

UNIVERSITE MOHAMMED V RABAT

**Faculté des Sciences Juridiques,
Economiques et Sociales
Rabat-Agdal**

Département des Sciences Economiques

Filière : Sciences Economiques et Gestion

Semestre 3

Module : Echantillonnage et Estimation

Prof : Adil ELMARHOUM

Année Universitaire 2018 - 2019

THEORIE D'ECHANTILLONNAGE

I. ROLE DE L'ECHANTILLONNAGE

Lorsqu'on souhaite collecter les informations sur une population, deux possibilités s'offrent : La première solution consiste à observer ou interroger tous les éléments de la population, c'est ce qu'on appelle une enquête complète ou enquête exhaustive ou recensement. La seconde solution consiste à observer ou interroger une partie de la population, c'est ce qu'on appelle enquête partielle ou sondage. Les éléments de la population qui sont réellement observés constituent l'échantillon et l'opération qui consiste à choisir ces éléments est appelée échantillonnage.

L'alternative décrite ci-dessus se présente dans beaucoup de situations et le recours à la deuxième solution c'est à dire l'enquête partielle et la pratique la plus courante.

Par rapport à l'enquête complète, l'enquête partielle offre une série d'avantages. Le coût global de l'enquête partielle est en général plus réduit que le coût global d'une enquête complète. L'enquête par sondage est plus rapide que l'enquête complète, surtout lorsque la caractéristique étudiée présente des modifications assez importantes au cours du temps. Les erreurs d'observations sont plus réduites que dans l'enquête exhaustive. En fin dans certaines situations particulières, l'enquête partielle est la seule solution possible, c'est le cas lorsque l'observation présente un caractère destructif.

II. VOCABULAIRE

Enquête : ensemble des opérations de collecte et de traitement de données relatives à quelques domaines que ce soit.

Population : rassemblement de tous les cas qui répondent à un ensemble de caractères spécifiques. Appelée aussi univers ou ensemble statistique, c'est l'ensemble des éléments auxquels on s'intéresse.

Unité de base : unité d'échantillonnage ou unité de sondage, c'est l'élément pris en considération dans l'enquête.

Recensement : Enquête complète ou enquête exhaustive, c'est une enquête au cours de laquelle toutes les unités de base de la population sont observées.

Sondage : Enquête incomplète, enquête partielle ou enquête par échantillonnage, c'est une enquête au cours de laquelle seulement une partie des unités de base de la population sont observées.

Echantillon : ensemble des unités de base sélectionnées et réellement observées au cours d'un sondage.

Echantillonnage : ensemble des opérations qui permettent de sélectionner de façon organisée les éléments de l'échantillon.

Base de sondage : énumération ou présentation ordonnée de toutes les unités de base constituant la population.

Erreur d'échantillonnage : écart entre les résultats obtenus auprès d'un échantillon et ce que nous apprendrait un recensement comparable de la population. Plus la taille de l'échantillon est grande plus l'erreur d'échantillonnage diminue.

Fraction ou taux de sondage : proportion des unités de la population qui font partie de l'échantillon. C'est le rapport entre la taille de l'échantillon n , et la taille de la population N .

$$f = \frac{n}{N} \times 100$$

III. DETERMINATION DE LA TAILLE DE L'ECHANTILLON

Le nombre n n'est pas une garantie absolue de représentativité. La détermination de la taille d'échantillon dépend essentiellement de deux facteurs :

- La précision souhaitée : plus on souhaite des résultats précis, plus l'échantillon nécessaire est important.
- Le budget disponible : plus on augmente la taille, plus le coût de l'enquête s'accroît.

La taille de l'échantillon doit être celle qui permet d'atteindre le meilleur équilibre entre le risque de commettre des erreurs d'échantillonnage, le coût induit par ces erreurs, et le coût de l'échantillonnage lui-même.

Afin de déterminer la taille de l'échantillon, nous utiliserons l'inégalité de Bienaymé Tchebycheff ou la loi normale.

3.1. UTILISATION DE L'INEGALITE DE BIENAYME TCHEBYCHEFF

Cette inégalité n'est utilisée que si la loi de la variable aléatoire est complètement inconnue. Elle aboutit à des échantillons de taille élevée.

3.1.1. Taille d'échantillon pour estimer une moyenne.

- La taille de l'échantillon dépend de la précision souhaitée pour la généralisation des résultats.
- La précision (ou erreur d'échantillonnage) s'exprime en valeur absolue ou relative. Elle représente la largeur de l'intervalle de confiance de la moyenne. Soit ε la moitié de cette largeur.

L'inégalité de Bienaymé Tchebycheff dans le cas de la moyenne s'écrit :

$$P\left(\left|\bar{X}-m\right| < \varepsilon\right) \geq 1-\frac{\sigma^2}{n\varepsilon^2}$$

avec :

n : taille de l'échantillon ;

ε : précision souhaitée ;

\bar{X} : moyenne de l'échantillon ;

m : moyenne de la population.

σ : Ecart- type d'échantillon, il est souvent inconnu, il faut avoir des informations antérieures ou mener une étude pilote.

Pour obtenir un maximum de fiabilité dans les résultats, on commence par se fixer une marge d'erreur " ε " que l'on accepte. On se fixe ensuite un seuil de confiance $(1-\alpha)$, qui représente la probabilité minimale pour que la moyenne calculée à partir de l'échantillon ne s'écarte pas de la moyenne de la population de plus de ε . Ceci s'écrit :

$$P\left(\left|\bar{X}-m\right| < \varepsilon\right) \geq 1-\alpha$$

En rapprochant les deux formules on obtient :

$$1-\frac{\sigma^2}{n\varepsilon^2} = 1-\alpha$$

et donc :

$$n = \frac{\sigma^2}{\varepsilon^2 \times \alpha}$$

Exemple :

Un parc de loisirs souhaite estimer à 10dh près le montant moyen d'achats effectués par chaque visiteur, c'est à dire on se fixe une marge d'erreur de 10 dans l'analyse des résultats :

$$\varepsilon = 10$$

Une étude pilote menée sur 50 visiteurs choisis au hasard a montré que l'écart- type des achats est : $\sigma = 100$ dh.

Si on se fixe un seuil de confiance $(1-\alpha) = 95\%$, La taille de l'échantillon est donc :

$$n = \frac{100^2}{10^2 \times 0,05} = 2000$$

3.1.2. Taille d'échantillon pour estimer une proportion

- La taille de l'échantillon dépend de la précision souhaitée pour la généralisation des résultats.
- La précision (ou erreur d'échantillonnage) s'exprime en valeur absolue ou relative. Elle représente la largeur de l'intervalle de confiance de la proportion. Soit ε la moitié de cette largeur.

l'inégalité de Bienaymé Tchebycheff dans le cas de la proportion s'écrit :

$$P(|f_n - p| < \varepsilon) \geq 1 - \frac{pq}{n\varepsilon^2}$$

avec :

n : taille de l'échantillon ;

ε : précision souhaitée ;

f_n : proportion ou fréquence relative dans l'échantillon ;

p : proportion dans la population ($q = 1 - p$). Elle est souvent inconnue, il faut avoir des informations antérieures ou mener une étude pilote, sinon on utilise une proportion de 50 %.

Pour obtenir un maximum de fiabilité dans les résultats, on commence par se fixer une marge d'erreur " ε " que l'on accepte. On se fixe ensuite un seuil de confiance $(1-\alpha)$, qui représente la probabilité minimale pour que la fréquence calculée à partir de l'échantillon ne s'écarte pas de la proportion dans la population de plus de ε . Ceci s'écrit :

$$P(|f_n - p| < \varepsilon) \geq 1 - \alpha$$

En rapprochant les deux formules on obtient :

$$1 - \frac{pq}{n\varepsilon^2} = 1 - \alpha$$

et donc :

$$n = \frac{pq}{\varepsilon^2 \times \alpha}$$

Exemple :

Le parc souhaite estimer la proportion des visiteurs qui font des achats à cinq points près, c'est à dire on se fixe une marge d'erreur de 5% dans l'analyse des résultats :

$$\varepsilon = 0,05$$

L'enquête pilote a estimé cette proportion à 65%, c'est à dire $p = 0,65$

Si on se fixe un seuil de confiance $(1 - \alpha) = 95\%$, la taille de l'échantillon est donc :

$$n = \frac{0,65 \times 0,35}{0,05^2 \times 0,05} = 1820$$

3.2. UTILISATION DE LA LOI NORMALE

On applique cette méthode si la variable suit une loi normale ou si elle peut être approchée par la loi normale.

3.2.1. Taille d'échantillon pour estimer une moyenne

a) Cas des prélèvements dans une population finie avec remise ou dans une population infinie sans remise :

Pour obtenir un maximum de fiabilité dans les résultats, on commence par se fixer une marge d'erreur " ε " que l'on accepte. On se fixe ensuite un seuil de confiance $(1 - \alpha)$, qui représente la probabilité minimale pour que la moyenne calculée à partir de l'échantillon ne s'écarte pas de la moyenne de la population de plus de ε . Ceci s'écrit :

$$P\left(\left|\bar{X} - m\right| < \varepsilon\right) \geq 1 - \alpha$$

avec :

- ε : précision souhaitée ;
- \bar{X} : moyenne de l'échantillon ;
- m : moyenne de la population.

D'après le théorème central limite, la variable aléatoire \bar{X} suit une loi normale dont les paramètres sont :

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

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

L'écart type de la moyenne est donc : $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Déterminer la taille de l'échantillon consiste à résoudre l'équation :

$$P(|\bar{X} - m| < \varepsilon) \geq 1 - \alpha$$

$$P(-\varepsilon < \bar{X} - m < \varepsilon) \geq 1 - \alpha$$

$$P\left(-\frac{\varepsilon}{\frac{\sigma}{\sqrt{n}}} < \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} < \frac{\varepsilon}{\frac{\sigma}{\sqrt{n}}}\right) \geq 1 - \alpha$$

$$P\left(-\frac{\varepsilon\sqrt{n}}{\sigma} < Z < \frac{\varepsilon\sqrt{n}}{\sigma}\right) \geq 1 - \alpha$$

$$\Pi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - \Pi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right) \geq 1 - \alpha$$

$$\Pi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - [1 - \Pi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)] \geq 1 - \alpha$$

$$2\Pi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1 \geq 1 - \alpha$$

$$\Pi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) \geq 1 - \frac{\alpha}{2}$$

On se reporte à la table de distribution de la loi Normale centrée réduite, et on cherche la valeur correspondante à une probabilité égale à $1 - \frac{\alpha}{2}$, cette valeur de z sera désignée par $Z_{1-\frac{\alpha}{2}}$

On a alors :

$$\frac{\varepsilon\sqrt{n}}{\sigma} = Z_{1-\frac{\alpha}{2}}$$

$$n = Z_{1-\frac{\alpha}{2}}^2 \frac{\sigma^2}{\varepsilon^2}$$

Exemple :

Reprenons l'exemple du parc de loisirs qui souhaite estimer à 10dh près le montant moyen d'achats effectués par chaque visiteur, c'est à dire on se fixe une marge d'erreur de 10 dans l'analyse des résultats : $\varepsilon = 10$

Une étude pilote menée sur 50 visiteurs choisis au hasard a montré que l'écart-type des achats est : $\sigma = 100$ dh.

Si on se fixe un seuil de confiance $(1-\alpha) = 95\%$, La taille de l'échantillon est donc :

$$n = 1,96^2 \frac{100^2}{10^2} = 384,16 = 385$$

b) Cas des prélèvements dans une population finie sans remise :

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

$$V(\bar{X}_n) = \frac{N-n}{N-1} \frac{\sigma^2}{n}$$

L'écart type de la moyenne est donc : $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}} \approx \frac{\sigma}{\sqrt{n}} \sqrt{1-\frac{n}{N}}$

De la même manière, on arrive à :

$$\frac{\varepsilon \sqrt{n}}{\sigma} \sqrt{\frac{N}{N-n}} = Z_{1-\frac{\alpha}{2}}$$

$$\sqrt{\frac{n}{N-n}} = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\varepsilon \sqrt{N}}$$

$$\frac{n}{N-n} = Z_{1-\frac{\alpha}{2}}^2 \times \frac{\sigma^2}{\varepsilon^2 N}$$

$$n = Z_{1-\frac{\alpha}{2}}^2 \frac{\sigma^2}{\varepsilon^2} - n Z_{1-\frac{\alpha}{2}}^2 \frac{\sigma^2}{\varepsilon^2 N}$$

$$n(1 + Z_{1-\frac{\alpha}{2}}^2 \frac{\sigma^2}{\varepsilon^2 N}) = Z_{1-\frac{\alpha}{2}}^2 \frac{\sigma^2}{\varepsilon^2}$$

$$n = \frac{Z_{1-\frac{\alpha}{2}}^2 \sigma^2 N}{\varepsilon^2 N + Z_{1-\frac{\alpha}{2}}^2 \sigma^2}$$

3.2.2. Taille d'échantillon pour estimer une proportion.

Pour obtenir un maximum de fiabilité dans les résultats, on commence par se fixer une marge d'erreur " ε " que l'on accepte. On se fixe ensuite un seuil de confiance $(1-\alpha)$, qui représente la probabilité minimale pour que la fréquence calculée à partir de l'échantillon ne s'écarte pas de la proportion dans la population de plus de ε . Ceci s'écrit :

$$P(|f_n - p| < \varepsilon) \geq 1 - \alpha$$

avec :

n : taille de l'échantillon ;

ε : précision souhaitée ;

f_n : proportion ou fréquence relative dans l'échantillon ;

p : proportion dans la population ($q = 1 - p$). Elle est souvent inconnue, il faut avoir des informations antérieures ou mener une étude pilote, sinon on utilise une proportion de 50 %.

D'après le théorème central limite, la variable aléatoire f_n suit une loi normale dont les paramètres sont :

a) Cas des prélèvements dans une population finie avec remise ou dans une population infinie sans remise :

$$E(f_n) = p$$

$$V(f_n) = \frac{pq}{n}$$

L'écart type de la fréquence est donc : $\sigma_{f_n} = \frac{\sqrt{pq}}{\sqrt{n}}$

Déterminer la taille de l'échantillon consiste à résoudre l'équation :

$$P(|f_n - p| < \varepsilon) \geq 1 - \alpha$$

$$P(-\varepsilon < f_n - p < \varepsilon) \geq 1 - \alpha$$

$$P\left(-\frac{\varepsilon}{\frac{\sqrt{pq}}{\sqrt{n}}} < \frac{f_n - p}{\frac{\sqrt{pq}}{\sqrt{n}}} < \frac{\varepsilon}{\frac{\sqrt{pq}}{\sqrt{n}}}\right) \geq 1 - \alpha$$

$$P\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{pq}} < Z < \frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) \geq 1-\alpha$$

$$\Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) - \Pi\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) \geq 1-\alpha$$

$$\Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) - [1 - \Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right)] \geq 1-\alpha$$

$$2\Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) - 1 \geq 1-\alpha$$

$$\Pi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{pq}}\right) \geq 1 - \frac{\alpha}{2}$$

On se reporte à la table de distribution de la loi Normale centrée réduite, et on cherche la valeur correspondante à une probabilité égale à $1 - \frac{\alpha}{2}$, cette valeur de z sera désignée par $Z_{1-\frac{\alpha}{2}}$

On a alors :

$$\frac{\varepsilon\sqrt{n}}{\sqrt{pq}} = Z_{1-\frac{\alpha}{2}}$$

$$n = Z_{1-\frac{\alpha}{2}}^2 \frac{pq}{\varepsilon^2}$$

Exemple :

Reprenons l'exemple du parc de loisirs qui souhaite estimer la proportion des visiteurs qui font des achats à cinq points près, c'est à dire on se fixe une marge d'erreur de 5% dans l'analyse des résultats :

$$\varepsilon = 0,05$$

L'enquête pilote a estimé cette proportion à 65%, c'est à dire $p = 0,65$

Si on se fixe un seuil de confiance $(1-\alpha) = 95\%$, on se reporte à la table de distribution de la loi Normale, et on cherche la valeur correspondante à une probabilité $(1-\alpha/2) = 0,975$, ce qui donne $Z = 1,96$.

La taille de l'échantillon est donc :

$$n = 1,96^2 \frac{0,65 \times 0,35}{0,05^2} = 349,58 = 350$$

b) Cas des prélèvements dans une population finie sans remise :

$$E(f_n) = p$$

$$V(f_n) = \frac{N-n}{N-1} \frac{pq}{n}$$

L'écart type de la fréquence est donc : $\sigma_{f_n} = \frac{\sqrt{pq}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \approx \frac{\sqrt{pq}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$

De la même manière, on arrive à :

$$\frac{\varepsilon \sqrt{n}}{\sqrt{pq}} \frac{\sqrt{N}}{\sqrt{N-n}} = Z_{1-\frac{\alpha}{2}}$$

$$\sqrt{\frac{n}{N-n}} = Z_{1-\frac{\alpha}{2}} \frac{\sqrt{pq}}{\varepsilon \sqrt{N}}$$

$$\frac{n}{N-n} = Z_{1-\frac{\alpha}{2}}^2 \frac{pq}{\varepsilon^2 N}$$

$$n = Z_{1-\frac{\alpha}{2}}^2 \frac{pq}{\varepsilon^2} - n Z_{1-\frac{\alpha}{2}}^2 \frac{pq}{\varepsilon^2 N}$$

$$n(1 + Z_{1-\frac{\alpha}{2}}^2 \frac{pq}{\varepsilon^2 N}) = Z_{1-\frac{\alpha}{2}}^2 \frac{pq}{\varepsilon^2}$$

$$n = \frac{Z_{1-\frac{\alpha}{2}}^2 pq N}{\varepsilon^2 N + Z_{1-\frac{\alpha}{2}}^2 pq}$$

IV. METHODES D'ECHANTILLONNAGE

Pour que les résultats d'une enquête par sondage puissent être extrapolés à l'ensemble de la population faisant l'objet de l'étude, il est indispensable que cette enquête soit conduite selon des règles bien définies et que les calculs conduisant à ces extrapolations soient conformes à la procédure d'échantillonnage utilisée.

L'échantillon choisi doit être le plus représentatif possible de la population étudiée, c'est à dire le degré de correspondance entre l'information recueillie et ce que nous apprendrait un recensement comparable de la population dépend en grande partie de la façon dont l'échantillon a été choisi.

La théorie moderne de l'échantillonnage nous propose une distinction fondamentale entre échantillons basés sur la probabilité : échantillons probabilistes; et échantillons non basés sur la probabilité : échantillons non probabilistes ou empiriques.

4.1. METHODES D'ECHANTILLONNAGE PROBABILISTES

4.1.1. Echantillonnage aléatoire et simple

Un échantillonnage est aléatoire si tous les individus de la population ont la même chance de faire partie de l'échantillon; il est simple si les prélèvements des individus sont réalisés indépendamment les uns des autres.

En particulier, si la population est finie, cette définition correspond au tirage aléatoire avec remise, qui permet de traiter les populations finies comme des populations infinies.

Pour prélever un échantillon aléatoire et simple il faut :

- Constituer la base de sondage qui correspond à la liste complète et sans répétition des éléments de la population ;
- Numéroté ces éléments de 1 à N ;
- Procéder, à l'aide d'une table de nombres aléatoires ou d'un générateur de nombres pseudo aléatoires à la sélection des unités différentes qui constitueront l'échantillon.

Exemple :

On souhaite avoir un échantillon aléatoire et simple de 5 entreprises parmi une population de 22 entreprises. On dispose de la base de sondage c'est à dire la liste complète et sans répétitions des 22 entreprises numérotées de 1 à 22. On prend un extrait d'une table de nombre aléatoire par exemple :

10480	15011	01536	02011	81647	91646
22368	46573	25595	85393	30995	89198
24130	48390	22527	97265	76393	64809
42167	93093	06243	61680	07856	16376
37570	39975	81837	16656	06121	91782
77921	06907	11008	42751	27756	53498

On choisit au hasard un nombre de la table, supposons ce nombre 06121. Comme $N=22$, on va retenir le premier groupe de 2 chiffres, ce qui donne les N° : 06, ensuite 12 ; 19 ; 17 ; les nombres (82,77 et 92) sont inutilisables. La cinquième entreprise sera le N° 10.

4.1.2. Echantillonnage stratifié

L'échantillonnage stratifié est une technique qui consiste à subdiviser une population hétérogène, d'effectif N , en P sous populations ou « strates » plus homogènes d'effectif N_i de telle sorte que $N = N_1 + N_2 + \dots + N_p$. Un échantillon, d'effectif n_i , est par la suite, prélevé indépendamment au sein de chacune des strates en appliquant un plan d'échantillonnage au choix de l'utilisateur. Le plus souvent, on procédera par un échantillonnage aléatoire et simple à l'intérieur de chaque strate.

La stratification peut entraîner des gains de précision appréciables, elle facilite en outre les opérations de collecte des données et fournit des informations pour différentes parties de la population.

Pour la répartition de l'effectif total n de l'échantillon dans les différentes strates, on utilise une répartition proportionnelle, elle consiste à conserver la même fraction d'échantillonnage dans chaque strate ou à tenir compte du poids de chaque strate.

Désignons par w_i le poids de la strate et par f la fraction de sondage constante.

$$f = \frac{n}{N} \qquad w_i = \frac{N_i}{N}$$

le nombre d'unités à choisir dans chacune des strates est donc :

$$n_i = w_i \times n = f \times N_i$$

Exemple :

Dans une population de 10000 entreprises, réparties en 500 petites entreprises, 3000 moyennes entreprises et 2000 grandes entreprises, on souhaite avoir un échantillon de 500 entreprises.

Fraction de sondage constante : $f = 500 / 1000 = 0.05 \%$

Strate	Effectif de la strate	Taille de l'échantillon
Petite	5000	$5000 * 0,05 = 250$
Moyenne	3000	$3000 * 0,05 = 150$
Grande	2000	$2000 * 0,05 = 100$
Total	10000	500

4.1.3. Echantillonnage par degrés

L'échantillonnage par degrés regroupe toute une série de plans d'échantillonnage caractérisés par un système ramifié et hiérarchisé d'unités.

Dans le cas de deux degrés, par exemple, on considère que la population est constituée d'un certain nombre d'unités de sondage du premier degré (unités primaires), chacune de ces unités étant constituée d'un certain nombre d'unités du second degré. (unités secondaires)

On réalise d'abord un échantillonnage d'unités du premier degré. Ensuite, dans chaque unité sélectionnée au premier degré, on prélève un échantillon d'unités du second degré. Le mode de sélection pouvant varier d'un degré à l'autre.

L'échantillonnage par degrés s'impose lorsqu'il est impossible d'inventorier les éléments de toute la population et qu'il est possible d'énumérer les unités prélevées au premier degré. Il permet une concentration du travail sur le terrain et donc une réduction des coûts.

Pour un même nombre total d'observations, il faut citer sa plus faible efficacité que l'échantillonnage aléatoire et simple.

Exemple :

Pour étudier le niveau de consommation des ménages d'une ville, on a tiré aléatoirement 5 quartiers. Dans chaque quartier sélectionné, on retient une rue sur 5, dans chaque rue retenue, on retient un immeuble sur 3, et dans chaque immeuble, un ménage par étage sera questionné.

4.1.4. Echantillonnage systématique

L'échantillonnage systématique est une technique qui consiste à prélever des unités d'échantillonnage situées à intervalles égaux. Le choix du premier individu détermine la composition de tout l'échantillon.

Si on connaît l'effectif total de la population N et qu'on souhaite prélever un échantillon d'effectif n , l'intervalle entre deux unités successives à sélectionner est donné par :

$$k = \frac{N}{n} \text{ (arrondi à l'entier le plus proche)}$$

Connaissant k , on choisit le plus souvent, pour débiter, un nombre aléatoire, i , compris entre 1 et k . le rang des unités sélectionnées est alors $i, i+2k, i+3k, \dots$

L'échantillonnage systématique est facile à préparer et, en général facile à exécuter, il réduit le temps consacré à la localisation des unités sélectionnées.

Si les éléments de la population se présentent dans un ordre aléatoire (pas de tendance) l'échantillonnage systématique est équivalent à l'échantillonnage aléatoire et simple. Par contre si les éléments de la population présentent une tendance, l'échantillonnage systématique est plus précis que l'échantillonnage aléatoire.

Exemple :

On veut sélectionner un échantillon de 30 entreprises au sein d'une population de 1800 entreprises.

$$k = \frac{1800}{30} = 60$$

Ainsi on va tirer une entreprise toutes les 60 en partant d'un nombre tiré aléatoirement entre 1 et 60.

Supposons ce nombre est le 15. On va donc sélectionner la 15^{ème} entreprise puis la 75^{ème}, la 135^{ème}, jusqu'à la 1755^{ème} ce qui nous donnera l'échantillon de 30 entreprises.

4.2. METHODES D'ECHANTILLONNAGE EMPIRIQUES

4.2.1 Echantillonnage accidentel (De convenance)

Il s'agit d'un échantillon constitué d'individus qui se trouvaient accidentellement à l'endroit et au moment où l'information a été collectée.

Exemple :

- Enquêtes réalisées dans la rue, les lieux publics, en sortie de super marché ...
- Questionnaires figurant dans les magazines et renvoyés spontanément.

Les échantillons accidentels ne peuvent être considérés représentatifs d'aucune population. Il est risqué de généraliser à une population donnée des résultats obtenus par un échantillon accidentel.

4.2.2. Echantillonnage à priori

C'est un échantillonnage par jugement à priori. Il consiste à sélectionner des individus dont on pense, avant de les interroger, qu'ils peuvent détenir l'information.

Le risque de ce type d'échantillonnage est de considérer des individus, apparemment représentatifs de la population étudiée.

4.2.3. Echantillonnage « Boule de neige »

Cette méthode est réservée aux populations composées d'individus dont l'identification est difficile ou qui possèdent des caractéristiques rares.

La méthode consiste à faire construire l'échantillon par les individus eux-mêmes. Il suffit d'en identifier un petit nombre initial et de leur demander de faire appel à d'autres individus possédant les mêmes caractéristiques.

4.2.4. Echantillonnage par Quotas.

L'échantillonnage par quotas est l'échantillonnage non probabiliste le plus connu, et finalement le mieux accepté comme substitut aux méthodes probabilistes dans le cas où ces dernières rencontreraient des contraintes de base de sondage. Mais la représentativité de la population étudiée reste douteuse.

L'échantillonnage par quotas consiste à étudier la structure de la population selon des critères choisis (quotas) empiriquement. L'échantillon est ensuite construit de manière à constituer une reproduction en miniature de la population sur ces critères.

L'échantillonnage par quotas est une forme simplifiée de l'échantillonnage stratifié à fraction de sondage constante. Les quotas représentent les variables de stratification.

Une fois les quotas sont fixés, les individus sont sélectionnés à la convenance de l'enquêteur.

Les critères servant de base à la définition des quotas ne doivent pas être nombreux. Au-delà de 3 critères, la démarche devient complexe. Les quotas doivent être construits sur une base de données fiable (statistiques disponibles) indiquant la répartition de la population sur les critères choisis. Les critères les plus utilisés dans les études de marché sont économiques et socio-démographiques en particulier l'âge, le sexe, la catégorie socioprofessionnelle, ...

Exemple :

On souhaite avoir un échantillon de 1000 individus. La structure de la population selon trois critères est la suivante :

1) Age

Age	Structure de la population	Répartition de l'échantillon
20 à 29 ans	40 %	400
30 à 49 ans	35 %	350
50 à 60 ans	25 %	250
Total	100 %	1000

2) Sexe x Age

Structure de la population

Age	Sexe	Masculin	Féminin	Total
20 à 29 ans		48 %	52 %	100 %
30 à 49 ans		49 %	51 %	100 %
50 à 60 ans		45 %	55 %	100 %

Répartition de l'échantillon

Age	Sexe	Masculin	Féminin	Total
20 à 29 ans		192	208	400
30 à 49 ans		172	178	350
50 à 60 ans		113	137	250

3) Age x Sexe x Catégorie socioprofessionnelle

Structure de la population

AGE	CSP Sexe	Sans	Etudiant	Agric	Artisans	Prof libérales	Employés	Ouvriers	Total
20-29	M	10%	30%	5%	6%	9%	25%	15%	100%
	F	15%	25%	2%	10%	8%	30%	10%	100%
30-49	M	8%	5%	15%	22%	15%	15%	20%	100%
	F	20%	4%	10%	16%	14%	24%	12%	100%
50-60	M	6%	2%	25%	22%	18%	17%	10%	100%
	F	35%	1%	20%	20%	6%	13%	5%	100%

Répartition de l'échantillon

AGE	CSP Sexe	Sans	Etudiant	Agric	Artisans	Prof libérales	Employés	Ouvriers	Total
20-29	M	19	58	10	12	17	48	28	192
	F	31	52	4	21	17	62	21	208
30-49	M	14	9	26	38	26	26	33	172
	F	36	7	18	28	25	43	21	178
50-60	M	7	2	28	25	20	19	12	113
	F	48	1	27	27	8	18	8	137

V. DISTRIBUTIONS D'ECHANTILLONNAGE

La notion de distribution d'échantillonnage est à la base des méthodes d'inférence statistique dont les deux principales applications sont les problèmes d'estimation et les tests d'hypothèses. Les premiers ont pour but d'estimer, à partir d'un échantillon, la valeur numérique d'un ou de plusieurs paramètres de la population, et de déterminer la précision de cette ou de ces estimations. Les seconds ont pour but de vérifier la véracité d'une hypothèse émise au départ au sujet d'une ou de plusieurs populations.

A tout paramètre de population θ , on peut associer une série infinie de valeurs observées t, t', t'', \dots , calculées à partir d'échantillons successifs de même effectif, prélevés dans des conditions identiques. Ces valeurs peuvent être considérées comme des valeurs observées d'une même variable aléatoire T , et cette variable est fonction des différentes variables aléatoires correspondant à chacun des individus de l'échantillon :

$$T = f(X_1, X_2, \dots, X_n)$$

En supposant que l'échantillon est aléatoire et simple, la variable aléatoire T possède une distribution de probabilité, dite distribution d'échantillonnage. On peut donc calculer l'espérance $E(T)$ et la variance $V(T)$ de cette distribution.

La distribution d'échantillonnage est donc la distribution des différentes valeurs que peut prendre la variable aléatoire T , pour les différents échantillons possibles. Son écart type σ_T est appelé erreur standard.

Les principales distributions d'échantillonnage sont la distribution d'échantillonnage de la moyenne, la distribution d'échantillonnage de la variance et la distribution d'échantillonnage de la proportion.

5.1. DISTRIBUTION D'ECHANTILLONNAGE DE LA MOYENNE

Supposons que dans une population infinie quelconque, on ait prélevé au hasard un premier échantillon de n observations :

$$X_1, X_2, X_3, \dots, X_n$$

et qu'on ait calculé la moyenne :

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

Si on prélève, dans les mêmes conditions, un deuxième échantillon de même effectif :

$$X_1', X_2', X_3', \dots, X_n'$$

La moyenne correspondante $\bar{x}' = \frac{\sum_{i=1}^n x_i'}{n}$ sera généralement différente de la première moyenne observée.

Il en sera de même pour les moyennes d'autres échantillons prélevés dans les mêmes conditions :

$$x_1'', x_2'', x_3'', \dots, x_n''$$

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

On peut considérer la suite des premières observations x_1, x_1', x_1'', \dots des différents échantillons comme des valeurs observées d'une même variable aléatoire X_1 , la suite des deuxièmes observations des différents échantillons comme des valeurs observées d'une même variable aléatoire X_2 , etc.

Les moyennes observées $\bar{x}, \bar{x}', \bar{x}'', \dots$ sont alors des valeurs observées d'une même variable aléatoire \bar{X} qui est fonction de X_1, X_2, \dots, X_n .

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

Comme X_1, X_2, \dots, X_n , la variable aléatoire \bar{X} possède une distribution de probabilité, dite distribution d'échantillonnage de la moyenne. On peut donc calculer l'espérance et la variance de cette distribution, en supposant que l'échantillon est aléatoire et simple, les variables aléatoires X_1, X_2, \dots, X_n ont toutes la même distribution de probabilité, dont la moyenne est désignée par m et la variance par σ^2 .

$$E(X_i) = m \quad \text{et} \quad V(X_i) = \sigma^2$$

On démontre alors :

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

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

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ est appelé erreur standard de la moyenne d'un échantillon aléatoire est simple

Dans le cas d'une population finie d'effectif N , au sein de laquelle est prélevé, sans remise, un échantillon aléatoire est simple d'effectif n , la variance de la moyenne est :

$$v(\bar{X}) = \frac{N-n}{N-1} \frac{\sigma^2}{n}$$

L'erreur standard est alors : $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

En plus des caractéristiques de la distribution d'échantillonnage de la moyenne, on peut aussi rechercher la forme de cette distribution.

Si par exemple, la population mère possède une distribution normale, on peut affirmer que la distribution de la moyenne est elle-même normale de moyenne m et d'écart type $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Si la distribution de la population parent est inconnue, le théorème central limite permet d'affirmer que la distribution de la moyenne est asymptotiquement normale. Pour un effectif suffisamment élevé, la moyenne d'un échantillon peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque l'effectif est supérieur à 30. Dans le cas contraire ($n < 30$), la moyenne d'un échantillon peut toujours être considérée comme une variable de Student à $(n-1)$ degré de liberté.

5.2. DISTRIBUTION D'ECHANTILLONNAGE DE LA VARIANCE

De la même manière que la moyenne, chacun des échantillons possède une variance :

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

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

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

Ces variances peuvent être considérées comme des valeurs observées d'une même variable aléatoire :

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

Comme X_1, X_2, \dots, X_n , la variable aléatoire $V(X)$ possède une distribution de probabilité, dite distribution d'échantillonnage de la variance. On peut donc calculer l'espérance mathématique et la variance de cette distribution, en supposant que l'échantillon est aléatoire et simple, les variables aléatoires X_1, X_2, \dots, X_n ont toutes la même distribution de probabilité, dont la moyenne est désignée par m et la variance par σ^2 .

$$E(X_i) = m \quad \text{et} \quad V(X_i) = \sigma^2$$

on peut démontrer alors :

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

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

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

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

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

$$E(V(X)) = E(\sigma^2 - (\bar{X} - m)^2)$$

$$E(V(X)) = E(\sigma^2) - E((\bar{X} - m)^2)$$

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

$$E(V(X)) = \frac{n-1}{n} \times \sigma^2$$

Pour la variance de la distribution d'échantillonnage de la variance, on démontre, dans le cas d'une population normale :

$$V(V(X)) = E[(V(X) - E(V(X)))^2] = E[(V(X) - \frac{n-1}{n}\sigma^2)^2] = \frac{2(n-1)}{n^2} \sigma^4.$$

Dans le cas d'une population finie d'effectif N, au sein de laquelle est prélevé, sans remise, un échantillon aléatoire est simple d'effectif n, l'espérance mathématique de la variance est :

$$E(V(X)) = \frac{N}{N-1} \times \frac{n-1}{n} \times \sigma^2$$

En ce qui concerne la forme de la distribution d'échantillonnage de la variance, on peut

démontrer que dans le cas particulier d'une population normale, la variable aléatoire $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2}$ possède une distribution khi deux à (n-1) degré de liberté.

5.3. DISTRIBUTION D'ECHANTILLONNAGE DE LA PROPORTION

Si on considère une population infinie et si on y prélève un échantillon aléatoire et simple d'effectif n, on désigne par x le nombre d'individus possédant, dans l'échantillon, le caractère étudié.

$f_n = \frac{X_n}{n}$ est la fréquence ou proportion des individus possédant, dans l'échantillon, le caractère étudié.

On désigne par p la proportion des individus possédant, dans la population, le caractère étudié.

De la même manière que la moyenne et la variance, chacun des échantillons possède une fréquence :

$$f_n = \frac{X_n}{n}$$

$$f_n' = \frac{X_n'}{n}$$

$$f_n'' = \frac{X_n''}{n}$$

Ces fréquences peuvent être considérées comme des valeurs observées d'une même variable aléatoire :

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

La variable aléatoire F_n possède une distribution de probabilité, dite distribution d'échantillonnage de la proportion. On peut donc calculer l'espérance et la variance de cette distribution, en supposant que l'échantillon est aléatoire et simple.

On peut démontrer alors :

$$E(F_n) = E\left(\frac{X_n}{n}\right) = \frac{1}{n} E(X_n) = \frac{1}{n} n p = p$$

$$V(F_n) = V\left(\frac{X_n}{n}\right) = \frac{1}{n^2} V(X_n) = \frac{1}{n^2} n p q = \frac{pq}{n}$$

$\sigma_{F_n} = \frac{\sqrt{pq}}{\sqrt{n}}$ est appelé erreur standard de la fréquence d'un échantillon aléatoire est simple

Dans le cas d'une population finie d'effectif N , au sein de laquelle est prélevé, sans remise, un échantillon aléatoire est simple d'effectif n , la variance de la fréquence est :

$$V(F_n) = V\left(\frac{X_n}{n}\right) = \frac{1}{n^2} V(X_n) = \frac{1}{n^2} \frac{N-n}{N-1} n p q = \frac{N-n}{N-1} \frac{pq}{n}$$

L'erreur standard est alors : $\sigma_{F_n} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sqrt{pq}}{\sqrt{n}}$

En ce qui concerne la forme de cette distribution, on peut affirmer que la distribution de la proportion suit une loi normale de moyenne p et d'écart type $\sigma_{F_n} = \frac{\sqrt{pq}}{\sqrt{n}}$ à condition que la taille de l'échantillon soit supérieure ou égale à 30 ($n \geq 30$) et le produit $n p \geq 5$.

EXERCICES SUR LA THEORIE D'ECHANTILLONNAGE

1. Quelle est l'espérance mathématique et quelle est la variance des résultats qu'on peut obtenir quand on choisit au hasard et indépendamment dix nombres entiers de 1 à 9 et qu'on en calcule la moyenne, en supposant que chacun des nombres de 1 à 9 a une même probabilité d'être choisi et qu'un même nombre peut être choisi plusieurs fois sans aucune restriction ?
2. Quelle est la probabilité que la moyenne d'un échantillon de 12 observations provenant d'une population de distribution uniforme définie dans l'intervalle (0, 1) soit comprise entre 0,4 et 0,6?
3. Calculez la moyenne et l'écart type de la variance S^2 , ainsi que la probabilité $P(10 < S^2 < 20)$, en supposant que S^2 désigne la variance observée d'échantillons aléatoires et simples d'effectif 10 extraits d'une population normale de moyenne égale à 15 et écart type égal à 4.
4. On suppose que les poids de 3000 étudiants d'une université suivent une loi normale de moyenne 68,0 kilogrammes et écart type 3,0 kilogrammes. Si l'on extrait 80 échantillons de 25 étudiants chacun, quelle est la moyenne et écart type théoriques de la distribution d'échantillonnage des moyennes pour (a) un échantillonnage non exhaustif, (b) un échantillonnage exhaustif ?
5. 500 pignons ont un poids moyen de 5,02 grammes et un écart type de 0,30 grammes. Trouver la probabilité pour qu'un échantillon de 100 pignons choisi au hasard ait un poids total (a) compris entre 496 et 500 grammes. (b) plus grand que 510 grammes.
6. Chacune des personnes d'un groupe de 500 individus lance 120 fois une pièce de monnaie parfaite. Combien de personnes signaleront-elles que (a) le nombre de faces qu'elles obtiennent se trouve compris entre 40 et 60. (b) 5 sur 8 ou plus de leurs jets correspondent à des faces ?
7. Les ampoules électriques d'un fabricant A ont une durée de vie moyenne de 1400 heures avec un écart-type de 200 heures, et celles d'un fabricant B ont une durée de vie moyenne de 1200 heures avec un écart-type de 100 heures. Si l'on teste des échantillons de 125 ampoules pour chaque marque, quelle est la probabilité pour que la marque d'ampoules A ait une durée de vie moyenne qui soit au moins supérieure de (a) 160 heures, (b) 250 heures à celle de la marque d'ampoules B ?
8. Un certain type d'ampoule électrique a une durée de vie moyenne de 1500 heures et un écart type de 150 heures. Trois ampoules sont branchées de telle manière que, si l'une d'elles est grillée, les autres continuent à fonctionner. En supposant que les durées de vie suivent une loi de Laplace Gauss, quelle est la probabilité pour que l'éclairage fonctionne (a) au moins pendant 5000 heures. (b) au plus pendant 4200 heures ?
9. Les poids des colis reçus dans un grand magasin ont une moyenne de 300 kg et un écart-type de 50 kg, Quelle est la probabilité pour que 25 colis reçus au hasard et chargés sur un monte-charge dépassent la limite de sécurité du monte-charge, qui est 8200 kilogrammes.

10. Trouver la probabilité pour que parmi les 200 prochains enfants à naître (a) il y ait moins de 40 % de garçons, (b) il y ait entre 43 % et 57 % de filles, (c) il y ait plus de 54 % de garçons. On suppose que la naissance d'un garçon et la naissance d'une fille sont équiprobables.
11. Un fabricant expédie 1000 lots de 100 ampoules électriques chacun. Si 5 % des ampoules sont normalement défectueuses, dans combien de lots peut-on avoir (a) moins de 90 bonnes ampoules, (b) 98 bonnes ampoules ou davantage ?
12. Dans le but d'étudier l'intention d'achat d'un produit, on décide de réaliser un sondage. Combien de personnes doit-on interroger pour que la fréquence empirique ne s'éloigne pas de la vraie proportion de 1% et ce avec une probabilité au moins égale à 95% ?
13. Une enquête sur l'emploi a pour but d'estimer le taux d'activité dans un pays. Dans les statistiques disponibles, la population active du pays est estimée à 10000000 personnes sur une population totale de 40 millions de personnes. Déterminer la taille de l'échantillon si l'on accepte une erreur de 1% . avec une probabilité de 0,95.
14. Afin d'estimer le revenu mensuel moyen dans un secteur de production. Quelle doit être la taille de l'échantillon de salariés à interroger pour que la moyenne empirique ne s'éloigne pas de la moyenne de la population de 100 dh avec une probabilité au moins égale à 0,95 sachant que l'écart type est de 500 dh par salarié ?
15. On souhaite réaliser une enquête sur la consommation des ménages afin d'estimer la dépense moyenne par ménage. Quel doit être la taille de l'échantillon de ménages si la population est composée de 5 millions de ménages et que l'erreur admise ne doit pas dépasser 100 dh avec une probabilité de 0,99 ? l'écart type de la dépense des ménages est de 2000 dh.
16. On souhaite réaliser une enquête sur l'emploi afin d'estimer le taux de chômage. La population active est de 5 millions de personnes. Quelle doit être la taille de l'échantillon pour que la fréquence empirique ne s'éloigne pas du vrai taux de chômage et ce avec une probabilité de 0,95 de 2%. Une enquête récente avait donné un taux de chômage de 12 %
17. Dans le cadre d'une étude socio-économique, on s'intéresse aux habitants de 18 unités urbaines, réparties en deux régions. L'enquête devrait comporter 500 interviews. Comme on dispose de 10 enquêteurs et qu'on souhaite que chaque enquêteur n'opère que dans une seule unité urbaine, on souhaite se limiter à l'étude de 10 unités urbaines. On considère qu'un enquêteur peut réaliser 10 interviews dans la même journée. En fonction de la répartition des unités urbaines par région et de leurs nombres d'habitants, expliquez, de façon aussi détaillée que possible la manière dont on pourrait organiser cette enquête, en précisant notamment dans quelles unités urbaines il y aurait lieu d'envoyer les enquêteurs.

Région 1		Région 2	
Unités urbaines	Nombres d'habitants	Unités urbaines	Nombres d'habitants
1	93600	9	117100
2	45400	10	107100
3	38900	11	61200
4	36500	12	51000
5	35100	13	43800
6	32900	14	38900
7	28100	15	37800
8	26400	16	33500
		17	25800
		18	25300

18. Dans une région regroupant environ 3 millions d'habitants réunis en un peu plus de 1500 communes, on désire réaliser une enquête au cours de laquelle 0,5 pour mille des habitants devraient être interrogés. En effectuant une stratification basée sur la distribution de fréquences donnée ci-dessous, combien d'interviews devrait-on réaliser dans chacune des catégories de communes. Si de plus pour des raisons de facilité, on décidait de ne pas effectuer moins de 10 interviews par commune, dans combien de communes différentes de chacune des catégories les enquêteurs devraient-ils se rendre ?

Nombre d'habitants	Nombre de communes
Moins de 1000	900
1000 – 2000	300
2000 – 5000	200
5000 – 10000	80
10000 – 20000	40
plus de 20000	10
Total	1530

19. Un sondage vise à étudier la notoriété d'une marque. Pour cela on dispose de 12 enquêteurs durant un mois. (a) Sachant que le rendement par jour et par enquêteur est distribué selon une loi normale de moyenne 5, et écart type 1, déterminer la taille de l'échantillon retenue no telle que : $P(n > n_0) = 0,025$. (b) On propose de stratifier la population selon l'âge. Sachant que la population se répartit comme suit, déterminer la répartition de l'échantillon:

Age	moins de 20 ans	entre 20 et 30 ans	entre 30 et 60 ans	plus de 60ans
Effectifs	5500 000	2500 000	1250 000	250 000

20. On s'intéresse au pourcentage de fusibles défectueux dans un lot de 50 sacs contenant chacun 10000 fusibles. Les sacs proviennent de différents fournisseurs qui affirment en général que la proportion de fusibles défectueux ne dépasse pas 1%. L'erreur acceptée sur ce pourcentage est de 0,1% au niveau de confiance 0,95. (a) Déterminer la taille de cet échantillon en utilisant l'IBT, et en supposant la normalité de la variable. Laquelle de ces deux tailles doit-on retenir ? et pourquoi ? (b) Préciser de quel type de sondage s'agit-il : Si on tire n fusibles en prélevant $n/50$ par sac. Si on choisit d'abord K sacs et on tire ensuite n_i fusibles par sac. Si on mélange le contenu des 50 sacs, et on tire n fusibles. (c) Quel est le procédé de tirage, le mieux adapté ?
21. Le budget alloué à une enquête est de 132500 dh. Cette enquête est destinée à estimer le taux de chômage qu'on a estimé a priori égal à 10 %. Les frais de déplacement quotidien sont évalués à 1000 dh par enquêteur. La rémunération d'un enquêteur est de 170 dh par jour. Les charges fixes sont de 20000 dh. (a) Déterminer la taille de l'échantillon si on tolère une erreur de moins de 1 % avec un niveau de confiance de 95% (b) Déterminer la taille maximale permise par le budget allouée si le rendement par enquêteur est de 6 questionnaires par jour. (c) Quel niveau d'erreur faut-il accepter si on réalise l'enquête avec le budget alloué ?
22. Une machine automatique fabrique des entretoises destinées à un montage de roulements. La longueur de ces entretoises doit être comprise, au sens large, entre 37,45 et 37,55 mm. La variable aléatoire X , qui associe à chaque entretoise sa longueur, est une variable gaussienne de moyenne 37,50 mm.
- 1) Quel doit être l'écart type de la variable aléatoire X pour que 998 sur 1000 des pièces fabriquées soient bonnes ?
 - 2) On prélève un échantillon non exhaustif dans la production. Quel doit être l'effectif de cet échantillon pour que la moyenne des longueurs des pièces prélevées appartienne à l'intervalle $[37,495 ; 37,505]$ avec une probabilité de 0,95 ?

LES PROBLEMES D'ESTIMATION

Les premiers problèmes d'inférence statistique auxquels s'applique la théorie des distributions d'échantillonnage sont les problèmes d'estimations. Le but poursuivi est d'estimer, à partir d'un échantillon, la ou les valeurs numériques d'un ou de plusieurs paramètres de la population considérée et de déterminer la précision de cette ou de ces estimations.

On distingue deux formes d'estimations : l'estimation ponctuelle et l'estimation par intervalle de confiance.

I. ESTIMATION PONCTUELLE

L'estimation ponctuelle ou l'estimation de point d'un paramètre est la connaissance de la seule valeur estimée de ce paramètre. Les paramètres les plus recherchés sont la moyenne, la variance et la proportion.

1.1. PRINCIPES GENERAUX DE L'ESTIMATION

Soit une population quelconque, dont la distribution de probabilité $L(X)$ est fonction d'un paramètre θ : $L(X) = f(X, \theta)$ et un échantillon aléatoire et simple d'effectif n extrait de cette population.

On appelle estimateur du paramètre θ , toute fonction aléatoire des valeurs observées, X_1, X_2, \dots, X_n , susceptibles de servir à estimer θ .

$$T_n = f(X_1, X_2, \dots, X_n)$$

On appelle estimations les valeurs numériques t_1, t_2, \dots de cette variable aléatoire T_n .

1.1.1. Les principales qualités d'un estimateur

a) l'absence de biais

La première qualité d'un bon estimateur est l'absence d'erreur systématique ou de biais. Cette qualité implique que la vraie valeur θ doit être retrouvée en moyenne :

$$E(T_n) = \theta$$

Tout estimateur qui satisfait cette condition est dit sans biais ou non biaisé.

b) la variance minimale

Une deuxième qualité d'un bon estimateur est de posséder une précision suffisante. Cette précision peut être mesurée par le moment d'ordre deux par rapport à θ .

$$E[(T_n - \theta)^2]$$

Pour les estimateurs non biaisés, ce moment se confond avec la variance :

$$E[(T_n - \theta)^2] = V(T_n)$$

On peut démontrer qu'à tout paramètre θ correspond une valeur minimum de $E[(T_n - \theta)^2]$.

La fonction qui correspond à ce minimum définit l'estimateur de variance minimum.

Dans le cas des estimateurs non biaisés, cette variance vaut :

$$\frac{1}{nE\left[\left(\frac{d \log f(x, \theta)}{d\theta}\right)^2\right]}$$

Un estimateur non biaisé dont la variance est égale à ce minimum est appelé estimateur non biaisé de variance minimum ou estimateur efficace.

c) convergence en probabilité

un estimateur T_n converge en probabilité vers θ si :

$$\forall \varepsilon > 0, \quad \lim_{n \rightarrow \infty} P(|T_n - \theta| > \varepsilon) = 0$$

Ce ci signifie que l'écart entre le paramètre calculé à partir de l'échantillon et la vraie valeur du paramètre de la population est très faible quand la taille de l'échantillon est grande. Cet écart peut être mesuré par la variance. Ainsi on parle de convergence en probabilité si :

$$\lim_{n \rightarrow \infty} V(T_n) = 0$$

Un estimateur qui converge en probabilité est dit consistant.

1.1.2. la méthode du maximum de vraisemblance

Ayant défini les principales qualités des estimateurs, la méthode du maximum de vraisemblance permet le plus souvent d'obtenir des estimateurs possédant ces qualités. Le principe de cette méthode est de choisir comme estimation de tout paramètre θ la valeur la plus vraisemblable, c'est à dire celle qui a la plus grande probabilité de provoquer l'apparition des valeurs observées dans l'échantillon. Cette probabilité est appelée fonction de vraisemblance. C'est la probabilité ou la densité de probabilité relative aux valeurs observées x_1, x_2, \dots, x_n , exprimée en fonction du paramètre de la population.

Pour un échantillon aléatoire et simple et pour une population définie par un seul paramètre θ , la fonction de vraisemblance est :

$$L(\theta) = p(x_1, x_2, \dots, x_n; \theta) = p(x_1; \theta) p(x_2; \theta) \dots p(x_n; \theta)$$

Ou

$$L(\theta) = f(x_1, x_2, \dots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta)$$

Les estimateurs du maximum de vraisemblance correspondent par définition au maximum de cette fonction. On cherche ce maximum en annulant la dérivée de la fonction par rapport à θ :

$$\frac{dL(\theta)}{d\theta} = 0$$

ou en annulant la dérivée de son logarithme :

$$\frac{d \log L(\theta)}{d\theta} = 0$$

1.2. Estimation de la moyenne

La meilleure estimation de la moyenne m d'une population, qui puisse être déduite d'un échantillon aléatoire et simple, est la moyenne de l'échantillon.

$$\hat{m} = \bar{x}$$

La dispersion des différentes estimations possibles autour de cette moyenne générale, est mesurée par l'erreur standard de la moyenne :

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

Estimateur du maximum de vraisemblance :

Pour une population normale, la densité de probabilité est :

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

la fonction de vraisemblance est :

$$L(m) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-m}{\sigma}\right)^2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_2-m}{\sigma}\right)^2} \dots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_n-m}{\sigma}\right)^2}$$

$$L(m) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i-m}{\sigma}\right)^2}$$

$$\text{Log } L(m) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^n \left(\frac{x_i-m}{\sigma}\right)^2$$

La dérivée de cette fonction par rapport à m est :

$$\frac{d\log L(m)}{dm} = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i-m)$$

l'estimation du maximum de vraisemblance de la moyenne de la population, est telle que :

$$\frac{1}{\sigma^2}\sum_{i=1}^n (x_i-m) = 0$$

$$\sum_{i=1}^n (x_i-m) = \sum_{i=1}^n x_i - n \times m = 0$$

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

On retrouve la moyenne de l'échantillon définie précédemment.

Les qualités de cet estimateur

a) l'absence de biais

$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ est un estimateur non biaisé de la moyenne m de la population puisqu'on a démontré que :

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

En effet, pour l'ensemble des échantillons qui peuvent être rencontrés, on doit retrouver, en moyenne, la vraie valeur de la population.

b) la variance minimale

Pour une population normale, la densité de probabilité est :

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

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

$$\frac{d \log f(x, m)}{dm} = \frac{x-m}{\sigma^2}$$

$$nE\left[\left(\frac{d \log f(x, m)}{dm}\right)^2\right] = nE\left[\left(\frac{x-m}{\sigma^2}\right)^2\right] = \frac{n}{\sigma^4} E[(X-m)^2] = \frac{n}{\sigma^2}$$

le minimum de la variance des estimateurs de la moyenne est donc :

$$\frac{1}{nE\left[\left(\frac{d \log f(x, m)}{dm}\right)^2\right]} = \frac{\sigma^2}{n}$$

Comme cette valeur est aussi la variance de la distribution d'échantillonnage de la moyenne, il

en résulte que la moyenne $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ d'un échantillon aléatoire et simple est un estimateur de variance minimale. Il est donc un estimateur efficace de la moyenne m de la population.

c) convergence en probabilité

$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ est un estimateur consistant de la moyenne m de la population puisqu'on a démontré que :

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

$$\lim_{n \rightarrow \infty} V(\bar{X}_n) = 0$$

la moyenne $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ calculée à partir d'un échantillon de taille n converge en probabilité vers m .

1.3. Estimation de la variance

Estimateur du maximum de vraisemblance :

Pour une population normale, la densité de probabilité est :

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

la fonction de vraisemblance est :

$$L(\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_1-m}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_2-m}{\sigma}\right)^2} \dots \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_n-m}{\sigma}\right)^2}$$

$$L(\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_{i=1}^n (x_i-m)^2}{2\sigma^2}}$$

$$\text{Log } L(\sigma^2) = \frac{-n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2$$

La dérivée de cette fonction par rapport à σ^2 est :

$$\frac{d \log L(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - m)^2$$

l'estimation du maximum de vraisemblance de la variance de la population, est telle que :

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - m)^2 = 0$$

$$-n\sigma^2 + \sum_{i=1}^n (x_i - m)^2 = 0$$

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

On retrouve la variance de l'échantillon $V(X)$.

Les qualités de cet estimateur

$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n}$ est un estimateur biaisé de la variance σ^2 de la population puisqu'on a démontré que :

$$E(V(X)) = \frac{n-1}{n} \times \sigma^2$$

Contrairement à la moyenne, la meilleure estimation de la variance σ^2 d'une population, qui puisse être déduite d'un échantillon aléatoire et simple, n'est pas la variance de l'échantillon $v(x)$. En effet, pour l'ensemble des échantillons qui peuvent être rencontrés, on ne retrouve pas, en moyenne, la vraie valeur de la population, on obtient ainsi, en moyenne, une valeur inférieure à la variance de la population.

le biais est :

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

Ce biais peut être corrigé en multipliant la variance de l'échantillon par le facteur $\frac{n}{n-1}$. On obtient alors l'estimation :

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

dont l'espérance mathématique est bien σ^2 .

$$E(\hat{\sigma}^2) = E\left(\frac{n}{n-1} \times v(x)\right) = \frac{n}{n-1} E(v(x)) = \frac{n}{n-1} \frac{n-1}{n} \times \sigma^2 = \sigma^2$$

$\hat{\sigma}^2 = \frac{n}{n-1} \times v(x)$ est appelée quasi-variance, c'est un estimateur sans biais de la variance σ^2 de la population. La quasi-variance est désignée par $\hat{\sigma}^2_{n-1}$

L'erreur standard de cette estimation est, dans le cas d'une population normale :

$$\sqrt{v(\hat{\sigma}^2)} = \sqrt{v\left(\frac{n}{n-1} \sigma^2\right)} = \frac{n}{n-1} \sqrt{\frac{2(n-1)\sigma^4}{n^2}} = \sigma^2 \sqrt{\frac{2}{n-1}}$$

1.4. Estimation de la proportion

La meilleure estimation de la proportion p d'une population, qui puisse être déduite d'un échantillon aléatoire et simple, est la fréquence de l'échantillon f_n .

$$\hat{p} = f_n$$

La dispersion des différentes estimations possibles autour de cette proportion générale, est mesurée par l'erreur standard de la proportion :

$$\sigma_{f_n} = \sqrt{\frac{f_n(1-f_n)}{n}}$$

Estimateur du maximum de vraisemblance :

Pour un échantillon aléatoire et simple d'effectif n , dont x individus possèdent le caractère étudié, la fonction de vraisemblance est :

$$L(p) = C_n^x p^x (1-p)^{n-x}$$

$$\text{Log } L(p) = \log C_n^x + x \log p + (n-x) \log (1-p)$$

La dérivée de cette fonction par rapport à p est :

$$\frac{d \log L(p)}{dp} = \frac{x}{p} - \frac{n-x}{1-p}$$

l'estimation du maximum de vraisemblance de la variance de la population, est telle que :

$$\frac{x}{p} - \frac{n-x}{1-p} = 0$$

$$(1-p)x - p(n-x) = 0$$

$$x - np = 0$$

$$\hat{p} = \frac{x}{n}$$

La fréquence f_n de l'échantillon est donc un estimateur du maximum de vraisemblance de la proportion de la population.

Les qualités de cet estimateur**a) l'absence de biais**

$F_n = \frac{X_n}{n}$ est un estimateur non biaisé de la proportion p de la population puisqu'on a démontré que :

$$E(F_n) = p$$

En effet, pour l'ensemble des échantillons qui peuvent être rencontrés, on doit retrouver, en moyenne, la vraie valeur de la population.

b) convergence en probabilité

$F_n = \frac{X_n}{n}$ est un estimateur consistant de la proportion p de la population puisqu'on a démontré que :

$$V(f_n) = \frac{pq}{n}$$

$$\lim_{n \rightarrow \infty} V(f_n) = 0$$

la fréquence relative $F_n = \frac{X_n}{n}$ calculée à partir d'un échantillon de taille n converge en probabilité vers p .

II. ESTIMATION PAR INTERVALLE DE CONFIANCE

L'estimation par intervalle de confiance consiste à déterminer autour de la valeur estimée un intervalle dont on a de fortes chances de croire qu'il contient la vraie valeur du paramètre recherché.

Si on s'intéresse à un paramètre θ , dont on possède un estimateur T , l'estimation par intervalle de confiance consiste à déterminer de part et d'autre de T les bornes $T1$ et $T2$ d'un intervalle qui a une forte probabilité de contenir θ . Cette probabilité est appelée niveau de confiance et désignée par $(1-\alpha)$. α est alors un risque d'erreur.

Les limites $T1$ et $T2$ sont telles que :

$$p(T1 \leq \theta \leq T2) = 1 - \alpha$$

L'intervalle $[T1, T2]$ est appelé intervalle de confiance.

La probabilité que le paramètre θ se trouve à l'extérieur de cet intervalle est donc :

$$p(\theta < T1) + p(\theta > T2) = \alpha$$

Le risque total α peut être réparti d'une infinité de manière. Généralement, on divise le risque α en deux parties égales, Les limites $T1$ et $T2$ sont telles que :

$$p(\theta < T1) = p(\theta > T2) = \alpha/2$$

2.1. Intervalle de confiance de la moyenne

2.1.1. cas d'une population normale

Si on s'intéresse à la moyenne inconnue m d'une population normale d'écart type connu σ , l'estimation par intervalle de confiance consiste à déterminer de part et d'autre de l'estimateur \bar{X} les bornes \bar{X}_1 et \bar{X}_2 d'un intervalle qui a un niveau de confiance $(1-\alpha)$ de contenir m .

Les limites \bar{X}_1 et \bar{X}_2 sont telles que :

$$p(\bar{X}_1 \leq m \leq \bar{X}_2) = 1 - \alpha$$

ou d'une autre façon :

$$p(m < \bar{X}_1) = p(m > \bar{X}_2) = \alpha/2$$

les limites de confiance peuvent être écrites :

$$\bar{X}_1 = \bar{X} - d1 \quad \text{et} \quad \bar{X}_2 = \bar{X} + d2$$

on peut alors écrire :

$$p(m < \bar{X} - d1) = p(m > \bar{X} + d2) = \alpha/2$$

$$p(\bar{X} - m > d1) = p(m - \bar{X} > d2) = \alpha/2$$

Comme, pour une population normale, la variable \bar{X} est elle-même normale de moyenne m et d'écart type $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, on peut écrire :

$$p\left(\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} > \frac{d1}{\frac{\sigma}{\sqrt{n}}}\right) = p\left(\frac{m - \bar{X}}{\frac{\sigma}{\sqrt{n}}} > \frac{d2}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{\alpha}{2}$$

$$p(Z1 > \frac{d1}{\frac{\sigma}{\sqrt{n}}}) = p(Z2 > \frac{d2}{\frac{\sigma}{\sqrt{n}}}) = \frac{\alpha}{2}$$

$$p(Z1 < \frac{d1}{\frac{\sigma}{\sqrt{n}}}) = p(Z2 < \frac{d2}{\frac{\sigma}{\sqrt{n}}}) = 1 - \frac{\alpha}{2}$$

$$\Pi\left(\frac{d1}{\frac{\sigma}{\sqrt{n}}}\right) = \Pi\left(\frac{d2}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \frac{\alpha}{2}$$

Si on désigne par $Z_{1-\frac{\alpha}{2}}$ la valeur de la variable normale réduite lue dans la table :

$$\frac{d1}{\frac{\sigma}{\sqrt{n}}} = \frac{d2}{\frac{\sigma}{\sqrt{n}}} = Z_{1-\frac{\alpha}{2}}$$

il en résulte :

$$d1 = d2 = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Les limites de confiances sont donc :

$$\bar{X}_1 = \bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \text{et} \quad \bar{X}_2 = \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

On notera l'intervalle de confiance :

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

C'est un intervalle symétrique par rapport à la moyenne.

2.1.2. cas d'une population de distribution inconnue

Pour une population de distribution de probabilité inconnue (écart type σ inconnu), on utilise la quasi-variance comme estimation de la variance de la population. L'intervalle de confiance de la moyenne sera défini selon les cas.

Cas d'un échantillon d'effectif inférieur à 30 ($n < 30$) :

Dans ce cas, la moyenne d'un échantillon peut toujours être considérée comme une variable T de Student à $(n-1)$ degré de liberté. La valeur $Z_{1-\frac{\alpha}{2}}$ sera remplacée par la valeur $T_{1-\frac{\alpha}{2}}$ à $(n-1)$

degré de liberté. L'intervalle de confiance est alors :

$$\bar{X} \pm T_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

Cas d'un échantillon d'effectif supérieur ou égal à 30 ($n \geq 30$) :

Dans ce cas, la moyenne d'un échantillon peut toujours être considérée comme une variable approximativement normale. L'intervalle de confiance est alors :

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

2.2. Intervalle de confiance de la variance

Si on s'intéresse à la variance σ^2 d'une population normale, l'estimation par intervalle de confiance consiste à déterminer les bornes σ^2_1 et σ^2_2 d'un intervalle qui a un niveau de confiance $(1-\alpha)$ de contenir σ^2 .

Les limites σ^2_1 et σ^2_2 sont telles que :

$$p(\sigma^2_1 \leq \sigma^2 \leq \sigma^2_2) = 1 - \alpha$$

Comme, pour une population normale, la variable aléatoire $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2}$ possède une distribution khi deux à $(n-1)$ degré de liberté, on peut alors écrire :

$$p\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2_2} \leq \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} \leq \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2_1}\right) = 1 - \alpha$$

ou encore :

$$p\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} < \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2_2}\right) = p\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} > \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2_1}\right) = \alpha/2$$

$$p\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} \leq \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2_1}\right) = 1 - \alpha/2 \quad \Rightarrow \quad \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2_1} = \chi^2_{1-\frac{\alpha}{2}}$$

$$p\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} < \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2_2}\right) = \alpha/2 \quad \Rightarrow \quad \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2_2} = \chi^2_{\frac{\alpha}{2}}$$

Les limites de confiances sont alors :

$$\sigma^2_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad \text{et} \quad \sigma^2_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\chi^2_{\frac{\alpha}{2}}}$$

Les valeurs de $\chi^2_{\frac{\alpha}{2}}$ et $\chi^2_{1-\frac{\alpha}{2}}$ sont à (n-1) degré de liberté.

2.3. Intervalle de confiance de la proportion

Si on s'intéresse à la proportion p, l'estimation par intervalle de confiance consiste à déterminer de part et d'autre de l'estimateur f_n les bornes p_1 et p_2 d'un intervalle qui a un niveau de confiance $(1-\alpha)$ de contenir p.

Les limites p_1 et p_2 sont telles que :

$$p(p_1 \leq p \leq p_2) = 1 - \alpha$$

ou d'une autre façon :

$$p(p < p_1) = p(p > p_2) = \alpha/2$$

les limites de confiance peuvent être écrites :

$$p_1 = f_n - d_1 \quad \text{et} \quad p_2 = f_n + d_2$$

on peut alors écrire :

$$p(p < f_n - d_1) = p(p > f_n + d_2) = \alpha/2$$

$$p(f_n - p > d_1) = p(p - f_n > d_2) = \alpha/2$$

Comme, la distribution de la proportion suit une loi normale de moyenne p et d'écart type

$\sigma_{f_n} = \frac{\sqrt{pq}}{\sqrt{n}}$ à condition que la taille de l'échantillon soit supérieure ou égale à 30 ($n \geq 30$) et le produit $np \geq 5$, on peut écrire :

$$p\left(\frac{f_n - p}{\sqrt{p(1-p)}} > \frac{d1}{\sqrt{n}}\right) = p\left(\frac{p - f_n}{\sqrt{p(1-p)}} > \frac{d2}{\sqrt{n}}\right) = \frac{\alpha}{2}$$

$$p\left(Z1 > \frac{d1}{\sqrt{p(1-p)}}\right) = p\left(Z2 > \frac{d2}{\sqrt{p(1-p)}}\right) = \frac{\alpha}{2}$$

$$p\left(Z1 < \frac{d1}{\sqrt{p(1-p)}}\right) = p\left(Z2 < \frac{d2}{\sqrt{p(1-p)}}\right) = 1 - \frac{\alpha}{2}$$

$$\Pi\left(\frac{d1}{\sqrt{p(1-p)}}\right) = \Pi\left(\frac{d2}{\sqrt{p(1-p)}}\right) = 1 - \frac{\alpha}{2}$$

Si on désigne par $Z_{1-\frac{\alpha}{2}}$ la valeur de la variable normale réduite lue dans la table :

$$\frac{d1}{\sqrt{p(1-p)}} = \frac{d2}{\sqrt{p(1-p)}} = Z_{1-\frac{\alpha}{2}}$$

il en résulte :

$$d1 = d2 = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

Les limites de confiances sont donc :

$$p1 = f_n - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \quad \text{et} \quad p2 = f_n + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

On notera l'intervalle de confiance :

$$f_n \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

La proportion p de la population sera estimée par la fréquence f_n de l'échantillon. On obtient ainsi un intervalle symétrique par rapport à la proportion.

Exemple 1 : intervalle de confiance de la moyenne et de l'écart type

Dans une entreprise produisant un article déterminé on veut estimer sa durée de vie en heures. À cette fin on a observé un échantillon aléatoire et simple de 16 unités dont les résultats sont (en 1000 heures) :

1,10	1,05	1,25	1,08	1,35	1,15	1,30	1,25
1,30	1,35	1,15	1,32	1,05	1,25	1,10	1,15

L'estimation ponctuelle de la moyenne de la population est :

$$\hat{m} = \bar{x} = \frac{\sum_{i=1}^{16} x_i}{16} = 1,2$$

L'estimation ponctuelle de l'écart type de la population de la population est :

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{16} (x_i - \bar{x})^2}{16-1}} = 0,11$$

L'intervalle de confiance de la moyenne à un niveau de confiance de 95 % ($\alpha=5\%$):

La distribution de la population parent étant inconnue et la taille de l'échantillon inférieure à 30, l'intervalle de confiance de la moyenne est défini par :

$$\bar{X} \pm T_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

La valeur de $T_{1-\frac{\alpha}{2}}$ à 15 degrés de liberté est : $t_{0,975} = 2,131$

l'intervalle de confiance est :

$$\bar{X} \pm T_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} = 1,2 \pm 2,131 \frac{0,11}{\sqrt{16}}$$

$$\bar{X}_1 = 1,2 - 2,131 \frac{0,11}{\sqrt{16}} = 1,14 \quad \text{et} \quad \bar{X}_2 = 1,2 + 2,131 \frac{0,11}{\sqrt{16}} = 1,26$$

L'intervalle [1,14 ; 1,26] a une probabilité de 95 % de contenir la vraie valeur de la moyenne de la population.

L'intervalle de confiance de l'écart type à un niveau de confiance de 95 % ($\alpha=5\%$):

Les limites de confiances de la variance sont :

$$\sigma^2_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad \text{et} \quad \sigma^2_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\chi^2_{\frac{\alpha}{2}}}$$

les valeurs de $\chi^2_{\frac{\alpha}{2}}$ et $\chi^2_{1-\frac{\alpha}{2}}$ sont à 15 degrés de liberté :

$$\chi^2_{0,025} = 6,26 \quad \text{et} \quad \chi^2_{0,975} = 27,49$$

L'écart type est la racine carrée de la variance, ses limites de confiance sont donc :

$$\hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{16} (x_i - \bar{x})^2}{\chi^2_{1-\frac{\alpha}{2}}} = \sqrt{\frac{0,11^2 \times 15}{27,49}} = 0,08 \quad \hat{\sigma}_2 = \sqrt{\frac{\sum_{i=1}^{16} (x_i - \bar{x})^2}{\chi^2_{\frac{\alpha}{2}}} = \sqrt{\frac{0,11^2 \times 15}{6,26}} = 0,17$$

Exemple 2 : intervalle de confiance de la proportion

On étudie le pourcentage d'utilisation d'une machine. 400 observations ont été effectuées qui ont donné le résultat suivant :

- Machine marche : 320 observations.
- Machine arrêtée : 80 observations.

L'estimation ponctuelle de la proportion d'utilisation de la machine est :

$$\hat{p} = f_n = \frac{320}{400} = 0,8$$

Le taux d'utilisation de la machine est estimé à 80 %.

L'intervalle de confiance de la proportion à un niveau de confiance de 95 % est défini par :

$$f_n \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

La valeur de $Z_{1-\frac{\alpha}{2}}$ est : $Z_{0,975} = 1,96$

Les limites de confiances de la proportion sont :

$$p_1 = f_n - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} = 0,80 - 1,96 \sqrt{\frac{0,8(1-0,8)}{400}} = 0,76$$

$$p_2 = f_n + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} = 0,80 + 1,96 \sqrt{\frac{0,8(1-0,8)}{400}} = 0,84$$

L'intervalle [76 % ; 84 %] a une probabilité de 95% de contenir le vrai taux d'utilisation de la machine.

EXERCICES SUR LES PROBLEMES DE L'ESTIMATION

Ex 1 : Soit X une variable de Poisson de paramètre (inconnu) m et (X_1, \dots, X_n) les observations d'un échantillon de taille n . Écrire la fonction du maximum de vraisemblance associée à la moyenne. Quel est l'estimateur du maximum de vraisemblance de la moyenne de la population ? Cet estimateur précédent est-il un estimateur efficace ?

Ex 2 : Soit X une variable aléatoire dont la densité de probabilité f est ainsi définie:

$$f(x, \lambda) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right) \quad \text{si } x > 0$$

$$f(x, \lambda) = 0 \quad \text{si } x < 0$$

Où λ est le paramètre (positif) de la loi.

- a) Calculer l'espérance mathématique et la variance de X .
- b) Pour estimer le paramètre λ , on considère un échantillon aléatoire de taille n . Quel est l'estimateur du maximum de vraisemblance de λ ?
- c) L'estimateur de λ est-il un estimateur efficace ?

Ex 3 : Le tableau suivant donne la distribution du nombre de pannes observées dans le fonctionnement d'une machine au cours de 100 journées de travail. Déduisez-en une estimation du nombre moyen de pannes par jour, en supposant que la distribution théorique du nombre de pannes est une loi de poisson. Donner l'erreur standard du résultat obtenu.

Nombres de pannes	Nombres de jours
0	53
1	32
2	11
3	3
4	1
Total	100

Ex 4 : lors d'un concours radiophonique, on note X le nombre de réponses reçues chaque jour. On suppose que X suit une loi normale de paramètres m et σ . Durant les 10 premiers jours, on a obtenu : $x_1 = 200$; $x_2 = 240$; $x_3 = 190$; $x_4 = 150$; $x_5 = 220$; $x_6 = 180$; $x_7 = 170$; $x_8 = 230$; $x_9 = 210$ et $x_{10} = 210$. Déterminer une estimation ponctuelle de m et σ .

Ex 5 : Un échantillon de 15 étudiants d'une faculté a donné les notes suivantes :

13 ; 06 ; 12 ; 10 ; 10 ; 16 ; 02 ; 04 ; 11 ; 12 ; 12 ; 05 ; 07 ; 08 ; 13

- a) Estimer la note moyenne et l'écart type des notes pour l'ensemble des étudiants de la faculté.
- b) Donner des estimations par intervalle de confiance pour la moyenne et l'écart type. ($\alpha=5\%$).

Ex 6 : Dans une entreprise produisant un article déterminé on veut estimer sa durée de vie en heures. À cette fin on a observé un échantillon de 16 unités dont les résultats sont (en 1000 heures) :

1,10 1,05 1,25 1,08 1,35 1,15 1,30 1,25
1,30 1,35 1,15 1,32 1,05 1,25 1,10 1,15

- a) Estimer la durée de vie moyenne et l'écart type d'un article.
- b) Donner des estimations par intervalle de confiance pour la moyenne et l'écart type. ($\alpha=5\%$).

Ex 7 : dans une population d'étudiants en sociologie, on a prélevé, indépendamment, deux échantillons de taille $n_1 = 120$ et $n_2 = 150$. On constate que 48 étudiants de l'échantillon 1 et 66 étudiants de l'échantillon 2 ont une formation secondaire scientifique; Soit p la proportion d'étudiants de la population ayant une formation scientifique ; calculer trois estimations ponctuelles de p .

Ex 8 : dans une station service, on suppose que le montant des chèques essence suit une loi normale de paramètres m et σ . On considère un échantillon de taille $n = 50$ et on obtient une moyenne de 130 Dh et un écart-type de 28 Dh. Donner une estimation de m et σ par un intervalle de confiance au niveau de confiance 95%.

Ex 9 : on donne la répartition des masses de 219 ressorts provenant d'une même fabrication :

masses (g)	[8,2 ; 8,4[[8,4 ; 8,6[[8,6 ; 8,8[[8,8 ; 9[[9 ; 9,2[[9,2 ; 9,4[[9,4 ; 9,6[
Nbre de ressorts	9	21	39	63	45	27	15

X donnant le poids d'un ressort provenant de cette fabrication, donner une estimation de $E(X)$ et $V(X)$. Donner pour $E(X)$ et $V(X)$ un intervalle de confiance au niveau de confiance 95%.

Ex 10 : on veut estimer l'espérance mathématique m d'une variable aléatoire gaussienne X dont on connaît l'écart type $\sigma = 2,3$. Quelle est la taille minimum de l'échantillon de X qui est à prendre si l'on veut obtenir pour m un intervalle de confiance de seuil 0,95 et dont la longueur ne dépasse pas 0,1 ?

Ex 11 : un confiseur vend des boites de bonbons d'un certain modèle. On note X la masse d'une boite pleine. Les pesées de 8 boites ont conduit aux masses (en kg) :

1,22 ; 1,23 ; 1,21 ; 1,19 ; 1,23 ; 1,24 ; 1,18 ; 1,21.

- Donner pour $E(X)$ un intervalle de confiance au risque de 5%.
- En supposant que la variance de X soit connue et égale à la variance observée, donner pour $E(X)$ un intervalle de confiance au seuil de confiance 95% et comparer avec le a).
- On suppose maintenant que l'on a trouvé la même moyenne et la même variance qu'observées mais avec 16 observations au lieu de 8. Reprendre les questions a) et b).

Ex 12 : après avoir pesé 12 pamplemousses d'une même provenance, on donne pour l'espérance mathématique m du poids X d'un pamplemousse, l'intervalle de confiance au niveau de confiance 95% : $390 \text{ g} \leq m \leq 520 \text{ g}$. En déduire la moyenne observée et l'écart type observé.

Ex 13 : Un promoteur désire étudier le nombre de garage qu'il est souhaitable de construire avec un ensemble de logements, afin que les occupants puissent y ranger leur voiture. Pour cela il fait effectuer une enquête par sondage auprès d'un échantillon de ménages susceptibles d'habiter ces appartements.

- On interroge un échantillon de 3238 ménages. On trouve parmi eux 1943 possesseurs d'une voiture. Estimez, à partir de cet échantillon, la proportion des ménages ayant une voiture. Degré de confiance 99 %.
- À partir de la proportion estimée, combien de ménages faudrait-il interroger pour construire, avec un risque d'erreur de 5 %, un intervalle de confiance d'amplitude 0,04 ?

Ex 14 : On étudie le pourcentage d'utilisation d'une machine. 400 observations ont été effectuées qui ont donné le résultat suivant :

- Machine marche : 320 observations.
- Machine arrêtée : 80 observations.

- Entre quelles limites peut-on fixer le taux d'utilisation de la machine avec un degré de confiance de 95 % ?
- On fait un plus grand nombre d'observations. On obtient le même pourcentage d'utilisation ce qui permet, avec un risque d'erreur de 5 %, de fixer les limites de confiance à [78,4 % ; 81,6 %]. Combien a-t-on fait d'observations ?

Ex 15 : Un échantillon aléatoire de 50 notes (sur 100) dans une population de 200 a donné une moyenne de 75 et un écart type de 10.

- Quelles sont les limites de confiance à 95 % pour estimer la moyenne des 200 notes ?
- Avec quel degré de confiance peut-on dire que la moyenne des 200 notes est de 75 plus ou moins 1 ?

Ex 16 : Un échantillon de 150 lampes de marque A a donné une durée de vie moyenne de 1400 heures et un écart type de 120 heures. Un échantillon de 200 lampes de marque B a donné une durée de vie moyenne de 1200 heures et un écart type de 80 heures. Déterminer les limites de confiance à 95 % de la différence des durées de vie moyennes des marques A et B.

Ex 17 : Sur un échantillon de 400 adultes et de 600 adolescents ayant regardé un certain programme de télévision, 100 adultes et de 300 adolescents l'ont apprécié. Calculer les limites de confiance à 99 % de la différence des fréquences des adultes et des adolescents qui ont regardé et apprécié le programme.

Ex 18 : Une compagnie fabrique des roulements à billes ayant un poids moyen de 0,638 Kg et un écart type de 0,012 Kg Calculer les limites de confiance à 95 % des poids de lots comprenant 100 roulements chacun.

Ex 19 : Dans une population de 579 individus, divisée en quatre strates comprenant respectivement 53 ; 190 ; 231 ; et 105 individus, on a prélevé un échantillon de 58 individus, dont 10 dans la première strate, 14 dans la deuxième, 21 dans la troisième et 13 dans la quatrième. En fonction des résultats suivants, estimez la moyenne de la population globale et l'erreur standard de cette moyenne, en considérant l'échantillon comme :

- a) Aléatoire et simple ;
- b) Stratifié.

Strates	Sommes	Sommes des carrés
1	54	1004
2	127	3081
3	388	13270
4	553	39667

Ex 20 : Soit une variable aléatoire X de densité de probabilité $f(x,\lambda)$ définie par :

$$f(x,\lambda) = \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{x^2}{2\lambda}} \quad \text{pour tout nombre réel } x.$$

- a) Reconnaître la loi de la variable X et en déduire, sans calcul, l'espérance mathématique et la variance de X.
- b) Déterminer un estimateur de maximum de vraisemblance de λ associé à un échantillon aléatoire de taille n.
- c) L'estimateur précédent est-il un estimateur sans biais ?

LES TESTS STATISTIQUES

I. INTRODUCTION

Un test statistique est une méthode permettant de prendre une **décision** à partir d'informations fournies par un **échantillon**.

Les tests statistiques ou les tests d'hypothèses ont pour but de vérifier, à partir de données observées dans un ou plusieurs échantillons, la validité de certaines hypothèses relatives à une ou plusieurs populations.

On peut distinguer différents types de tests, en fonction des hypothèses auxquelles on a affaire.

Les tests de comparaison à une norme ou tests de conformité sont destinés à comparer entre eux une population théorique et un échantillon observé. Ils servent à vérifier si un échantillon donné peut être considéré comme extrait d'une population possédant telle caractéristique particulière (telle moyenne, telle variance, ...). Le test se fait en vérifiant si la différence entre la valeur observée et la valeur théorique du paramètre considéré peut être attribuée au hasard ou non.

Les tests d'homogénéité ou d'égalité ont pour but de comparer entre elles un certain nombre de populations, à l'aide d'un même nombre d'échantillons.

Les tests d'ajustement sont destinés à vérifier si un échantillon observé peut être extrait d'une population donnée.

Les tests d'indépendance ont pour but de contrôler, à partir d'un échantillon, l'indépendance de deux ou plusieurs critères de classification, généralement qualitatifs.

II. LE PRINCIPE D'UN TEST STATISTIQUE

Pour commencer, on émet une certaine hypothèse à tester, appelée **hypothèse nulle**, généralement désignée par H_0 . Celle-ci suppose toujours l'égalité des caractéristiques comparées.

L'hypothèse qui diffère de H_0 est dite **hypothèse alternative**, généralement désignée par H_1 .

On mesure ensuite l'écart observé entre les caractéristiques comparées, et on calcule la probabilité d'observer, si l'hypothèse nulle est vraie, un écart aussi important.

Si cette probabilité est relativement élevée, on considère l'hypothèse nulle comme plausible et on l'accepte. Par contre si la probabilité calculée est faible, l'écart observé apparaît comme peu compatible avec l'hypothèse nulle et on rejette celle-ci.

L'ensemble des valeurs observées pour lesquelles l'hypothèse nulle est admissible forme la région d'acceptation. Les autres valeurs constituent la région de rejet. Les valeurs limites sont appelées valeurs critiques.

La décision dépend donc de l'échantillon. Ainsi qu'elle que soit la décision prise, le hasard de l'échantillonnage peut fausser les conclusions. Quatre situations doivent en effet être envisagées:

L'acceptation de l'hypothèse nulle alors qu'elle est vraie, le rejet de l'hypothèse nulle alors qu'elle est vraie, l'acceptation de l'hypothèse nulle alors qu'elle est fautive, le rejet de l'hypothèse nulle alors qu'elle est fautive.

Dans le premier et le dernier cas, la conclusion obtenue est correcte, mais il n'en est malheureusement pas de même dans les deux cas intermédiaires. L'erreur qui consiste à rejeter une hypothèse vraie est appelée erreur de première espèce et désignée par RH_0/H_0 . Accepter une hypothèse fautive est une erreur de seconde espèce, elle est désignée par AH_0/H_1 .

Les probabilités d'aboutir à de telles conclusions erronées sont les risques de première et de deuxième espèce, désignés respectivement par α et β .

$$\alpha = p(RH_0/H_0) \qquad \beta = p(AH_0/H_1)$$

Le risque de première espèce α est appelé aussi seuil de signification du test, fixé très souvent à 5 %. La probabilité contraire de α désigne le niveau de confiance du test.

$$1-\alpha = p(AH_0/H_0)$$

La probabilité contraire de β désigne la **puissance du test**.

$$1-\beta = p(RH_0/H_1)$$

On peut présenter une table de décision comme suit :

		Décision prise	
		Accepter H_0	Accepter H_1
Hypothèse Vraie	H_0	$1-\alpha$ Niveau de confiance	α : erreur de première espèce
	H_1	β : erreur de deuxième espèce	$1-\beta$ Puissance du test

La détermination des valeurs limites de la région d'acceptation de l'hypothèse nulle dépend de l'hypothèse alternative H_1 , ainsi on distingue le test bilatéral et le test unilatéral.

2.1. Test bilatéral

Un test est dit bilatéral si la condition de rejet est indépendante du signe de l'écart observé entre les caractéristiques comparées. Les hypothèses formulées du test bilatéral sont :

$$H_0 : \theta = t_0 \quad \text{et} \quad H_1 : \theta \neq t_0$$

θ et t_0 sont les caractéristiques comparées.

La règle de décision peut être représentée ainsi :

$\theta \neq t_0$	$\theta = t_0$	$\theta \neq t_0$
Région de rejet de H_0	Région d'acceptation de H_0	Région de rejet de H_0
A1	A2	

A1 et A2 sont les valeurs critiques qui délimitent la région d'acceptation.

La région d'acceptation est donc l'intervalle $[A1 ; A2]$.

$$p(A1 \leq t_0 \leq A2) = 1 - \alpha$$

$$p(t_0 < A1) = p(t_0 > A2) = \alpha/2$$

2.2. Test unilatéral

Un test est dit unilatéral si l'hypothèse alternative désigne qu'une caractéristique est strictement supérieure ou inférieure à l'autre. On parle respectivement de test unilatéral à droite ou à gauche.

2.2.1 Test unilatéral à droite

Les hypothèses formulées du test unilatéral à droite sont :

$$H_0 : \theta = t_0 \quad \text{et} \quad H_1 : \theta > t_0$$

La règle de décision peut être représentée ainsi :

$\theta \leq t_0$	$\theta > t_0$
Région d'acceptation de H_0	Région de rejet de H_0
A	

A désigne la valeur critique qui délimite la région d'acceptation.

La région d'acceptation est donc l'intervalle $]-\infty ; A]$.

$$p(t_0 \leq A) = 1 - \alpha$$

$$p(t_0 > A) = \alpha$$

2.2.2. Test unilatéral à gauche

Les hypothèses formulées du test unilatéral à gauche sont :

$$H_0 : \theta = t_0 \quad \text{et} \quad H_1 : \theta < t_0$$

La règle de décision peut être représentée ainsi :

$\theta < t_0$	$\theta \geq t_0$
Région de rejet de H_0	Région d'acceptation de H_0

A

A désigne la valeur critique qui délimite la région d'acceptation.

La région d'acceptation est donc l'intervalle $[A ; +\infty [$.

$$p(t_0 < A) = \alpha$$

$$p(t_0 \geq A) = 1 - \alpha$$

pour récapituler, la démarche d'un test statistique est formée des étapes suivantes :

1. Formuler les hypothèses H_0 et H_1 ;
2. Fixer le seuil de signification α ;
3. Préciser la loi de probabilité de l'écart observé, appelé aussi variable de décision ;
4. Calculer la valeur numérique de la variable de décision ;
5. Déterminer les valeurs critiques qui délimitent la région d'acceptation ;
6. Prendre la décision et conclure.

III. TESTS STATISTIQUES SUR LES MOYENNES

3.1. Test de conformité d'une moyenne

Formulation de l'hypothèse nulle :

On attribue la valeur m_0 pour moyenne dans une population dont la vraie moyenne m est inconnue, et on veut juger la validité de cette hypothèse.

Ce test a pour but de vérifier si la moyenne m d'une population est ou n'est pas égale à une valeur donnée m_0 , appelée norme.

L'hypothèse nulle est donc : **H_0** **$m = m_0$**

Variable de décision :

On extrait de la population un échantillon aléatoire et simple dans lequel la moyenne observée

\bar{x} est en général différente de m_0 , il s'agit d'expliquer cette différence.

La variable de décision du test correspond à l'estimation de m qui est la moyenne de l'échantillon :

$$VD = \bar{x}$$

Pour une population normale d'écart type σ connu, la variable de décision est elle-même normale de moyenne m_0 et d'écart type. La variable de décision centrée réduite est donc :

$$VDR = \frac{\bar{x} - m_0}{\frac{\sigma}{\sqrt{n}}}$$

VDR est alors une variable normale réduite $N(0 ; 1)$.

Si la distribution de la population parent est inconnue, la quasi-variance sera utilisée comme estimation de la variance de la population. Pour un effectif suffisamment élevé, la variable de décision peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque l'effectif est supérieur à 30. Dans le cas contraire ($n < 30$), la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à $(n-1)$ degré de liberté.

Région d'acceptation :

La région d'acceptation dépend de l'hypothèse alternative H_1 .

a) Test bilatéral :

$$H_0 : m = m_0 \quad \text{et} \quad H_1 : m \neq m_0$$

Les valeurs critiques qui délimitent la région d'acceptation sont, pour une distribution normale réduite ou asymptotiquement normale réduite, Z_1 et Z_2 telles que :

$$p(Z_1 \leq \text{VDR} \leq Z_2) = 1 - \alpha$$

$$p(\text{VDR} < Z_1) = \alpha/2 \quad \Rightarrow \quad Z_1 = Z_{\frac{\alpha}{2}}$$

$$p(\text{VDR} > Z_2) = \alpha/2 \quad \Rightarrow \quad p(\text{VDR} \leq Z_2) = 1 - \alpha/2 \Rightarrow \quad Z_2 = Z_{1 - \frac{\alpha}{2}}$$

La région d'acceptation est donc l'intervalle $[Z_{\frac{\alpha}{2}} ; Z_{1 - \frac{\alpha}{2}}]$.

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|\text{VDR}| > Z_{1 - \frac{\alpha}{2}}$$

b) Test unilatéral à droite :

$$H_0 : m = m_0 \quad \text{et} \quad H_1 : m > m_0$$

La valeur critique qui délimitent la région d'acceptation est, pour une distribution normale réduite ou asymptotiquement normale réduite, Z telle que :

$$p(\text{VDR} \leq Z) = 1 - \alpha \quad \Rightarrow \quad Z = Z_{1 - \alpha}$$

La région d'acceptation est donc l'intervalle $]-\infty ; Z_{1 - \alpha}]$.

c) Test unilatéral à gauche :

$$H_0 : m = m_0 \text{ et } H_1 : m < m_0$$

La valeur critique qui délimitent la région d'acceptation est, pour une distribution normale réduite ou asymptotiquement normale réduite, Z telle que :

$$p(\text{VDR} < Z) = \alpha \Rightarrow Z = Z_\alpha$$

La région d'acceptation est donc l'intervalle $[Z_\alpha ; +\infty[$.

Remarque :

Pour une distribution de probabilité inconnue, et lorsque l'effectif de l'échantillon est inférieur à 30, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à $(n-1)$ degré de liberté. Les valeurs de Z sont remplacées par les valeurs de T de la loi de Student avec $(n-1)$ degré de liberté.

Exemple :

Le diamètre des billes fabriquées par une machine est en moyenne de 6 mm. Pour contrôler si la machine est bien réglée, on a prélevé un échantillon de 50 billes et on a mesuré leur diamètre. On a trouvé :

$$\sum x_i = 350 \qquad \sum x_i^2 = 2462$$

La machine est-elle bien réglée au seuil de signification de 95 % ?

Pour répondre à cette question, on doit vérifier si le diamètre moyen des 50 billes observées, est conforme à la norme de 6 mm. Il s'agit donc de faire un test de conformité de la moyenne.

Hypothèse nulle :

Il s'agit d'un test bilatéral $H_0 : m = 6 \quad H_1 : m \neq 6$

Variable de décision :

La variable de décision du test correspond à l'estimation de m qui est la moyenne de l'échantillon :

$$VD = \frac{\sum x_i}{50} = \frac{350}{50} = 7$$

La variable de décision peut être considérée comme une variable approximativement normale.

La variance de la population peut être estimée par la quasi-variance.

$$\hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{n\sigma^2}{n-1} = \frac{50(2462 - 7^2)}{49} = 0,24$$

$$\hat{\sigma} = \sqrt{0,24} = 0,49$$

$$\text{VDR} = \frac{\bar{x} - m_0}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{7-6}{\frac{0,49}{\sqrt{50}}} = 14,43$$

Région d'acceptation :

La région d'acceptation est l'intervalle $[Z_{\frac{\alpha}{2}} ; Z_{1-\frac{\alpha}{2}}]$.

Au seuil de signification de 95 % ($\alpha = 0,05$), les valeurs critiques qui délimitent la région d'acceptation sont :

$$Z_{\frac{\alpha}{2}} = Z_{0,025} = -1,96$$

$$Z_{1-\frac{\alpha}{2}} = Z_{0,975} = 1,96$$

La région d'acceptation est donc l'intervalle $[-1,96 ; 1,96]$.

On rejette l'hypothèse nulle car la variable de décision réduite n'appartient pas à la région d'acceptation. La machine n'est donc pas bien réglée au seuil de signification de 95 %

3.2. Test de comparaison des moyennes de deux échantillons indépendants

Ce test a pour but de comparer les moyennes de deux populations à l'aide de deux échantillons.

Soient deux échantillons aléatoires et non exhaustifs prélevés respectivement dans une population 1 de moyenne inconnue m_1 et dans une population 2 de moyenne inconnue m_2 . Les moyennes observées des deux échantillons x_1 et x_2 sont en général différentes, il s'agit d'expliquer cette différence.

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si la moyenne m_1 d'une population est ou n'est pas égale à la moyenne m_2 d'une autre population.

L'hypothèse nulle est donc : **H_0** **$m_1 = m_2$**

Variable de décision :

La variable de décision du test correspond à la différence entre les moyennes observées des deux échantillons :

$$VD = \bar{x}_1 - \bar{x}_2$$

Une distinction est faite entre le cas de deux populations de variances inégales et le cas de deux populations de variances égales.

a) cas de deux populations de variances inégales

Pour des populations normales (variances connues), les variables $\bar{x}_1 - \bar{x}_2$ sont des variables normales de moyennes respectivement m_1 et m_2 et d'écart type respectivement $\frac{\sigma_1}{\sqrt{n_1}}$ et $\frac{\sigma_2}{\sqrt{n_2}}$.

La variable de décision est elle-même normale de moyenne $(m_1 - m_2)$ et d'écart type $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Sous l'hypothèse nulle, $(m_1 - m_2) = 0$. La variable de décision centrée réduite :

$$VDR = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Est donc une variable normale réduite $N(0 ; 1)$.

Si les distributions des populations parents sont inconnues, pour des effectifs suffisamment élevés, la variable de décision peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque les effectifs sont supérieurs à 30. Dans le cas contraire, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à $(n_1 + n_2 - 2)$ degré de liberté.

b) cas de deux populations de variances inégales

Dans le cas où les populations sont de variances égales, une estimation de la variance commune aux deux populations est donnée par :

$$\hat{\sigma}^2 = \frac{\sum (x_i - \bar{x}_1)^2 + \sum (x_i - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

la variable de décision réduite devient :

$$\text{VDR} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\text{VDR} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sum (x_i - \bar{x}_1)^2 + \sum (x_i - \bar{x}_2)^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Si les distributions des populations parents sont inconnues, pour des effectifs suffisamment élevés, la variable de décision peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque les effectifs sont supérieurs à 30. Dans le cas contraire, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à $(n_1 + n_2 - 2)$ degré de liberté.

Région d'acceptation :

La région d'acceptation dépend de l'hypothèse alternative H_1 .

a) Test bilatéral :

$$H_0 : m_1 = m_2 \quad \text{et} \quad H_1 : m_1 \neq m_2$$

Les valeurs critiques qui délimitent la région d'acceptation sont, pour des distributions normales réduites ou asymptotiquement normales réduites, Z_1 et Z_2 telles que :

$$p(Z_1 \leq \text{VDR} \leq Z_2) = 1 - \alpha$$

$$p(\text{VDR} < Z_1) = \alpha/2 \quad \Rightarrow \quad Z_1 = Z_{\frac{\alpha}{2}}$$

$$p(\text{VDR} > Z_2) = \alpha/2 \quad \Rightarrow \quad p(\text{VDR} \leq Z_2) = 1 - \alpha/2 \Rightarrow \quad Z_2 = Z_{1 - \frac{\alpha}{2}}$$

La région d'acceptation est donc l'intervalle $\left[Z_{\frac{\alpha}{2}} ; Z_{1 - \frac{\alpha}{2}} \right]$.

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|VDR| > Z_{1-\frac{\alpha}{2}}$$

b) Test unilatéral à droite :

$$H_0 : m_1 = m_2 \quad \text{et} \quad H_1 : m_1 > m_2$$

La valeur critique qui délimitent la région d'acceptation est, pour des distributions normales réduites ou asymptotiquement normales réduites, Z telle que :

$$p(VDR \leq Z) = 1 - \alpha \quad \Rightarrow \quad Z = Z_{1-\alpha}$$

La région d'acceptation est donc l'intervalle $]-\infty ; Z_{1-\alpha}]$.

c) Test unilatéral à gauche :

$$H_0 : m_1 = m_2 \quad \text{et} \quad H_1 : m_1 < m_2$$

La valeur critique qui délimitent la région d'acceptation est, pour des distributions normales réduites ou asymptotiquement normales réduites, Z telle que :

$$p(VDR < Z) = \alpha \quad \Rightarrow \quad Z = Z_{\alpha}$$

La région d'acceptation est donc l'intervalle $[Z_{\alpha} ; +\infty[$.

Remarque :

Pour des distributions de probabilités inconnues, et lorsque les effectifs des échantillons sont inférieurs à 30, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à $(n-1)$ degré de liberté. Les valeurs de Z sont remplacées par les valeurs de T de la loi de Student avec $(n-1)$ degré de liberté.

Exemple :

Pour savoir s'il existe une différence d'assiduité entre les filles et les garçons, on a choisi de manière aléatoire et simple un premier échantillon de 10 filles et de façon indépendante, un deuxième échantillon de 10 garçons. En fonction des résultats ci-dessous relatifs aux notes d'assiduités (note sur 100), et en supposant que les variances des deux populations sont égales, peut-on conclure, au seuil de 5 %, à l'existence d'une différence significative entre les deux sexes ?

Assiduité des filles	72	67	52	54	46	58	59	54	58	63
Assiduité des garçons	66	59	54	57	63	55	61	55	66	75

Pour répondre à cette question, on doit réaliser un test de comparaison de deux moyennes.

Hypothèse nulle :

Ce test a pour but de vérifier si l'assiduité moyenne m_1 des filles est ou n'est pas égale à l'assiduité moyenne m_2 des garçons.

Il s'agit d'un test bilatéral :

$$H_0 : m_1 = m_2 \quad \text{et} \quad H_1 : m_1 \neq m_2$$

Variable de décision :

Les deux échantillons sont indépendants, les populations sont de variances égales, la variable de décision centrée réduite est donc:

$$VDR = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sum (x_i - \bar{x}_1)^2 + \sum (x_i - \bar{x}_2)^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{58,3 - 61,1}{\sqrt{\frac{514,1 + 390,9}{10 + 10 - 2} \left(\frac{1}{10} + \frac{1}{10} \right)}} = -0,88$$

Région d'acceptation :

$$|VDR| = 0,88$$

Pour $\alpha = 0,05$, la valeur de $t_{1-\frac{\alpha}{2}}$ avec 18 degrés de liberté est : $t_{0,975} = 2,101$

$|VDR| < t_{1-\frac{\alpha}{2}}$, on accepte donc l'hypothèse nulle. C'est à dire, il n'y a pas de différence significative entre l'assiduité des deux sexes.

3.3. Test de comparaison des moyennes de deux échantillons appariés

Ce test a pour but de comparer les moyennes de deux populations à l'aide de deux échantillons associés par paires. C'est le cas où on soumet les mêmes individus, choisis dans une population donnée, à deux types d'observations.

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si une action a un effet sur une variable c'est-à-dire si la moyenne des différences est nulle.

L'hypothèse nulle est donc : $H_0 \quad \bar{d} = 0$

Variable de décision :

Soient deux séries de n observations chacune, x_1, x_2, \dots, x_n , et y_1, y_2, \dots, y_n . On travaille avec la série des différences :

$$d_i = x_i - y_i$$

La variable de décision du test correspond à la moyenne des différences :

$$VD = \bar{d}$$

Pour une population normale, la variable de décision est elle-même normale de moyenne. La variable de décision centrée réduite est donc :

$$VDR = \frac{\bar{d}}{\frac{\sigma_d}{\sqrt{n}}}$$

VDR est alors une variable normale réduite $N(0 ; 1)$.

Si la distribution de la population parent est inconnue, pour un effectif suffisamment élevé, la variable de décision peut toujours être considérée comme une variable approximativement normale. C'est généralement le cas lorsque l'effectif est supérieur à 30. Dans le cas contraire ($n < 30$), la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à $(n-1)$ degré de liberté.

Région d'acceptation :

La région d'acceptation est identique à celle du test précédent. Elle dépend toujours de l'hypothèse alternative H_1 .

a) Test bilatéral :

$$H_0 : \bar{d} = \mathbf{0} \quad \text{et} \quad H_1 : \bar{d} \neq \mathbf{0}$$

La région d'acceptation est l'intervalle $[\frac{Z\alpha}{2} ; Z_{1-\frac{\alpha}{2}}]$.

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|VDR| > Z_{1-\frac{\alpha}{2}}$$

b) Test unilatéral à droite :

$$H_0 : \bar{d} = \mathbf{0} \quad \text{et} \quad H_1 : \bar{d} > \mathbf{0}$$

La région d'acceptation est l'intervalle $] -\infty ; Z_{1-\alpha}]$.

c) Test unilatéral à gauche :

$$H_0 : \bar{d} = \mathbf{0} \quad \text{et} \quad H_1 : \bar{d} < \mathbf{0}$$

La région d'acceptation est l'intervalle $[Z\alpha ; +\infty[$.

Remarque :

Pour des distributions de probabilités inconnues, et lorsque les effectifs des échantillons sont inférieurs à 30, la variable de décision réduite VDR peut toujours être considérée comme une variable de Student à (n-1) degré de liberté. Les valeurs de Z sont remplacées par les valeurs de T de la loi de Student avec (n-1) degré de liberté.

Exemple :

Un chef de produit souhaite tester l'effet d'un nouvel emballage sur les ventes d'un produit. Un échantillon aléatoire de 20 magasins est constitué, puis scindé en deux échantillons de 10 unités, couplés sur la base de leurs ventes hebdomadaires. L'un des magasins de chaque couple propose le produit dans son nouvel emballage, tandis que l'autre magasin présente le produit dans l'ancien emballage. Les ventes enregistrées sont indiquées dans le tableau ci-dessous. Peut-on parler d'un effet positif du nouvel emballage ?

Couple	Nouvel emballage	Ancien emballage	Différence (di)
1	4580	3970	610
2	5190	4880	310
3	3940	4090	-150
4	6320	5870	450
5	7680	6930	750
6	3480	4000	-520
7	5720	5080	640
8	7040	6950	90
9	5270	4960	310
10	5840	5130	710

Pour répondre à cette question, on doit réaliser un test de comparaison de deux moyennes.

Hypothèse nulle :

Ce test a pour but de vérifier si, en moyenne, les ventes enregistrées avec le nouvel emballage m_1 sont ou ne sont pas égales aux ventes enregistrées avec l'ancien emballage m_2 .

Il s'agit d'un test unilatéral à droite :

$$H_0 : \bar{d} = 0 \quad \text{et} \quad H_1 : \bar{d} > 0$$

Variable de décision :

Les deux échantillons sont associés par paires, la variable de décision centrée réduite est donc:

$$VDR = \frac{\bar{d}}{\frac{\sigma_d}{\sqrt{n}}} = \frac{320}{\frac{410,96}{\sqrt{10}}} = 2,462$$

Région d'acceptation :

$$|VDR| = 2,462$$

Pour $\alpha = 0,05$, la valeur de $t_{1-\alpha}$ avec 9 degrés de liberté est : $t_{0,95} = 1,833$

$|VDR| > t_{1-\alpha}$, on rejette donc l'hypothèse nulle. C'est à dire, on peut conclure que le nouvel emballage est plus performant que l'ancien.

3.1. Analyse de la variance

C'est une **méthode statistique pour tester l'égalité de plusieurs moyennes**. La méthode repose sur les postulats suivants: les échantillons aléatoires proviennent de populations distribuées normalement et ayant la même variance. Comme ces suppositions de base ne sont pas toujours satisfaites en pratique, l'analyste dispose aussi de méthodes dites non paramétriques pour comparer les échantillons entre eux.

Formulation de l'hypothèse nulle

L'analyse de variance, sert à effectuer le test de l'égalité de plusieurs moyennes. On écrit comme suit les hypothèses:

$$H_0: m_1 = m_2 = \dots = m_J$$

H_1 : au moins une des moyennes est différente des autres.

En effet, l'analyse de variance est une technique d'analyse statistique qui permet de tester globalement l'égalité des moyennes de J populations normales dans lesquelles on suppose que les variances sont égales ($\sigma_1^2 = \sigma_2^2 = \dots = \sigma_J^2 = \sigma^2$), même si elles demeurent inconnues. L'analyse de variance constitue une extension à J populations normalement distribuées, $J \geq 2$, du test de comparaison des moyennes de deux échantillons indépendants.

ANALYSE DE VARIANCE À UN FACTEUR

On essaie de découvrir si un seul facteur peut expliquer ou non les **variations** constatées dans les observations Y_{ij} . Au départ, on dispose d'échantillons prélevés aléatoirement dans des populations normales dans lesquelles les variances sont supposées égales ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots = \sigma_J^2$). Le tableau suivant illustre la notation indicée: par exemple, Y_{21} représente la deuxième observation prélevée de la première population. Dans chaque échantillon, on a aussi calculé le total des observations, la moyenne et la variance.

Matrice des données

POPULATION	P1 : N(m1,σ1)	P1 : N(m2,σ2)	...	P1 : N(mj,σj)
	Y ₁₁	Y ₁₂		Y _{1j}
	Y ₂₁	Y ₂₂	...	Y _{2j}

	y _{n11}	y _{n22}		y _{nij}
TOTAL	T1	T2	...	Tj
Moyenne	\bar{Y}_1	\bar{Y}_2	...	\bar{Y}_j
Variance	S ² ₁	S ² ₂	...	S ² _j

$$N = n_1 + n_2 + \dots + n_j \quad \text{Grand total} = T \quad \text{Moyenne générale} = \bar{Y}$$

Équation fondamentale de l'analyse de la variance

L'analyse de la variance développée par Fisher repose sur la comparaison de deux estimateurs de la variance σ^2 commune aux J populations normales.

a) Estimation de σ^2 par $\hat{\sigma}_T^2$

Un premier estimateur de σ^2 , noté $\hat{\sigma}_T^2$, est obtenu à partir de l'ensemble des $N = n_1 + n_2 + \dots + n_j$ observations en divisant la somme totale des carrés, STC, par ses degrés de liberté, soit (N-1). La statistique qui en découle est donnée par l'expression suivante:

$$\hat{\sigma}_T^2 = \frac{STC}{N-1} = \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2}{N-1}$$

b) Estimation de σ^2 par $\hat{\sigma}_M^2$

Un deuxième estimateur de σ^2 , noté $\hat{\sigma}_M^2$, est obtenu cette fois en mesurant la variabilité existante entre les moyennes des échantillons. On l'appelle parfois la moyenne des carrés inter-groupes, ou la moyenne des carrés due aux traitements. Dans ce qui suit, on la nomme la moyenne des carrés due au facteur (MCF); elle est calculée en divisant la somme des carrés due au facteur (SCF) par ses degrés de liberté, (J-1):

$$\hat{\sigma}_M^2 = \text{MCF} = \frac{\text{SCF}}{J-1} = \frac{\sum_{j=1}^J n_j (\bar{Y}_j - \bar{Y})^2}{J-1}$$

c) Estimation de σ^2 par $\hat{\sigma}_C^2$

Un troisième estimateur de σ^2 est obtenu cette fois en combinant les variances intra-échantillons ($S_1^2, S_2^2, \dots, S_J^2$) déjà présentées dans le tableau des données. La pondération attribuée à S_j^2 sera égale aux degrés de liberté de cette statistique, soit $(n_j - 1)$, $j=1, 2, \dots, J$. L'estimateur est appelé la moyenne des carrés due à l'erreur (MCE) et il est donné par les expressions équivalentes suivantes:

$$\hat{\sigma}_C^2 = \text{MCE} = \frac{\text{SCE}}{N-J} = \frac{\sum_{j=1}^J (n_j - 1) S_j^2}{N-J} = \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{N-J}$$

Les trois sommes de carrés présentées plus haut ne sont pas totalement indépendantes les unes des autres. Il existe en effet un résultat important qui montre que la somme totale des carrés est égale à la somme des deux autres sommes de carrés:

$$\text{STC} = \text{SCF} + \text{SCE}$$

C'est cette relation qui s'appelle l'équation fondamentale de l'analyse de la variance. La variabilité totale entre les observations est décomposée en une part due aux différences entre les modalités du facteur et une part de variabilité résiduelle.

Formules équivalentes

Pour effectuer les calculs à l'aide d'une calculatrice électronique, il est préférable d'utiliser les formules suivantes qui sont algébriquement équivalentes aux précédentes:

$$STC = \sum_{j=1}^J \sum_{i=1}^{n_j} Y_{ij}^2 - \frac{T^2}{N}$$

$$SCF = \sum_{j=1}^J \frac{T_j^2}{n_j} - \frac{T^2}{N}$$

$$SCE = STC - SCF$$

Tableau d'analyse de variance à un seul facteur

Il est d'usage de présenter les résultats d'une analyse de variance à un seul facteur dans un tableau comme celui-ci:

Analyse de variance à un facteur

Source de variation	Somme des carrés	Degrés de liberté	Moyenne des carrés	
Facteur	SCF	J-1	MCF	$\frac{MCF}{MCE}$
Erreur	SCE	N-J	MCE	
Totale	SCT	N-1		

Quand H_0 est vraie, MCF et MCE constituent deux estimateurs indépendants de σ^2 de sorte que le rapport $\tilde{F} = \frac{MCF}{MCE}$ obéit à une loi de Fisher avec (J-1) et (N-J) degrés de liberté. En vertu même de la construction du rapport \tilde{F} , on devra rejeter l'hypothèse nulle de l'égalité des moyennes $H_0 : \mu_1 = \mu_2 = \dots = \mu_J$ au seuil α si et seulement si la valeur de $\tilde{F} = \frac{MCF}{MCE}$ est **plus grande** que la valeur critique de la table $F_{1-\alpha} \diamond (J-1) \text{ et } (N-J) \text{ dl}$.

Exemple :

Un manufacturier japonais de puces électroniques songe à implanter une nouvelle usine au Maroc afin de desservir tout le marché nord-africain. Il hésite entre trois villes: Tanger, Casablanca et Eljadida. Selon son point de vue, le critère le plus important à prendre en considération pour déterminer l'emplacement de cette nouvelle usine est l'assiduité au travail des ouvriers.

Le manufacturier a visité au hasard dans chacune des villes considérées cinq grandes usines de fabrication et il a obtenu des administrateurs le taux d'absentéisme par 3500 journées de travail. Les résultats sont reproduits dans le tableau ci-dessous.

Données numériques

Ville	Echantillon	Total
Tanger	141; 127 ; 111; 124 ; 144	T1 = 647
Casablanca	157; 131; 105; 132 ; 163	T2 = 688
Eljadida	183; 161; 145 ; 157 ; 189	T3 = 835
J = 3	N = 15	T = 2170

A un seuil de 5%, peut-on conclure que le taux d'absentéisme au travail est le même **en moyenne** dans ces 3 villes?

On calcule en premier lieu les trois sommes des carrés:

$$STC = \sum_{j=i=1}^J \sum_{n_j} Y_{ij}^2 - \frac{T^2}{N} = 141^2 + 127^2 + \dots + 189^2 - \frac{2170^2}{15} = 8149,33$$

$$SCF = \sum_{j=i}^J \frac{T_j^2}{n_j} - \frac{T^2}{N} = \frac{647^2}{5} + \frac{688^2}{5} + \frac{835^2}{5} - \frac{2170^2}{15} = 3908,93$$

$$SCE = 8149,33 - 3908,93 = 4240,40$$

Ce qui permet la construction du tableau de l'analyse de variance.

Tableau d'analyse de variance à un facteur

Source de variation	Somme des carrés	Degrés de liberté	Moyenne des carrés	
Facteur	3908,93	2	1954,467	5,53
Erreur	4240,40	12	353,367	
Totale	8149,33	14		

A un seuil $\alpha = 5\%$, on ne peut pas conclure que l'assiduité des travailleurs à leur travail soit la même en moyenne dans ces 3 villes puisque la valeur observée 5,53 de F est supérieure à la valeur critique $F_{0,95 \text{ à } 2 \text{ et } 12 \text{ dl}} = 3,89$ obtenue de la distribution de Fisher à 2 et 12 degrés de liberté.

IV. TESTS STATISTIQUES SUR LES VARIANCES

4.1. Test de conformité d'une variance

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si la variance σ^2 d'une population est ou n'est pas égale à une valeur donnée σ^2_0 , appelée norme.

L'hypothèse nulle est donc : $H_0 : \sigma^2 = \sigma^2_0$

Variable de décision :

On extrait un échantillon aléatoire non exhaustif de taille n . La variable de décision du test correspond à :

$$VD = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2_0}$$

La variable de décision possède une distribution khi deux à $(n-1)$ degrés de liberté.

Région d'acceptation :

La région d'acceptation dépend de l'hypothèse alternative H_1 .

Test bilatéral :

$$H_0 : \sigma^2 = \sigma^2_0 \quad \text{et} \quad H_1 : \sigma^2 \neq \sigma^2_0$$

Les valeurs critiques qui délimitent la région d'acceptation sont χ^2_1 et χ^2_2 telles que :

$$p(\chi^2_1 \leq VD \leq \chi^2_2) = 1 - \alpha$$

$$p(VD < \chi^2_1) = \alpha/2 \quad \Rightarrow \quad \chi^2_1 = \chi^2_{\frac{\alpha}{2}}$$

$$p(VD > \chi^2_2) = \alpha/2 \quad \Rightarrow \quad p(VD \leq \chi^2_2) = 1 - \alpha/2 \quad \Rightarrow \quad \chi^2_2 = \chi^2_{1 - \frac{\alpha}{2}}$$

La région d'acceptation est donc l'intervalle $[\chi^2_{\frac{\alpha}{2}} ; \chi^2_{1 - \frac{\alpha}{2}}]$.

On accepte l'hypothèse nulle si la variable de décision appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Test unilatéral à droite :

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{et} \quad H_1 : \sigma^2 > \sigma_0^2$$

La valeur critique qui délimitent la région d'acceptation est χ^2 telle que :

$$p(\text{VD} \leq \chi^2) = 1 - \alpha \quad \Rightarrow \quad \chi^2 = \chi^2_{1-\alpha}$$

La région d'acceptation est donc l'intervalle $]0 ; \chi^2_{1-\alpha}]$.

Test unilatéral à gauche :

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{et} \quad H_1 : \sigma^2 < \sigma_0^2$$

La valeur critique qui délimite la région d'acceptation est χ^2 telle que :

$$p(\text{VD} < \chi^2) = \alpha \quad \Rightarrow \quad \chi^2 = \chi^2_{\alpha}$$

La région d'acceptation est donc l'intervalle $[\chi^2_{\alpha} ; +\infty[$.

Exemple :

On souhaite vérifier, au seuil de signification de 95 %, si le peuplement, dans lequel on a mesuré la hauteur d'un échantillon de 12 arbres, appartient à un type de forêt dont l'écart type est de 1,4 m. Les résultats en mètre sont :

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

Pour répondre à cette question, on doit réaliser un test de conformité de la variance.

Hypothèse nulle :

Il s'agit d'un test bilatéral.

$$H_0 \quad \sigma^2 = 1,4^2 = 1,96 \quad H_1 : \sigma^2 \neq 1,96$$

Variable de décision :

La variable de décision du test correspond à :

$$\text{VD} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma_0^2} = \frac{6,6}{1,96} = 3,37$$

La variable de décision possède une distribution khi deux à 11 degrés de liberté.

Région d'acceptation :

Les valeurs critiques qui délimitent la région d'acceptation sont : $\chi^2_{\frac{\alpha}{2}}$ et $\chi^2_{1-\frac{\alpha}{2}}$

Au seuil de signification de 95 % ($\alpha = 0,05$)

$$\chi^2_{0,025} = 3,82 \quad \text{et} \quad \chi^2_{0,975} = 21,9$$

La région d'acceptation est donc l'intervalle [3,82 ; 21,9].

On rejette l'hypothèse nulle car la variable de décision n'appartient pas à la région d'acceptation.

4.2. Test de comparaison des deux variances de deux échantillons indépendants

Ce test a pour but de comparer les variances de deux populations à l'aide de deux échantillons indépendants.

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si la variance σ^2_1 d'une population est ou n'est pas égale à la variance σ^2_2 d'une autre population.

L'hypothèse nulle est donc : $H_0 \quad \sigma^2_1 = \sigma^2_2$

Variable de décision :

Soient deux échantillons aléatoires et non exhaustifs prélevés dans les deux populations. La variable de décision du test correspond au rapport des deux variances observées des deux échantillons :

$$VD = \frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_2}$$

La variable de décision suit une loi de Fisher avec (n_1-1) et (n_2-1) degré de liberté.

Les tables de la loi de Fisher ne donnent que des valeurs supérieures à l'unité. C'est la raison pour laquelle la variable de décision correspond au rapport de variances qui est supérieur à l'unité, d'où l'échantillon 1 est celui qui a la plus grande variance.

Région d'acceptation :

Le test d'égalité de deux variances est en général un test bilatéral. Il précède généralement le test de comparaison des moyennes de deux échantillons indépendants.

$$H_0 \quad \sigma^2_1 = \sigma^2_2 \quad \text{et} \quad H_1 \quad \sigma^2_1 \neq \sigma^2_2$$

Les valeurs critiques qui délimitent la région d'acceptation sont F_1 et F_2 telles que :

$$p(F_1 \leq VD \leq F_2) = 1 - \alpha$$

$$p(VD < F_1) = \alpha/2 \quad \Rightarrow \quad F_1 = F_{\frac{\alpha}{2}}$$

$$p(VD > F_2) = \alpha/2 \quad \Rightarrow \quad p(VD \leq F_2) = 1 - \alpha/2 \quad \Rightarrow \quad F_2 = F_{1 - \frac{\alpha}{2}}$$

La région d'acceptation est donc l'intervalle $[F_{\frac{\alpha}{2}} ; F_{1 - \frac{\alpha}{2}}]$.

Les tables de la loi de Fisher ne donnent que des valeurs supérieures à l'unité, de telle sorte que seule est possible la comparaison avec $F_{1 - \frac{\alpha}{2}}$, et on rejette l'hypothèse nulle si la variable de décision est supérieure ou égale à $F_{1 - \frac{\alpha}{2}}$.

Exemple :

Pour savoir si les filles sont plus assidues que les garçons ou non, on a choisi de manière aléatoire et simple un premier échantillon de 10 filles et de façon indépendante, un deuxième échantillon de 10 garçons. En fonction des résultats ci-dessous relatifs aux notes d'assiduités (note sur 100), peut-on supposer, au seuil de 5 %, que les variances des deux populations sont égales ?

Assiduité des filles	72	67	52	54	46	58	59	54	58	63
Assiduité des garçons	66	59	54	57	63	55	61	55	66	75

Pour répondre à cette question, on doit réaliser un test de comparaison de deux variances.

Hypothèse nulle :

Ce test a pour but de vérifier si la variance σ^2_1 de la population des filles est ou n'est pas égale à la variance σ^2_2 de la population des garçons.

Il s'agit d'un test bilatéral : $H_0 \quad \sigma^2_1 = \sigma^2_2 \quad \text{et} \quad H_1 \quad \sigma^2_1 \neq \sigma^2_2$

Variable de décision :

$$VD = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{57,12}{43,43} = 1,31$$

Région d'acceptation :

Pour $\alpha = 0,05$ la valeur de $F_{1-\frac{\alpha}{2}}$ avec 9 et 9 degrés de liberté est :

$$F_{0,975} = 4,03$$

La variable de décision est inférieure à $F_{1-\frac{\alpha}{2}}$, on accepte donc l'hypothèse d'égalité des variances des deux populations.

V. TESTS STATISTIQUES SUR LES PROPORTIONS

5.1. Test de conformité d'une proportion

Formulation de l'hypothèse nulle :

On attribue la valeur p_0 pour proportion dans une population dont la vraie proportion p est inconnue, et on veut juger la validité de cette hypothèse.

Ce test a pour but de vérifier si la proportion p d'une population est ou n'est pas égale à une valeur donnée p_0 , appelée norme.

L'hypothèse nulle est donc : **$H_0 \quad p = p_0$**

Variable de décision :

On extrait de la population un échantillon aléatoire et simple dans lequel la proportion observée f_n est en général différente de p_0 , il s'agit d'expliquer cette différence.

La variable de décision du test correspond à l'estimation de p qui est la fréquence de l'échantillon :

$$VD = f_n$$

Comme, la distribution de la proportion suit une loi normale de moyenne p et d'écart type $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ à condition que la taille de l'échantillon soit supérieure ou égale à 30 ($n \geq 30$) et le produit $n p \geq 5$, la variable de décision réduite :

$$VDR = \frac{fn - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

est donc une variable normale réduite $N(0 ; 1)$.

Région d'acceptation :

La région d'acceptation dépend de l'hypothèse alternative H_1 .

Test bilatéral :

$$H_0 : p = p_0 \quad \text{et} \quad H_1 : p \neq p_0$$

Les valeurs critiques qui délimitent la région d'acceptation sont les valeurs d'une variable normale réduite Z_1 et Z_2 telles que :

$$p(Z_1 \leq VDR \leq Z_2) = 1 - \alpha$$

$$p(VDR < Z_1) = \alpha/2 \Rightarrow Z_1 = Z_{\frac{\alpha}{2}}$$

$$p(VDR > Z_2) = \alpha/2 \Rightarrow p(VDR \leq Z_2) = 1 - \alpha/2 \Rightarrow Z_2 = Z_{1 - \frac{\alpha}{2}}$$

La région d'acceptation est donc l'intervalle $[Z_{\frac{\alpha}{2}} ; Z_{1 - \frac{\alpha}{2}}]$.

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|VDR| > Z_{1 - \frac{\alpha}{2}}$$

Test unilatéral à droite :

$$H_0 : p = p_0 \quad \text{et} \quad H_1 : p > p_0$$

La valeur critique qui délimite la région d'acceptation est la valeur d'une variable normale réduite Z telle que :

$$p(VDR \leq Z) = 1 - \alpha \Rightarrow Z = Z_{1 - \alpha}$$

La région d'acceptation est donc l'intervalle $]-\infty ; Z_{1 - \alpha}]$.

Test unilatéral à gauche :

$$H_0 : p = p_0 \quad \text{et} \quad H_1 : p < p_0$$

La valeur critique qui délimite la région d'acceptation est la valeur d'une variable normale réduite Z telle que :

$$p(\text{VDR} < Z) = \alpha \quad \Rightarrow \quad Z = Z_\alpha$$

La région d'acceptation est donc l'intervalle $[Z_\alpha ; +\infty[$.

Exemple :

Au cours des élections, un candidat est élu avec 52 % des voix. Plusieurs mois après l'élection, un institut de sondage interroge 1600 électeurs, dont 800 déclarent qu'ils voteraient en cas d'élection, pour le même candidat. Ce résultat est-il ou non significatif d'une désaffection des électeurs pour l'élu ?

Pour répondre à cette question, on doit vérifier si le nouveau pourcentage obtenu par le sondage, n'est pas inférieur à la norme de 52 %. Il s'agit donc de faire un test de conformité de la proportion.

Hypothèse nulle :

Il s'agit d'un test unilatéral à gauche $H_0 \quad p = 0,52 \quad H_1 : p < 0,52$

Variable de décision :

La variable de décision du test correspond à la fréquence f_n de l'échantillon :

$$\text{VD} = f_n = \frac{800}{1600} = 0,50$$

La distribution de la proportion suit une loi normale de moyenne p et d'écart type $\frac{\sqrt{pq}}{\sqrt{n}}$ (la taille de l'échantillon est supérieure à 30 et le produit $n p > 5$).

La variable de décision réduite est :

$$\text{VDR} = \frac{f_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0,50 - 0,52}{\sqrt{\frac{0,52(1-0,52)}{1600}}} = -1,60$$

Région d'acceptation :

La région d'acceptation est l'intervalle $[Z\alpha ; +\infty[$.

Au seuil de signification de 95 % ($\alpha = 0,05$) : $Z\alpha = Z_{0,05} = -1,65$

La région d'acceptation est donc l'intervalle $[-1,65 ; +\infty[$.

On accepte l'hypothèse nulle car la variable de décision réduite appartient à la région d'acceptation. Ce résultat n'est donc pas significatif d'une désaffection des électeurs pour ce candidat.

5.2. Test de comparaison des proportions de deux échantillons indépendants

Ce test a pour but de comparer les proportions de deux populations à l'aide de deux échantillons indépendants.

Formulation de l'hypothèse nulle :

Ce test a pour but de vérifier si la proportion p_1 d'une population est ou n'est pas égale à la proportion p_2 d'une autre population.

L'hypothèse nulle est donc : **$H_0 \quad p_1 = p_2$**

Variable de décision :

Il s'agit de comparer deux proportions observées. Soient deux échantillons aléatoires de taille respectivement n_1 et n_2 extraits de deux populations. Les fréquences observées f_{n1} et f_{n2} sont généralement différentes, il s'agit d'expliquer cette différence.

$$f_{n1} = \frac{X_1}{n_1} \quad \text{et} \quad f_{n2} = \frac{X_2}{n_2}$$

La variable de décision du test correspond à la différence entre les fréquences observées des deux échantillons :

$$VD = f_{n1} - f_{n2}$$

Comme, les distributions des deux proportions suivent des lois normales de moyennes respectivement p_1 et p_2 et d'écart types respectifs $\frac{\sqrt{p_1(1-p_1)}}{\sqrt{n_1}}$ et $\frac{\sqrt{p_2(1-p_2)}}{\sqrt{n_2}}$ à condition que la taille de l'échantillon soit supérieure ou égale à 30 ($n \geq 30$) et le produit $n p \geq 5$, la variable de décision est elle-même normale de moyenne (p_1-p_2) et d'écart type $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.

Sous l'hypothèse nulle $p_1 = p_2$, il y a la même proportion inconnue p dans les deux populations. Cette proportion peut être estimée par la fréquence observée $f_{n_1+n_2}$ dans l'échantillon unique qui est la réunion des deux échantillons.

$$f_{n_1+n_2} = \frac{X_1+X_2}{n_1+n_2} = \frac{n_1 f_{n_1} + n_2 f_{n_2}}{n_1+n_2}$$

Sous l'hypothèse nulle, la variable de décision suit une loi normale de moyenne $(p_1-p_2) = 0$ et d'écart type :

$$\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{f_{n_1+n_2}(1-f_{n_1+n_2})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

La variable de décision centrée réduite :

$$VDR = \frac{f_{n_1} - f_{n_2}}{\sqrt{f_{n_1+n_2}(1-f_{n_1+n_2})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

est donc une variable normale réduite $N(0 ; 1)$.

Région d'acceptation :

La région d'acceptation est identique à celle du test de conformité d'une proportion, elle dépend de l'hypothèse alternative H_1 .

Test bilatéral :

$$H_0 : p_1 = p_2 \quad \text{et} \quad H_1 : p_1 \neq p_2$$

La région d'acceptation est l'intervalle $\left[Z_{\frac{\alpha}{2}} ; Z_{1-\frac{\alpha}{2}} \right]$.

On accepte l'hypothèse nulle si la variable de décision réduite appartient à la région d'acceptation. Sinon, c'est l'hypothèse alternative qui est acceptée.

Remarque :

Puisque la région d'acceptation est symétrique, on rejette l'hypothèse nulle si :

$$|VDR| > Z_{1-\frac{\alpha}{2}}$$

Test unilatéral à droite :

$$H_0 : p_1 = p_2 \quad \text{et} \quad H_1 : p_1 > p_2$$

La région d'acceptation est donc l'intervalle $]-\infty ; Z_{1-\alpha}]$.

Test unilatéral à gauche :

$$H_0 : p_1 = p_2 \quad \text{et} \quad H_1 : p_1 < p_2$$

La région d'acceptation est donc l'intervalle $[Z_\alpha ; +\infty[$.

Exemple :

Une enquête sur l'emploi a concerné 220 personnes dont 115 dans le milieu rural et 105 dans le milieu urbain. Sur les 115 ruraux enquêtés, 74 se sont révélés actifs, alors que pour les enquêtés urbains, 81 sont actifs. Peut-on admettre, au seuil de 5 %, qu'il n'y a pas de différence significative entre les taux d'activités dans les deux milieux ?

Pour répondre à cette question, on doit réaliser un test de comparaison de deux proportions.

Hypothèse nulle :

Ce test a pour but de vérifier si la proportion p_1 des personnes actives dans le milieu rural est ou n'est pas égale à la proportion p_2 des personnes actives dans le milieu urbain.

Il s'agit d'un test bilatéral : $H_0 : p_1 = p_2$ et $H_1 : p_1 \neq p_2$

Variable de décision :

D'après les données :

$$f_{n1} = \frac{74}{115} = 0,64 \quad f_{n2} = \frac{81}{105} = 0,77 \quad f_{n2+n2} = \frac{155}{220} = 0,70$$

La variable de décision centrée réduite est :

$$VDR = \frac{fn_1 - fn_2}{\sqrt{fn_1 + n_2(1 - fn_1 + n_2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0,64 - 0,77}{\sqrt{0,70(1 - 0,70)\left(\frac{1}{115} + \frac{1}{105}\right)}} = -2,10$$

Région d'acceptation :

Pour $\alpha = 0,05$ la valeur de $Z_{1-\frac{\alpha}{2}}$ est : $Z_{0,975} = 1,96$

$|VDR| > Z_{1-\frac{\alpha}{2}}$, on rejette donc l'hypothèse nulle. C'est à dire, il y a une différence significative entre les taux d'activités dans les deux milieux.

5.3. Test de comparaison des proportions de plusieurs échantillons indépendants

Ce test a pour but de comparer les proportions d'un certain nombre de populations à l'aide du même nombre d'échantillons indépendants.

Formulation de l'hypothèse nulle

Ce test a pour but de vérifier si les proportions p_1, p_2, \dots, p_k de k populations sont égales. On écrit comme suit les hypothèses:

$$H_0: p_1 = p_2 = \dots = p_k$$

H_1 : au moins une des proportions est différente des autres.

Variable de décision :

Soient k échantillons aléatoires de taille respectivement n_1, n_2, \dots, n_k extraits de k populations. Il s'agit de comparer les effectifs observés n_{ij} dans les k échantillons et les effectifs attendus ou théoriques sous l'hypothèse nulle.

Effectifs observés

	Echantillon 1	Echantillon 2	...	Echantillon k
Avoir le caractère étudié	n11	n21	...	nk1
Ne pas avoir le caractère étudié	n12	n22	...	nk12
Total	n1.	n2.	...	nk.

Sous l'hypothèse nulle $p_1 = p_2 = \dots = p_k$, il y a la même proportion inconnue p dans les k populations. Cette proportion peut être estimée par la fréquence observée f dans l'échantillon unique qui est la réunion des k échantillons.

$$f = \frac{n11+n21+\dots+nk1}{n1.+n2.+...+nk.}$$

sous l'hypothèse nulle, les effectifs théoriques sont :

Effectifs théoriques

	Echantillon 1	Echantillon 2	...	Echantillon k
Avoir le caractère étudié	$f n1.$	$f n2.$...	$f nk.$
Ne pas avoir le caractère étudié	$(1 - f) n1.$	$(1 - f) n2.$...	$(1 - f) nk.$
Total	n1.	n2.	...	nk.

On est amené à confronter les effectifs observés et les effectifs théoriques. On calcule la variable de décision VD :

$$VD = \sum(\text{effectif observé} - \text{effectif théorique})^2 / \text{effectif théorique}$$

$$VD = \sum(\text{effectif observé}^2 / \text{effectif théorique}) - n$$

On peut démontrer que la variable de décision est une variable aléatoire Khi deux avec $(k-1)$ degré de liberté.

Région d'acceptation :

La variable de décision est nulle lorsque les effectifs observés sont tous égaux aux effectifs attendus, c'est à dire, lorsqu'il y a concordance absolue entre la distribution observée et la distribution théorique. La valeur de la variable de décision est d'autant plus grande que les écarts entre les effectifs observés et attendus sont plus grands. La valeur critique qui délimite la région d'acceptation est χ^2 telle que :

$$p(\text{VD} < \chi^2) = 1 - \alpha \Rightarrow \chi^2 = \chi^2_{1-\alpha}$$

Le test étant toujours unilatéral, la région d'acceptation est donc l'intervalle $[0 ; \chi^2_{1-\alpha}]$.

On rejettera donc l'hypothèse nulle lorsque la valeur de la variable de décision est supérieure ou égale à $\chi^2_{1-\alpha}$ avec $(k-1)$ degrés de liberté.

Exemple :

Lors d'une campagne électorale, un parti politique a effectué un sondage pour évaluer les intentions de vote en faveur de ce parti. Quatre échantillons indépendants ont été choisis dans quatre villes différentes. On a obtenu les résultats suivants :

	Rabat	Tanger	Oujda	Agadir
Voteront pour le parti	94	58	60	43
Ne Voteront pas pour le parti	240	230	252	197
Total	334	288	312	240

Au seuil de signification de 5 %, la proportion de la population des électeurs qui ont l'intention de voter pour ce parti est-elle identique dans les quatre villes ?

Formulation de l'hypothèse nulle

$$H_0: p_1 = p_2 = p_3 = p_4$$

H_1 : au moins une des proportions est différente des autres.

Variable de décision :

Sous l'hypothèse nulle : $p_1 = p_2 = p_3 = p_4$, il y a la même proportion inconnue p dans les 4 villes. Cette proportion peut être estimée par la fréquence observée f dans l'échantillon unique qui est la réunion des 4 échantillons.

$$f = \frac{94+58+60+43}{334+288+312+240} = 0,22$$

Sous l'hypothèse nulle, les effectifs théoriques sont :

Effectifs théoriques

	Rabat	Tanger	Oujda	Agadir
Voteront pour le parti	73,48	63,36	68,64	52,8
Ne Voteront pas pour le parti	260,52	224,64	243,36	187,2
Total	334	288	312	240

On calcule la variable de décision VD :

VD =

$$\frac{(94-73,48)^2}{73,48} + \frac{(240-260,52)^2}{260,52} + \frac{(58-63,36)^2}{63,36} + \frac{(230-224,64)^2}{224,64} + \frac{(60-68,64)^2}{68,64} + \frac{(252-243,36)^2}{243,36} + \frac{(43-52,8)^2}{52,8} + \frac{(197-187,2)^2}{187,2} = 11,65$$

La variable de décision est une variable aléatoire Khi deux avec 3 degrés de liberté.

Région d'acceptation :

La région d'acceptation est donc l'intervalle $[0 ; \chi^2_{1-\alpha}]$.

Au seuil de signification de 5 %, la valeur $\chi^2_{0,95}$ à 3 degrés de liberté est égale à 7,81.

La valeur de la variable de décision est supérieure à la valeur $\chi^2_{0,95}$ à 3 degrés de liberté, on rejettera donc l'hypothèse nulle, c'est à dire au seuil de signification de 5 %, la proportion de la population des électeurs qui ont l'intention de voter pour ce parti n'est pas identique dans les quatre villes.

VI. LES TESTS D'AJUSTEMENT

Les tests d'ajustement sont destinés à comparer une distribution observée et une distribution théorique donnée. D'une façon générale, on considère d'une part, une population infinie dont les individus sont classés en k catégories, en fonction d'un critère qualitatif ou quantitatif, et d'autre part, un échantillon aléatoire et simple d'effectif n , dont les individus sont classés de la même manière. Le but du test est de vérifier si la population possède une distribution de probabilité donnée :

$$p_1, p_2, p_3, \dots, p_k \quad \text{tel que : } \sum_{i=1}^k p_i = 1$$

Formulation de l'hypothèse nulle :

Pour comparer la distribution théorique et la distribution observée, on est amené à confronter les effectifs observés n_i et les effectifs attendus ou théoriques correspondants np_i .

L'hypothèse nulle est alors :

$$H_0 : n_i = np_i \quad \text{avec} \quad \sum_{i=1}^k n_i = \sum_{i=1}^k np_i = n$$

Variable de décision :

On distingue deux cas d'application de ces tests, selon que la distribution théorique est ou n'est pas complètement définie. Dans le premier cas, la variable de décision peut être calculée immédiatement. Dans le second cas, la distribution de probabilité de la population n'est définie qu'en fonction d'un ou de plusieurs paramètres, ceux-ci doivent préalablement être estimés à partir des données de l'échantillon.

Cas d'une distribution complètement définie :

Pour comparer la distribution théorique et la distribution observée, on est amené à confronter les effectifs observés n_i et les effectifs attendus ou théoriques correspondants np_i .

Les effectifs attendus doivent être tous supérieurs ou égaux à 5. quand cette condition n'est pas remplie, on peut regrouper des classes voisines, de manière à augmenter les effectifs attendus.

On calcule la variable de décision VD :

$$VD = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$$

On peut démontrer que la variable de décision est une variable aléatoire Khi deux avec $(k-1)$ degré de liberté. k correspond au nombre de classes après regroupement.

La variable de décision peut être calculée plus facilement par :

$$VD = \sum_{i=1}^k \frac{(ni - np_i)^2}{np_i} = \sum_{i=1}^k \frac{ni^2 - 2ninp_i + n^2p_i^2}{np_i} = \sum_{i=1}^k \frac{ni^2}{np_i} - \sum_{i=1}^k \frac{2ninp_i}{np_i} + \sum_{i=1}^k \frac{n^2p_i^2}{np_i} = \sum_{i=1}^k \frac{ni^2}{np_i} - 2n + n$$

$$VD = \sum_{i=1}^k \frac{ni^2}{np_i} - n$$

Cas d'une distribution incomplètement définie :

Lorsque la distribution théorique n'est pas complètement définie, le ou les paramètres qui caractérisent cette distribution doivent tout d'abord être estimés. On peut calculer ensuite les probabilités estimées \hat{p}_i , les effectifs attendus correspondants $n\hat{p}_i$, et la valeur de décision :

$$VD = \sum_{i=1}^k \frac{ni^2}{n\hat{p}_i} - n$$

Le nombre de degré de liberté $(k-1)$ doit être réduit du nombre de paramètres estimés.

Région d'acceptation :

La variable de décision est nulle lorsque les effectifs observés sont tous égaux aux effectifs attendus, c'est à dire, lorsqu'il y a concordance absolue entre la distribution observée et la distribution théorique. La valeur de la variable de décision est d'autant plus grande que les écarts entre les effectifs observés et attendus sont plus grands. La valeur critique qui délimite la région d'acceptation est χ^2 telle que :

$$p(VD < \chi^2) = 1 - \alpha \Rightarrow \chi^2 = \chi^2_{1-\alpha}$$

Le test étant toujours unilatéral, la région d'acceptation est donc l'intervalle $[0 ; \chi^2_{1-\alpha}[$.

On rejettera donc l'hypothèse nulle lorsque la valeur de la variable de décision est supérieure ou égale à $\chi^2_{1-\alpha}$.

Exemple :

Le tableau suivant donne la distribution de fréquences des nombres de garçons observés dans 1600 familles de 4 enfants, considérées comme choisies au hasard au sein d'une très large population. En fonction de ces résultats, peut-on affirmer, au seuil de 5 %, que le nombre de garçons suit une loi binomiale ?

Nombre de garçons	Nombre de familles
0	113
1	367
2	576
3	426
4	118
Total	1600

Pour répondre à cette question, on doit réaliser un test d'ajustement dans le but de comparer la distribution observée à la une distribution binomiale.

Hypothèse nulle :

$$H_0 : n_i = np_i \quad \text{avec} \quad \sum_{i=1}^k n_i = \sum_{i=1}^k np_i = n$$

Variable de décision :

Pour comparer la distribution théorique et la distribution observée, on est amené à confronter les effectifs observés n_i et les effectifs attendus ou théoriques correspondants np_i . on doit calculer alors les probabilités p_i en utilisant la loi binomiale.

La probabilité d'avoir un garçon est supposée égale à 0,5, la loi binomiale qui caractérise le nombre de garçons dans une famille de 4 enfants a pour paramètre 4 et 0,5.

En utilisant la formule de la loi binomiale, on trouve les probabilités suivantes :

$$p(x) = C_n^x p^x q^{n-x}$$

Distribution de la variable B(4 , 1/2)

x	p(x)
0	0,0625
1	0,2500
2	0,3750
3	0,2500
4	0,0625
Total	1

Le tableau suivant regroupe les effectifs observés n_i et les effectifs attendus ou théoriques correspondants np_i .

x	n_i	np_i
0	113	100
1	367	400
2	576	600
3	426	400
4	118	100
Total	1600	1600

Les effectifs théoriques sont tous supérieures à 5, on peut calculer la variable de décision :

$$VD = \sum_{i=1}^k \frac{n_i^2}{np_i} - n$$

$$VD = \frac{113^2}{100} + \frac{367^2}{400} + \frac{576^2}{600} + \frac{426^2}{400} + \frac{118^2}{100} - 1600 = 10,3$$

Région d'acceptation :

La région d'acceptation est l'intervalle $[0 ; \chi^2_{1-\alpha}]$.

Pour $\alpha = 0,05$, la valeur de $\chi^2_{1-\alpha}$ avec 4 degrés de liberté est : $\chi^2_{0,95} = 9,49$

La valeur de la variable de décision est supérieure à $\chi^2_{1-\alpha}$, on rejette donc l'hypothèse nulle.

VII. LES TESTS D'INDEPENDANCE

Les tests d'indépendance ont pour but de contrôler l'indépendance stochastique de deux ou plusieurs critères de classification. Ils permettent également d'effectuer des comparaisons de proportions.

Les tests d'indépendance concernent une population subdivisée en pq classes, en fonction de deux critères de classification. La distribution de probabilité correspondante est alors une distribution à deux dimensions, et les données relatives à tout échantillon sont présentées sous la forme d'un tableau de contingence.

Pour des échantillons aléatoires et simples, si les deux critères de classification sont indépendants, les probabilités p_{ij} de la distribution à deux dimensions peuvent être estimées par :

$$\hat{p}_{ij} = f_{i.} \times f_{.j} \quad \text{avec } f_{i.} = \frac{n_{i.}}{n} \quad \text{et} \quad f_{.j} = \frac{n_{.j}}{n} \text{ sont les fréquences relatives marginales.}$$

$n_{i.}$ et $n_{.j}$ sont les effectifs marginaux, et n_{ij} les effectifs conjoints.
Les effectifs attendus correspondants sont donc :

$$n \hat{p}_{ij} = n f_{i.} \times f_{.j} = n \frac{n_{i.}}{n} \frac{n_{.j}}{n} = \frac{n_{i.} \times n_{.j}}{n}$$

les effectifs attendus doivent tous être supérieurs ou égaux à 5.

Formulation de l'hypothèse nulle :

Pour comparer la distribution théorique et la distribution observée, on est amené à confronter les effectifs observés n_{ij} et les effectifs attendus ou théoriques correspondants $n \hat{p}_{ij}$.

L'hypothèse nulle est l'indépendance des deux critères de classification.

$$H_0 : n_{ij} = n \hat{p}_{ij}$$

Variable de décision :

la comparaison des effectifs observés et attendus se fait comme pour les tests d'ajustement, en calculant la variable de décision suivante :

$$VD = \sum_{i=1}^p \sum_{j=1}^q \frac{n_{ij}^2}{n \hat{p}_{ij}} - n$$

On démontre que la variable de décision est une variable aléatoire Khi deux avec $(p-1)(q-1)$ degré de liberté.

Région d'acceptation :

La valeur critique qui délimite la région d'acceptation est χ^2 telle que :

$$p(VD < \chi^2) = 1 - \alpha \quad \Rightarrow \quad \chi^2 = \chi^2_{1-\alpha}$$

Le test étant toujours unilatéral, la région d'acceptation est donc l'intervalle $[0 ; \chi^2_{1-\alpha}]$.

On rejettera donc l'hypothèse nulle lorsque la valeur de la variable de décision est supérieure ou égale à $\chi^2_{1-\alpha}$.

Exemple :

Un tour opérateur souhaite segmenter son marché. Il se demande s'il existe un lien entre le choix d'une destination de vacances et le niveau d'instruction. Les données recueillies ont été structurées sous forme de tableau de contingence.

Niveau d'instruction	Destination de vacances			Total
	Mer	Montagne	Désert	
Primaire	300	50	100	450
Secondaire	250	80	20	350
Supérieur	50	120	30	200
Total	600	250	150	1000

Hypothèse nulle :

L'hypothèse nulle est l'indépendance des deux critères de classification.

$$H_0 : n_{ij} = n \hat{p}_{ij}$$

Variable de décision :

Les effectifs attendus sont estimés par la formule : $n \hat{p}_{ij} = \frac{n_i \times n_j}{n}$

Niveau d'instruction	Destination de vacances			Total
	Mer	Montagne	Désert	
Primaire	270	112,5	67,5	450
Secondaire	210	87,5	52,5	350
Supérieur	120	50	30	200
Total	600	250	150	1000

$$VD = \sum_{i=1}^p \sum_{j=1}^q \frac{n_{ij}^2}{n \hat{p}_{ij}} - n = \frac{300^2}{270} + \frac{50^2}{112,5} + \frac{100^2}{67,5} + \frac{250^2}{210} + \dots + \frac{30^2}{30} - 1000 = 220,91$$

Région d'acceptation :

La région d'acceptation est l'intervalle $[0 ; \chi^2_{1-\alpha}]$.

Le nombre de degrés de liberté est égal à $(3-1)(3-1) = 4$.

Pour $\alpha = 0,05$, la valeur de $\chi^2_{1-\alpha}$ avec 4 degrés de liberté est $\chi^2_{0,95} = 9,49$.

La valeur de la variable de décision est supérieure à $\chi^2_{1-\alpha}$, on rejette donc l'hypothèse nulle. On conclut donc que le niveau d'instruction a une influence sur le choix d'une destination touristique.

EXERCICES SUR LES TESTS STATISTIQUES

Ex 1 : Un fabricant de tubes à essais pour laboratoire fonde sa publicité sur le fait que la durée de vie de ses tubes correspond à 1500 heures de chauffage à l'aide d'un bec Bunzen. Un laboratoire de contrôle de publicité constate que sur 100 tubes à essais, la durée moyenne de vie est de 1485 heures de chauffage avec un écart-type de 110 heures. Au risque 5%, la durée de vie des tubes à essais est-elle différente de 1500 heures de chauffage ?

Ex 2 : L'expérience suivante a été réalisée par Weldon : il a lancé un dé 315 672 fois, il a tiré 106 602 fois l'une des faces 5 ou 6. Peut-on accepter l'hypothèse selon laquelle le dé est équilibré, au risque de 5% ?

Ex 3 : Le directeur de ventes d'un laboratoire pharmaceutique veut savoir s'il existe des différences significatives entre les régions en terme de niveau d'accueil d'un nouveau produit. Les résultats suivants ont été obtenus auprès d'un échantillon aléatoire de clients :

Niveau d'accueil	Régions			
	Nord	Est	Sud	Ouest
Faible	22	35	0	5
Modéré	84	55	8	24
Elevé	25	17	22	12

Le niveau d'accueil dépend-t-il de la région ?

Ex 4 : Les moteurs des appareils électroménagers d'une marque M ont une durée de vie moyenne de 3000 heures avec un écart-type de 150 heures. À la suite d'une modification dans la fabrication des moteurs, le fabricant affirme que les nouveaux moteurs ont une durée de vie supérieure à celle des anciens. On a testé un échantillon de 50 nouveaux moteurs et on a trouvé une durée de vie moyenne de 3250 heures avec un écart-type égal à 150 heures. Les nouveaux moteurs apportent-ils une amélioration dans la durée de vie des appareils électroménagers au risque de 1% ?

Ex 5 : Dans une grande ville d'un pays donné, une enquête a été réalisée sur les dépenses mensuelles pour les loisirs. On a observé les résultats suivants:

- Sur 280 familles habitant le centre-ville, les dépenses mensuelles pour les loisirs sont en moyenne de 640 dh avec un écart-type de 120 dh.
- Sur 300 familles habitant la banlieue, les dépenses mensuelles pour les loisirs sont en moyenne de 610 dh avec un écart-type de 100 dh.

Peut-on dire au risque de 5 % que la part du budget familial consacré aux loisirs est différente suivant que la famille habite le centre-ville ou la banlieue ?

Ex 6 : Un fabricant affirme qu'au moins 95 % de l'équipement qu'il fournit à un dépositaire est conforme au cahier des charges. L'examen d'un échantillon de 200 pièces fournies montre que 18 pièces sont défectueuses. Que penser de l'affirmation du fabricant au seuil de confiance de 5 %?

Ex 7 : On prélève dans la production d'une machine, un échantillon de 100 tiges métalliques. La moyenne des longueurs des tiges de cet échantillon est 100,04 cm avec un écart-type de 0,16 cm. La machine est réglée en principe pour obtenir des tiges de 100 cm.

- 1°) Au risque de 5 %, peut-on dire que la machine est bien réglée ?
- 2°) Reprendre la question précédente avec un risque de 1 %.

Ex 8 : Pour une élection, on effectue un sondage pour évaluer les intentions de vote en faveur du candidat M. Dans la ville de casa, sur 450 personnes interrogées, 52% ont l'intention de voter pour M. Dans la ville de rabat, sur 300 personnes interrogées, 49 % ont l'intention de voter pour M. Au risque de 5%, y a-t-il une différence d'intention de vote dans ces deux villes?

Ex 9 : Un spécialiste en marketing a fait modifier la méthode traditionnellement utilisée pour effectuer la promotion d'un certain produit. A titre expérimental, il a observé dans 10 points de vente le nombre d'unités vendues en une semaine en utilisant la méthode existante. La semaine d'après, les mêmes points de vente ont utilisé la nouvelle méthode de promotion, on a observé le nombre d'unités vendues en cette semaine en utilisant. Les données recueillies sont comme suit :

Ancienne méthode: 48, 46, 47, 43, 46, 45, 49, 46, 47, 44.
 Nouvelle méthode: 56, 49, 53, 51, 48, 52, 55, 53, 49, 50.

La nouvelle méthode de promotion a-t-elle un effet positif sur les ventes ($\alpha = 5\%$)?

Ex 10 : Un chercheur a découvert un procédé efficace à 90 % pour prolonger la durée de vie des ballons à eau chaude. On teste son procédé sur 200 ballons. On constate qu'il est efficace pour 160 d'entre eux. L'affirmation du chercheur est-elle légitime au seuil de signification de 0,05 ?

Ex 11 : Un laboratoire annonce que l'un de ses médicaments est efficace à 95 %. Sur un échantillon de 400 personnes le traitement s'est révélé efficace sur 355 d'entre elles. Quel risque faut-il accepter si l'on considère que l'affirmation du laboratoire est légitime ?

Ex 12 : Dans le but de contrôler le poids net des sachets d'un produit alimentaire, on a prélevé deux échantillons respectivement de 10 et 12 sachets, on a obtenu les résultats suivant (en grammes) :

Éch 1	190	200	202	195	194	208	205	196	198	206		
Éch 2	210	204	203	189	194	195	206	205	200	201	198	197

Ces deux résultats sont-ils significativement différents en ce qui concerne le poids moyen %

Ex 13 : Au concours d'entrée à une école, l'épreuve de culture générale est notée de 0 à 50. on tire au hasard un échantillon de 100 candidats et l'on relève que les notes qu'ils ont obtenues se classent en cinq tranches de la manière suivante :

Tranches de notes	Nombre de candidats
Note ≤ 10	10
$10 < \text{Note} \leq 20$	20
$20 < \text{Note} \leq 30$	30
$30 < \text{Note} \leq 40$	20
$40 < \text{Note} \leq 50$	20

Le jury se demande s'il est justifié de considérer que la distribution des notes suit une loi normale dans la population de tous les candidats.

Ex 14 : 24 têtes d'ovins ont reçu 6 alimentations différentes pour constituer 4 répétitions et on a enregistré les gains moyens quotidiens en poids suivants :

Alim. 1	Alim. 2	Alim. 3	Alim. 4	Alim. 5	Alim. 6
590	460	600	640	690	690
760	430	460	660	600	650
700	540	610	720	550	680
640	470	510	580	480	740

Au seuil de 5 %, existe-t-il une différence significative quant à l'effet des différentes alimentations sur le gain moyen quotidien en poids des ovins ?

Ex 15 : Une entreprise commerciale à succursales multiples procède à un sondage dans ses magasins de rabat et casa. A rabat, sur 1000 clients interrogés, 350 déclarent souhaiter que le magasin reste ouvert jusqu'à 21 heures tandis qu'à casa, sur 900 clients, 280 ont émis ce même vœu. L'entreprise peut-elle, au seuil de signification de 5%, considérer que sa clientèle de rabat réagit comme celle de casa ?

Ex 16 : Une machine fabrique des pièces identiques. La moyenne des poids de 50 pièces prélevées dans la production est 68,2 grammes avec un écart-type de 2,5 grammes. On effectue un réglage sur la machine. On prélève un nouvel échantillon de 50 pièces. On trouve un poids moyen de 67, 5 grammes avec un écart-type de 2, 8 grammes. Peut-on affirmer, au risque 5 % que le réglage a modifié le poids des pièces ?

Ex 17 : Dans une population, soit p_1 , la proportion d'hommes possédant le baccalauréat et p_2 la proportion de femmes possédant le baccalauréat. Le tableau suivant correspond à la répartition de 200 individus choisis au hasard dans cette population.

	hommes	femmes
Possèdent le bac	32	26
ne possèdent pas le bac	64	78

Peut-on affirmer au risque 0,05, que p_1 et p_2 sont significativement différents ?

Ex 18 : Dans un pays M, le gouvernement a annoncé que le taux de chômage est de 15,6 %. Contestant ce chiffre, les députés de l'opposition ont fait appel à un institut de sondage. Celui-ci a réalisé une étude couvrant 4900 personnes en âge d'activité et a trouvé que le taux de chômage est de 16,4 %. Avec un niveau de confiance de 0,95 ; estimez-vous que l'opposition a raison de contester le chiffre annoncé par le gouvernement ?

Ex 19 : Une enquête a été réalisée au près d'un échantillon de 500 individus prélevé au sein d'une population cible de 4 millions d'individus. Les données que l'on possède sur cette population sont les suivantes :

Hommes 48% soit 1,92 millions d'hommes

Femmes 58% soit 2,08 millions de femmes

Sexe	Hommes		Femmes		Total
	%	Effectifs	%	Effectifs	
Niveau d'instruction					
Aucun	35	672000	50	1040000	1712000
Primaire	30	576000	25	520000	1096000
Secondaire	15	288000	10	208000	496000
Formation professionnelle	13	249600	10	208000	457600
Supérieur	7	134400	5	104000	238400
Total	100	1920000	100	2080000	4000000

Au dépouillement, on a trouvé que les individus qui ont formé l'échantillon ont les caractéristiques suivantes :

Sexe	Hommes		Femmes		Total
	%	Effectifs	%	Effectifs	
Niveau d'instruction					
Aucun	32	61	54	112	173
Primaire	28	54	23	48	102
Secondaire	18	35	12	25	60
Formation professionnelle	14	27	8	17	44
Supérieur	8	15	3	6	21
Total	100	192	100	208	400

L'échantillon prélevé est-il représentatif de la population étudiée ?

Ex 20 : Dans une population, on interroge un échantillon aléatoire de 400 personnes dont 160 sont âgées de 18 à 40 ans et 240 sont âgées de plus de 40 ans. On a trouvé que les pourcentages des personnes propriétaires de leur logement dans les deux groupes sont respectivement 35% et 45%. Ces deux résultats sont-ils significativement différents au seuil de signification de 5% ?

Ex 21 : On a enregistré plusieurs fois de suite le nombre de personnes qui se sont présenté à un guichet automatique bancaire, pendant des temps de 5 minutes.

Nombres d'arrivées	0	1	2	3	4	5	6	7	8	9	10
Fréquences absolues observées	1	4	12	18	22	17	11	6	4	3	2

Peut-on affirmer au seuil de signification de 5 % que le nombre de personnes qui se présentent à un guichet automatique bancaire, pendant un intervalle de temps de 5 minutes suit une loi de poisson ?

Ex 22 : Le tableau suivant donne le nombre d'étudiants qui ont été brillants et médiocres devant trois examinateurs :

	Examineur1	Examineur2	Examineur3	Total
Brillants	50	47	56	153
Médiocres	5	14	8	27
Total	55	61	64	180

Au seuil de 5 %, testez l'hypothèse selon laquelle le nombre d'étudiants médiocres est le même pour chaque examinateur.

TABLE DE LA FONCTION DE REPARTITION DE LA LOI NORMALE REDUITE

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99897	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997

Exemple :

La valeur de $\Pi(1,36)$ correspond à l'intersection entre la ligne correspondante à 1,3 et la colonne correspondante à 0,06, on peut lire la valeur 0,91309.

$$\Pi(-2,24) = 1 - \Pi(2,24) = 1 - 0,98745 = 0,01255$$

TABLE DE LA LOI KHI DEUX DE PEARSON

k / p	0,0005	0,001	0,005	0,01	0,025	0,05	0,1	0,2	0,3	0,4
1	0,0 ⁶ 393	0,0 ⁵ 157	0,0 ⁴ 393	0,0 ³ 157	0,0 ³ 982	0,0 ² 393	0,0158	0,0642	0,148	0,275
2	0,0 ² 100	0,0 ² 200	0,0100	0,0201	0,0506	0,103	0,211	0,446	0,713	1,02
3	0,0153	0,0243	0,0717	0,115	0,216	0,352	0,584	1,00	1,42	1,87
4	0,0639	0,0908	0,207	0,297	0,484	0,711	1,06	1,65	2,19	2,75
5	0,158	0,210	0,412	0,554	0,831	1,15	1,61	2,34	3,00	3,66
6	0,299	0,381	0,676	0,872	1,24	1,64	2,20	3,07	3,83	4,57
7	0,485	0,598	0,989	1,24	1,69	2,17	2,83	3,82	4,67	5,49
8	0,710	0,857	1,34	1,65	2,18	2,73	3,49	4,59	5,53	6,42
9	0,972	1,15	1,73	2,09	2,70	3,33	4,17	5,38	6,39	7,36
10	1,26	1,48	2,16	2,56	3,25	3,94	4,87	6,18	7,27	8,30
11	1,59	1,83	2,60	3,05	3,82	4,57	5,58	6,99	8,15	9,24
12	1,93	2,21	3,07	3,57	4,40	5,23	6,30	7,81	9,03	10,2
13	2,31	2,62	3,57	4,11	5,01	5,89	7,04	8,63	9,93	11,1
14	2,70	3,04	4,07	4,66	5,63	6,57	7,79	9,47	10,8	12,1
15	3,11	3,48	4,60	5,23	6,26	7,26	8,55	10,3	11,7	13,0
16	3,54	3,94	5,14	5,81	6,91	7,96	9,31	11,2	12,6	14,0
17	3,98	4,42	5,70	6,41	7,56	8,67	10,1	12,0	13,5	14,9
18	4,44	4,90	6,26	7,01	8,23	9,39	10,9	12,9	14,4	15,9
19	4,91	5,41	6,84	7,63	8,91	10,1	11,7	13,7	15,4	16,9
20	5,40	5,92	7,43	8,26	9,59	10,9	12,4	14,6	16,3	17,8
21	5,90	6,45	8,03	8,90	10,3	11,6	13,2	15,4	17,2	18,8
22	6,40	6,98	8,64	9,54	11,0	12,3	14,0	16,3	18,1	19,7
23	6,92	7,53	9,26	10,2	11,7	13,1	14,8	17,2	19,0	20,7
24	7,45	8,08	9,89	10,9	12,4	13,8	15,7	18,1	19,9	21,7
25	7,99	8,65	10,5	11,5	13,1	14,6	16,5	18,9	20,9	22,6
26	8,54	9,22	11,2	12,2	13,8	15,4	17,3	19,8	21,8	23,6
27	9,09	9,80	11,8	12,9	14,6	16,2	18,1	20,7	22,7	24,5
28	9,66	10,4	12,5	13,6	15,3	16,9	18,9	21,6	23,6	25,5
29	10,2	11,0	13,1	14,3	16,0	17,7	19,8	22,5	24,6	26,5
30	10,8	11,6	13,8	15,0	16,8	18,5	20,6	23,4	25,5	27,4

TABLE DE LA LOI KHI DEUX DE PEARSON (SUITE)

k / p	0,5	0,6	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,455	0,708	1,07	1,64	2,71	3,84	5,02	6,63	7,88	10,8	12,1
2	1,39	1,83	2,41	3,22	4,61	5,99	7,38	9,21	10,6	13,8	15,2
3	2,37	2,95	3,67	4,64	6,25	7,81	9,35	11,3	12,8	16,3	17,7
4	3,36	4,04	4,88	5,99	7,78	9,49	11,1	13,3	14,9	18,5	20,0
5	4,35	5,13	6,06	7,29	9,24	11,1	12,8	15,1	16,7	20,5	22,1
6	5,35	6,21	7,23	8,56	10,6	12,6	14,4	16,8	18,5	22,5	24,1
7	6,35	7,28	8,38	9,80	12,0	14,1	16,0	18,5	20,3	24,3	26,0
8	7,34	8,35	9,52	11,0	13,4	15,5	17,5	20,1	22,0	26,1	27,9
9	8,34	9,41	10,7	12,2	14,7	16,9	19,0	21,7	23,6	27,9	29,7
10	9,34	10,5	11,8	13,4	16,0	18,3	20,5	23,2	25,2	29,6	31,4
11	10,3	11,5	12,9	14,6	17,3	19,7	21,9	24,7	26,8	31,3	33,1
12	11,3	12,6	14,0	15,8	18,5	21,0	23,3	26,2	28,3	32,9	34,8
13	12,3	13,6	15,1	17,0	19,8	22,4	24,7	27,7	29,8	34,5	36,5
14	13,3	14,7	16,2	18,2	21,1	23,7	26,1	29,1	31,3	36,1	38,1
15	14,3	15,7	17,3	19,3	22,3	25,0	27,5	30,6	32,8	37,7	39,7
16	15,3	16,8	18,4	20,5	23,5	26,3	28,8	32,0	34,3	39,3	41,3
17	16,3	17,8	19,5	21,6	24,8	27,6	30,2	33,4	35,7	40,8	42,9
18	17,3	18,9	20,6	22,8	26,0	28,9	31,5	34,8	37,2	42,3	44,4
19	18,3	19,9	21,7	23,9	27,2	30,1	32,9	36,2	38,6	43,8	46,0
20	19,3	21,0	22,8	25,0	28,4	31,4	34,2	37,6	40,0	45,3	47,5
21	20,3	22,0	23,9	26,2	29,6	32,7	35,5	38,9	41,4	46,8	49,0
22	21,3	23,0	24,9	27,3	30,8	33,9	36,8	40,3	42,8	48,3	50,5
23	22,3	24,1	26,0	28,4	32,0	35,2	38,1	41,6	44,2	49,7	52,0
24	23,3	25,1	27,1	29,6	33,2	36,4	39,4	43,0	45,6	51,2	53,5
25	24,3	26,1	28,2	30,7	34,4	37,7	40,6	44,3	46,9	52,6	54,9
26	25,3	27,2	29,2	31,8	35,6	38,9	41,9	45,6	48,3	54,1	56,4
27	26,3	28,2	30,3	32,9	36,7	40,1	43,2	47,0	49,6	55,5	57,9
28	27,3	29,2	31,4	34,0	37,9	41,3	44,5	48,3	51,0	56,9	59,3
29	28,3	30,3	32,5	35,1	39,1	42,6	45,7	49,6	52,3	58,3	60,7
30	29,3	31,3	33,5	36,3	40,3	43,8	47,0	50,9	53,7	59,7	62,2

Pour lire une valeur $\chi^2_{\alpha, k, dl}$ dans la table, il suffit de lire l'intersection entre la colonne correspondante à la valeur de la probabilité cumulée $F(\chi^2_{\alpha, k, dl})$ et la ligne correspondante aux degrés de liberté k .

Exemple :

La valeur de $\chi^2_{0,95 \text{ à } 10 \text{ dl}}$ pour une probabilité de 0,95 correspond à l'intersection entre la colonne correspondante à 0,95 et la ligne correspondante à 10, on peut lire la valeur 18,3.

$$\chi^2_{0,95 \text{ à } 10 \text{ dl}} = 18,3$$

$$\chi^2_{0,05 \text{ à } 20 \text{ dl}} = 10,9$$

TABLE DE LA LOI T DE STUDENT

k / p	0,6	0,7	0,8	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Pour lire une valeur $t_{\alpha k dl}$ dans la table, il suffit de lire l'intersection entre la colonne correspondante à la valeur de la probabilité cumulée $F(t_{\alpha k dl})$ et la ligne correspondante aux degrés de liberté k .

Exemple :

La valeur de $t_{\alpha 10 dl}$ pour une probabilité de 0,95 correspond à l'intersection entre la colonne correspondante à 0,95 et la ligne correspondante à 10, on peut lire la valeur 1,812.

$$t_{0,95 \text{ à } 10 \text{ dl}} = 1,812$$

$$t_{0,7 \text{ à } 20 \text{ dl}} = 0,533$$

TABLE DE LA LOI F DE FISHER ($p = 0,95$)

K1 k2	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	246	248	250	252	253	254	254	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,62	8,58	8,55	8,54	8,53	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,75	5,70	5,66	5,65	5,64	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,50	4,44	4,41	4,39	4,37	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,81	3,75	3,71	3,69	3,68	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,38	3,32	3,27	3,25	3,24	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,08	3,02	2,97	2,95	2,94	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,86	2,80	2,76	2,73	2,72	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,70	2,64	2,59	2,56	2,55	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72	2,65	2,57	2,51	2,46	2,43	2,42	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,47	2,40	2,35	2,32	2,31	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,38	2,31	2,26	2,23	2,22	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,31	2,24	2,19	2,16	2,14	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,25	2,18	2,12	2,10	2,08	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,19	2,12	2,07	2,04	2,02	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,15	2,08	2,02	1,99	1,97	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,11	2,04	1,98	1,95	1,93	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,07	2,00	1,94	1,91	1,89	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,04	1,97	1,91	1,88	1,86	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,15	2,07	1,98	1,91	1,85	1,82	1,80	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,11	2,03	1,94	1,86	1,80	1,77	1,75	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,07	1,99	1,90	1,82	1,76	1,73	1,71	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,04	1,96	1,87	1,79	1,73	1,69	1,67	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,84	1,76	1,70	1,66	1,64	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,74	1,66	1,59	1,55	1,53	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,69	1,60	1,52	1,48	1,46	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,65	1,56	1,48	1,44	1,41	1,39
80	4,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,79	1,70	1,60	1,51	1,43	1,38	1,35	1,32
100	4,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,57	1,48	1,39	1,34	1,31	1,28
200	4,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,72	1,62	1,52	1,41	1,32	1,26	1,22	1,19
500	4,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,69	1,59	1,48	1,38	1,28	1,21	1,16	1,11
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,67	1,57	1,46	1,35	1,24	1,17	1,11	1,00

TABLE DE LA LOI F DE FISHER (p = 0,975)

K1 k2	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	∞
1	648	800	864	900	922	937	948	957	963	969	985	993	1001	1008	1013	1016	1017	1018
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,3	14,2	14,1	14,0	14,0	13,9	13,9	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,66	8,56	8,46	8,38	8,32	8,29	8,27	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,23	6,14	6,08	6,05	6,03	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,27	5,17	5,07	4,98	4,92	4,88	4,86	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,57	4,47	4,36	4,28	4,21	4,18	4,16	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,10	4,00	3,89	3,81	3,74	3,70	3,68	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,77	3,67	3,56	3,47	3,40	3,37	3,35	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42	3,31	3,22	3,15	3,12	3,09	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,33	3,23	3,12	3,03	2,96	2,92	2,90	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,18	3,07	2,96	2,87	2,80	2,76	2,74	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,05	2,95	2,84	2,74	2,67	2,63	2,61	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	2,95	2,84	2,73	2,64	2,56	2,53	2,50	2,49
15	6,20	4,76	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,86	2,76	2,64	2,55	2,47	2,44	2,41	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,79	2,68	2,57	2,47	2,40	2,36	2,33	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,72	2,62	2,50	2,41	2,33	2,29	2,26	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,67	2,56	2,44	2,35	2,27	2,23	2,20	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,62	2,51	2,39	2,30	2,22	2,18	2,15	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,57	2,46	2,35	2,25	2,17	2,13	2,10	2,09
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,50	2,39	2,27	2,17	2,09	2,05	2,02	2,00
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,44	2,33	2,21	2,11	2,02	1,98	1,95	1,94
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,39	2,28	2,16	2,05	1,97	1,92	1,90	1,88
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,34	2,23	2,11	2,01	1,92	1,88	1,85	1,83
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,31	2,20	2,07	1,97	1,88	1,84	1,81	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,18	2,07	1,94	1,83	1,74	1,69	1,66	1,64
50	5,34	3,98	3,39	3,06	2,83	2,67	2,55	2,46	2,38	2,32	2,11	1,99	1,87	1,75	1,66	1,60	1,57	1,55
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,06	1,94	1,82	1,70	1,60	1,54	1,51	1,48
80	5,22	3,86	3,28	2,95	2,73	2,57	2,45	2,36	2,28	2,21	2,00	1,88	1,75	1,63	1,53	1,47	1,43	1,40
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	1,97	1,85	1,71	1,59	1,48	1,42	1,38	1,35
200	5,10	3,76	3,18	2,85	2,63	2,47	2,35	2,26	2,18	2,11	1,90	1,78	1,64	1,51	1,39	1,32	1,27	1,23
500	5,05	3,72	3,14	2,81	2,59	2,43	2,31	2,22	2,14	2,07	1,86	1,74	1,60	1,46	1,34	1,25	1,19	1,14
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,83	1,71	1,57	1,43	1,30	1,21	1,13	1,00

Pour lire une valeur $F_{\alpha, k_1 \text{ et } k_2, dl}$ dans la table, il suffit de lire l'intersection entre la colonne correspondante à la valeur de k_1 et la ligne correspondante à la valeur de k_2 .

Exemple :

La valeur de $F_{\alpha, 10 \text{ et } 15, dl}$ pour une probabilité de 0,95 correspond dans la table de la loi F pour $p=0,95$, à l'intersection entre la colonne correspondante à 10 et la ligne correspondante à 15, on peut lire la valeur 2,54.

$$F_{0,95 \text{ à } 10 \text{ et } 15, dl} = 2,54$$

$$F_{0,975 \text{ à } 15 \text{ et } 20, dl} = 2,57$$