsqu'il s'agit des encadrements car la loi de khi-deux n'est pas une loi symétrique. En effet, à la différence des précédentes lois, la loi de khi-deux n'est pas symétriquement distribuée autour de 0. Par conséquent lorsqu'on veut procéder, par exemple, à un encadrement, on doit d'abord définir la probabilité associée chaque fractile constituant les bornes. Par exemple pour encadrer une valeur U dans la perspective de la détermination d'un intervalle de confiance etc.., on calcule d'abord deux probabilités :

$$p1 = \frac{\alpha}{2}\ et\ \ p2 = \frac{\alpha}{2} + (1 - \alpha).$$

Ensuite, on lit les fractiles correspondant à chaque probabilité (en utilisant les degrés de liberté). Ensuite, on encadre U telle que $q1 < U < q2$ où $q1$ et $q2$ représentent respectivement les fractiles correspondants à $p1$ et $p2$. Cet encadrement se fait donc de telle sorte que $P(q1 < U < q2) = 1 - \alpha$. Où $1 - \alpha$ est le seuil de confiance. Pour executer cette prodcéure sous excel, on procède comme suit : $= loi.khideux.inverse(p1; ddl)$ et $= loi.khideux.inverse(p2; ddl)$.

Les valeurs obtenues servent donc à construire l'intervalle de confiance.

Lecture des probabilités connaissant les fractiles

Là également, il n'y pas de différence entre la procédure de lecture d'une loi de khi-deux et une loi de Student pour ce qui concerne la recherche d'une probabilité simple. Par conséquent, on peut se référer à la méthode discutée pour la loi de Student.
En revanche lorsqu'il s'agit de déterminer la probabilité lorsque le fractile est fourni sous forme d'encadrement, la procédure est un peu particulière.
En effet, Puisque, nous avons deux bornes, pour obtenir la probabilité correspondant à la fractile inférieure (c'est-à-dire pour obtenir $\frac{\alpha}{2}$), on recherche juste cette valeur dans la colonne où l'on identifié le fractile. Ensuite, on multiplie par 2 pour obtenir α. De la même manière, on peut se servir de la fractile supérieure pour déterminer la probabilité correspondant à $\frac{\alpha}{2} + (1 - \alpha)$ qui permet ensuite d'obtenir α. Toutes ces procédures peuvent être mises en œuvre sous excel, en utilisant l'une des formules suivantes :

$$= loi.khideux(q1; ddl)$$
$$= loi.khideux(q2; ddl$$

L'une ou l'autre de ces deux valeurs obtenues permettra alors de calculer α et par conséquent 1-α .