Statistiques appliquées

Tests & Sondages

MAT 4103

Sandro Franceschi

INSTITUT
POLYTECHNIQUE
DE PARIS

Table des matières

Première partie

Tests

Chapitre 1

Généralités sur les tests d'hypothèses

1.1 Statistique descriptive versus statistique inférentielle

Définition 1.1 (Statistique descriptive)**.** Les statistiques descriptives sont un ensemble de techniques utilisées pour **décrire** un ensemble de données et **résumer l'information** qui y est contenue à l'aide :

— d'**indicateurs** (moyenne, médiane, quantiles, ...)

— de **graphiques** (histogramme, diagramme camembert, ...)

Remarque. Il n'y a pas besoin des probabilités pour faire de la statistique descriptive !

Définition 1.2 (Statistique inférentielle)**.** Les statistiques inférentielles cherchent à **déterminer les caractéristiques d'un groupe** (la population) **à partir d'un sous groupe** (l'échantillon), avec une mesure de certitude de la prédiction (probabilité d'erreur). Cette recherche d'information est effectuée par exemple à l'aide

— d'**estimateurs** (évaluer un paramètre inconnu, sondages, ...)

— de **tests** (répondre à une question, valider/réfuter une hypothèse, ...)

Remarque. La statistique inférentielle utilise des modèles probabilistes ! Elle est utile pour faire des prévisions et prendre des décisions au vu des observations.

1.2 Introduction aux tests d'hypothèses

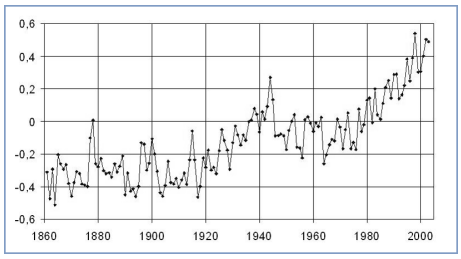
Pourquoi faire des tests ?

Exemple. Y-a-t-il bien, avec une erreur raisonnable, 1000mg de Vitamine C dans un comprimé industriel ? Sinon,

— y en a-t-il trop ? (industriel)

— en manque-t-il ? (consommateur)

Exemple. Y a-t'il un réchauffement climatique ? Le coefficient directeur de la droite de régression du graphe des températures au siècle dernier est-il significativement différent de 0 ?



Qu'est ce qu'un test d'hypothèse ? Un test d'hypothèse est un procédé qui a pour objectif de **déterminer la validité d'une hypothèse** relative à des populations à partir de l'étude d'échantillons aléatoires. Typiquement, un test a le choix entre deux hypothèses H_0 et H_1 (même si il peut en avoir parfois plus).

1.3 Tests d'hypothèses paramétriques/non paramétriques

Il y a deux grandes catégories de tests, les paramétriques et les non paramétriques.

Définition 1.3 (Tests paramétriques)**.** Test relatif au paramètre d'une variable aléatoire de loi spécifiée. Ce type de test nécessite une **hypothèse paramétrique**, on suppose que la loi des échantillons appartient à une famille de lois $(\mathbb{P}_\theta)_{\theta\in\Theta}$ indexée par un paramètre. (Typiquement une loi normale ou alors via une approximation normale pour des grands échantillons grâce au théorème central limite)

Exemple. Test de conformité, test de comparaison, ...

Remarque. Un test est dit **robuste** si il n'est pas sensible à des écarts aux hypothèses faites. Typiquement si le test reste valide sans l'hypothèse de normalité.

Définition 1.4 (Tests non paramétriques)**.** Test relatif à la loi d'une variable aléatoire (sans paramètre). Ce type de test ne nécessite pas d'hypothèse sur les lois des échantillons (pas d'hypothèse de normalité par exemple)

Exemple. Test d'adéquation à une loi, indépendance de variables aléatoires, homogénéité d'échantillons, ...

Remarque. Les tests paramétriques sont plus puissants que les tests non paramétriques mais sont à fortes contraintes (hypothèse de normalité, ...).

Remarque. Les tests non paramétriques, comme les tests du χ^2, peuvent s'utiliser même pour des échantillons de taille très faible.

1.4 Panorama des différents types de tests

Voici une classification des différents types de tests. Ce panorama n'est bien sur pas exhaustif.

— **Test de conformité** : comparer la valeur d'un paramètre θ de la loi d'1 échantillon, à une valeur pré-établie θ_0.

Exemple. Tests sur les moyennes, les variances, les proportions.

— **Test d'adéquation ou d'ajustement** : vérifier qu'1 échantillon suit bien une distribution choisie a priori.

Exemple. Test d'ajustement à la loi normale.

— **Test d'homogénéité ou de comparaison** : vérifier que 2 échantillons (ou plus) proviennent de la même population
-homogénéité faible : $\mathbb{E}[X] = \mathbb{E}[Y]$ (même espérance)
-homogénéité forte : $\mathcal{L}(X) = \mathcal{L}(Y)$ (même distribution)

Exemple. Comparaison du taux de glucose moyen d'individus ayant reçu des traitements différents.

— **Test d'indépendance** des lois de 2 échantillons (ou plus).

Exemple. La couleur des yeux est elle indépendante de la taille ?

Remarque. Certains de ces tests sont paramétriques (conformité, homogénéité faible/comparaison) et d'autres non paramétriques (adéquation/ajustement, homogénéité forte, indépendance).

Remarque. Certains de ces tests nécessitent un seul échantillon (conformité, adéquation/ajustement) et d'autres plusieurs (homogénéité/comparaison, indépendance).

1.5 Hypothèse nulle/hypothèse alternative

Hypothèse nulle H_0. C'est l'hypothèse que l'on désire contrôler. Elle est considérée comme vraie à priori et est **testée dans le but d'être potentiellement rejetée**.

Typiquement cette hypothèse consiste à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative et est due aux fluctuations d'échantillonnage.

Hypothèse alternative H_1. C'est l'hypothèse complémentaire de H_0.

Remarque. Une hypothèse est dite ***simple*** si, lorsqu'elle est vérifiée, elle permet de déterminer de manière unique la loi de l'échantillon. Elle est ***composite*** sinon.

Dissymétrie des hypothèses.

— **Rejeter** H_0 signifie que H_1 est accepté et très probablement vraie.

— **Conserver** H_0 n'est pas équivalente à H_0 est vraie mais juste qu'il n'y a pas d'évidence nette pour que H_0 soit fausse.

Remarque. Un test rejette ou à ne rejette pas une hypothèse nulle, mais ne l'accepte pas totalement d'emblée. Lorsqu'on ne la rejette pas, on peut dire qu'on la conserve en attendant de pousser l'étude plus loin.

Exemple (Test de comparaison / homogénéité faible)**.** On cherche à démontrer qu'un médicament modifie le taux de glucose dans le sang.

On compare μ_1 et μ_2 les moyennes du taux de glucose de deux populations, une sous traitement et l'autre sous placébo.

$$\begin{cases} H_0 : \mu_1 = \mu_2 & \text{(il est crédible de penser que } \mu_1 = \mu_2) \\ H_1 : \mu_1 \neq \mu_2 & (\mu_1 \text{ est significativement différente de } \mu_2) \end{cases}$$

— **Si H_0 est rejeté** alors le médicament modifie très probablement le taux de glucose dans le sang.

— **Si on ne peut pas rejeter** H_0 alors on n'a pas mis en évidence que le médicament modifie le taux de glucose, on dit qu'on conserve H_0 mais il faut pousser l'étude plus loin si besoin.

1.6 Tests, statistiques et régions de rejet

Soit $X = (X_1, \cdots, X_n)$ un n-échantillon (c'est à dire une **observation**) et considérons un choix entre deux hypothèses H_0 et H_1.

Définition 1.5 (Test d'hypothèse)**. Un test $\phi(X)$ est une application à valeur dans $\{0, 1\}$:**

— si $\phi(X) = 0$ on conserve H_0

— si $\phi(X) = 1$ on rejette H_0 et on accepte H_1

Définition 1.6 (Statistique et région de rejet)**.** En règle générale un test peut donc s'écrire

$$\boxed{\phi(X) = \mathbb{1}_{\{h(X)\in\mathcal{R}\}}}$$

On dit que $S = h(X)$ est la **statistique de test** et $\mathcal{R}$ la **région de rejet** (ou zone de rejet ou région critique). Une statistique est ainsi une fonction de l'échantillon dont la valeur numérique permet de déterminer si l'on conserve ou si l'on rejette H_0.

1.7 Risques d'erreurs

Soit H_0 une hypothèse nulle simple et H_1 son alternative. On considère un test $\phi(X) = \mathbb{1}_{\{S\in\mathcal{R}\}}$, une statistique S et une région de rejet $\mathcal{R}$.

Définition 1.7 (Risque de première espèce/niveau du test/seuil de signification)**.** Probabilité de choisir H_1 sachant que l'hypothèse nulle H_0 est vraie :

$$\alpha = \mathbb{P}(\phi(X) = 1|H_0) = \mathbb{P}(S \in \mathcal{R}|H_0).$$

Définition 1.8 (Risque de deuxième espèce)**.** Probabilité de choisir H_0 sachant que l'hypothèse alternative H_1 est vraie :

$$\beta = \mathbb{P}(\phi(X) = 0|H_1) = \mathbb{P}(S \notin \mathcal{R}|H_1).$$

Définition 1.9 (Puissance d'un test)**.** La **puissance du test** est $1 - \beta = \mathbb{P}(\phi(X) = 1|H_1)$.

Remarque. Si ϕ et ϕ' sont deux tests de niveau α, on dit que ϕ est (uniformément) **plus puissant** que ϕ' si $1 - \beta \geqslant 1 - \beta'$.

Risques d'erreurs

Réalité / Décision	H_0 vraie	H_0 fausse
On conserve H_0	**Vrai positif** Confiance $1 - \alpha$	**Faux positif** Risque de second espèce β
On rejette H_0	**Faux négatif** Risque de première espèce α	**Vrai négatif** Puissance $1 - \beta$

En général, il n'est pas possible de minimiser simultanément les risques d'erreur α et β :

— on minimise α (pour éviter les conséquences les plus graves),

— on "contrôle" β (pour que le test garde du sens).

Exemple. Suite à un problème sur la chaîne de production d'un médicament, une petite dose de produit toxique a été ajoutée

On décide de **tester la toxicité** du médicament sur des animaux pour choisir si

— on vend le médicament,

— ou on détruit le stock.

On prend donc pour hypothèse
H_0 : le médicament est toxique **versus** H_1 : il n'est pas toxique
Les risques à minimiser sont

— $\alpha = \mathbb{P}(\text{choisir } H_1 | H_0) \rightsquigarrow$ **tuer des innocents** (pour faire des économies),

— $\beta = \mathbb{P}(\text{choisir } H_0 | H_1) \rightsquigarrow$ **perdre de l'argent inutilement**.

Exercice (Pièce truquée)**.** On cherche à tester l'hypothèse qu'une pièce de monnaie n'est pas « **truquée** ». Nous faisons le test suivant :

— Statistique : $S =$ nombre de faces sur 100 lancés

— Hypothèses :
H_0 : pièce pas truquée **versus** H_1 : pièce truquée

— Région de rejet : $\mathcal{R} = [0, 40[\cup]60, 100]$

— La règle est donc : H_0 est

  — acceptée si $S \in [40, 60]$

  — rejetée si $S \notin [40, 60]$

1. Quel est le risque d'**erreur de première espèce** α ?
2. On suppose maintenant qu'une pièce truquée à une proba $0,6$ d'obtenir face. Quel est le **risque de seconde espèce** β ?

Solution. Cf cours en classe. □

Valeur seuil Soit un test $\phi(X) = \mathbb{1}_{\{S \in \mathcal{R}\}}$ et une statistique S

Pour un niveau de signification du test α donné (ex : 5%, 1%, 0,1%), on détermine une région de rejet $\mathcal{R}_\alpha$ telle que

$$\alpha = \mathbb{P}(S \in \mathcal{R}_\alpha | H_0) \quad \text{risque de 1ère espèce}$$

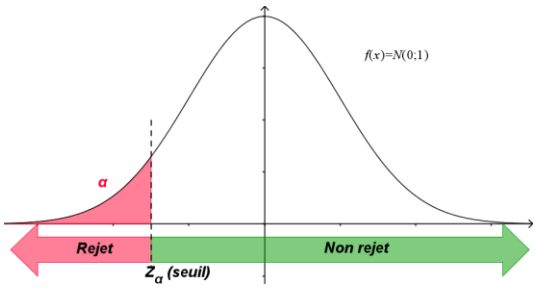
Pour déterminer $\mathcal{R}_\alpha$, on peut établir une **valeur seuil** s_α, il y a trois cas typiques :

— **Test unilatéral à droite** : $\mathcal{R}_\alpha = [s_\alpha, \infty[$ et $\alpha = \mathbb{P}(s_\alpha < S | H_0)$,

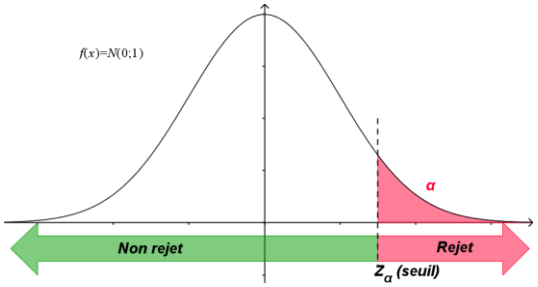
— **Test unilatéral à gauche** : $\mathcal{R}_\alpha =]-\infty, s_\alpha]$ et $\alpha = \mathbb{P}(S < s_\alpha | H_0)$,

— **Test bilatéral** : $\mathcal{R}_\alpha =]-\infty, -s_\alpha] \cup [s_\alpha, \infty[$ et $\alpha = \mathbb{P}(|S| > s_\alpha | H_0)$.

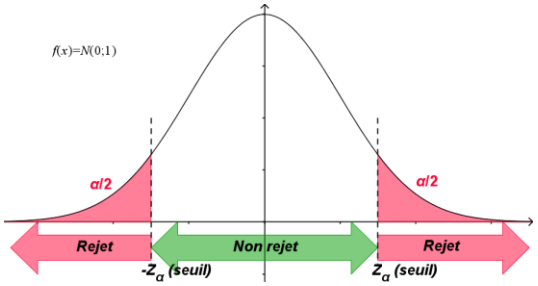
Illustration La courbe ci dessous est la densité de la statistique S **sous l'hypothèse** H_0 (par exemple un loi normale). La zone rouge est la région de rejet de H_0, on tombe dedans avec une probabilité α (erreur de première espèce ou niveau).



Test unilatéral à gauche



Test unilatéral à droite



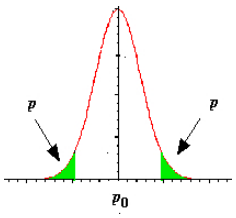
Test bilatéral

Test bilatéral/unilatéral La nature de H_0 détermine la façon de formuler H_1 et par conséquent la nature unilatérale ou bilatérale du test.

Exemple (Fumeurs)**.** H_0 : proportion de fumeurs étudiants p égal à proportion de fumeur dans la population générale p_0.

Test bilatéral H_1 se décompose en deux parties

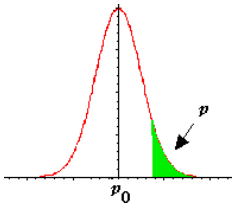
$H_0 : p = p_0$ **versus** $H_1 : p \neq p_0$



Test unilatéral H_1 se compose d'une seule partie

On suppose que la fréquence de fumeurs chez les étudiants p, ne peut pas être inférieure à la proportion de fumeurs dans la population générale p_0. On a alors

$H_0 : p = p_0$ **versus** $H_1 : p > p_0$



Risques d'erreurs Le graphique suivant illustre le fait qu'en règle générale il est impossible de diminuer simultanément les deux risques d'erreurs en jouant sur la valeur seuil.

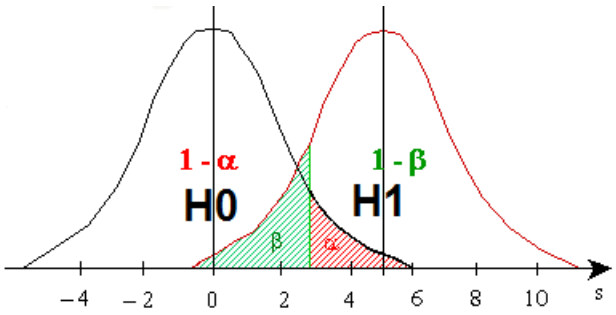
— Courbe noire : densité de la statistique S sous l'hypothèse H_0

— Courbe rouge : densité de la statistique S sous l'hypothèse H_1

— En rouge la région de rejet, risque 1ère espèce α

— En vert le risque de 2ème espèce β

— La valeur seuil se trouve entre la zone verte et la zone rouge



▶ **En général, si α diminue alors β augmente**

Remarque. La seule solution pour diminuer les deux risques d'erreur est d'augmenter la taille de l'échantillon (afin de réduire la variance).

Méthodologie Voici les étapes à suivre pour tester une hypothèse.

1. Définir l'hypothèse nulle H_0 et l'alternative H_1
2. Déterminer une statistique de test $h(X)$ pour contrôler H_0
3. Déterminer la loi de la statistique $h(X)$ sous H_0
4. Définir α le niveau du test et la région de rejet $\mathcal{R}_\alpha$ ou la valeur seuil s_α
5. Calculer avec l'échantillon x la valeur de la statistique $h(x)$
6. Conserver/rejeter H_0 en fonction de l'appartenance de $h(x)$ à $\mathcal{R}_\alpha$
7. Interprétation et commentaire sur le crédit à accorder au résultat (puissance du test, p-valeur)

Test de conformité de la moyenne, variance connue *Exemple le plus classique*

-Population de loi connue $\mathcal{N}(\mu_0, \sigma_0^2)$
-$X_1, \cdots, X_n$ échantillon de loi inconnue $\mathcal{N}(\mu, \sigma^2)$
Test de conformité : tester si l'échantillon appartient ou non à la population de loi connue

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

On suppose que $\sigma = \sigma_0$

— **Statistique** : $S = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}$ où $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$
sous H_0 on a $S \sim \mathcal{N}(0,1)$
Si $n > 30$ test robuste : on peut se passer de l'hypothèse de normalité (TCL)

— Région de rejet bilatérale $\mathcal{R}_\alpha =]-\infty, -s_\alpha] \cup [s_\alpha, \infty[$

— **Seuil** de niveau α égal au **quantile** de $1 - \alpha/2$

$$\text{i.e. } \boxed{s_\alpha = q_{1-\alpha/2}} \text{ où } \mathbb{P}(N < q_{1-\alpha/2}) = 1 - \alpha/2$$

— Niveau $\alpha = \mathbb{P}(|S| > q_{1-\alpha/2}|H_0)$

Exercice (Glycémie)**.** — La glycémie d'une population suit une loi normale d'espérance $\mu_0 = 1\text{g/l}$ et d'écart-type $\sigma_0 = 0,1\text{g/l}$.

— On relève les glycémies chez 9 patients qui ont un traitement. On trouve une moyenne empirique $\overline{x}_n = 1,12\text{g/l}$.

Cet échantillon est-il représentatif de la population, avec un risque d'erreur de 5% ?

Solution. Cf cours en classe. □

1.8 Probabilité critique ou p-valeur

La p-valeur ou probabilité critique sert à mesurer la qualité d'un test. C'est la probabilité, **sous** H_0, que la statistique soit au moins aussi éloignée de son espérance que la valeur observée. En d'autres termes, c'est la probabilité d'observer quelque chose d'au moins aussi surprenant que ce que l'on observe.

Définition 1.10 (p-valeur ou probabilité critique)**.** Soit un test $\phi(X) = \mathbb{1}_{\{S\in\mathcal{R}\}}$ et une statistique $S = h(X)$.

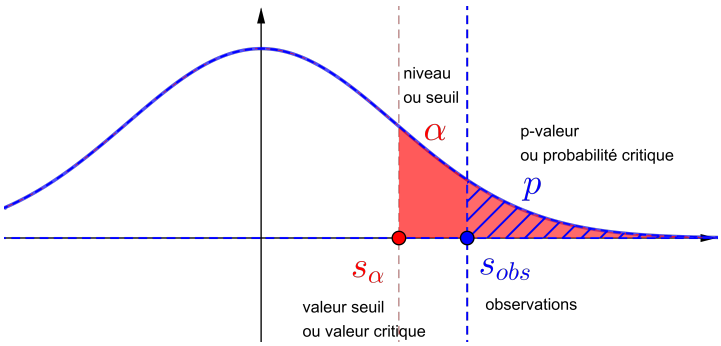
On note x la valeur observée et $s_{obs} = h(x)$.

La p-valeur vaut :

— Test bilatéral $p = \mathbb{P}(|S| > |s_{obs}||H_0)$

— Test unilatéral à droite $p = \mathbb{P}(S > s_{obs}|H_0)$

— Test unilatéral à gauche $p = \mathbb{P}(S < s_{obs}|H_0)$



Soit une famille de test $\phi_\alpha(X) = \mathbb{1}_{\{S\in\mathcal{R}_\alpha\}}$ de niveau α et soit x l'observation. La p-valeur peut aussi se voir comme :

— Plus haut niveau α permettant la conservation de H_0

$$p = \sup\{\alpha : \phi_\alpha(x) = 0\}$$

— Plus bas niveau α permettant le rejet de H_0

$$p = \inf\{\alpha : \phi_\alpha(x) = 1\}$$

Pour tester une hypothèse on peut donc calculer la p-valeur et la comparer au niveau α prédéterminé :

— $p > \alpha$ on **conserve** H_0,

— $p < \alpha$ on **rejette** H_0.

1.9 Lemme de Neyman-Pearson

La **vraisemblance** $V_\theta(X)$ est la **densité de l'échantillon** $X = (X_1, \cdots, X_n)$ de loi $\mathbb{P}_\theta$.

Test du rapport de vraisemblance

— $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$

— Statistique de test = **rapport de vraisemblance**

$$\boxed{S = \frac{V_{\theta_1}(X)}{V_{\theta_0}(X)}}$$

— Test de niveau α, valeur seuil s_α

— $\phi_\alpha(X) = \mathbb{1}_{\left\{\frac{V_{\theta_1}(X)}{V_{\theta_0}(X)} > s_\alpha\right\}}$

Proposition 1.11 (Lemme de Neyman-Pearson)**.** *Le test du rapport de vraisemblance est le plus puissant de niveau α. On dit qu'il est $UPP(\alpha)$ (uniformément plus puissant de niveau α).*

Exercice (Preuve du Lemme de Neyman-Pearson)**.** Considérons les deux hypothèses $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Soit ϕ le test du rapport de vraisemblance de niveau α et ϕ' un autre test de niveau $\alpha' \leqslant \alpha$.

1. Montrer que pour toute fonction g on a
$$\mathbb{E}[g(X)|H_1] = \mathbb{E}[g(X)\frac{V_{\theta_1}(X)}{V_{\theta_0}(X)}|H_0].$$
2. Montrer que
$$\mathbb{E}[\phi(X) - \phi'(X)|H_1] \geqslant s_\alpha \, \mathbb{E}[\phi(X) - \phi'(X)|H_0].$$
3. En déduire que $1 - \beta \geqslant 1 - \beta'$ et conclure.

Démonstration - Solution de l'exercice. 1. On a
$$\mathbb{E}[g(X)|H_1] = \int_{\mathbb{R}^n} g(x)V_{\theta_1}(x)\mathrm{d}x = \int_{\mathbb{R}^n} g(x)\frac{V_{\theta_1}(x)}{V_{\theta_0}(x)}V_{\theta_0}(x)\mathrm{d}x = \mathbb{E}[g(X)\frac{V_{\theta_1}(X)}{V_{\theta_0}(X)}|H_0].$$

2. D'après le point précédent on a ainsi
$$\mathbb{E}[\phi(X)-\phi'(X)|H_1]-s_\alpha \, \mathbb{E}[\phi(X)-\phi'(X)|H_0] = \mathbb{E}[(\phi(X)-\phi'(X))\left(\frac{V_{\theta_1}(X)}{V_{\theta_0}(X)} - s_\alpha\right)|H_0] \geqslant 0.$$
En effet, $(\phi(X) - \phi'(X))\left(\frac{V_{\theta_1}(X)}{V_{\theta_0}(X)} - s_\alpha\right) \geqslant 0$ car
   — si $\left(\frac{V_{\theta_1}(X)}{V_{\theta_0}(X)} - s_\alpha\right) > 0$ alors par définition $\phi(X) = 1$ et donc $(\phi(X) - \phi'(X)) \geqslant 0$.
   — si $\left(\frac{V_{\theta_1}(X)}{V_{\theta_0}(X)} - s_\alpha\right) \leqslant 0$ alors par définition $\phi(X) = 0$ et donc $(\phi(X) - \phi'(X)) \leqslant 0$.
3. Remarquons que $1 - \beta = \mathbb{E}[\phi(X)|H_1]$, $1 - \beta' = \mathbb{E}[\phi'(X)|H_1]$, $\alpha = \mathbb{E}[\phi(X)|H_0]$ et $\alpha = \mathbb{E}[\phi'(X)|H_0]$. L'inégalité établie à la question précédente et le fait qu'on suppose $\alpha \geqslant \alpha'$ donne
$$(1 - \beta) - (1 - \beta') \geqslant s_\alpha(\alpha - \alpha') \geqslant 0$$
en remarquant que $s_\alpha \geqslant 0$ car $S \geqslant 0$. Ainsi, $1 - \beta \geqslant 1 - \beta'$ et le test ϕ est donc plus puissant que ϕ'.

□

Chapitre 2

Lois de statistiques classiques

Les lois suivantes (normale $\mathcal{N}$, chi-deux χ^2, Student $\mathcal{T}$, Fisher-Snedecor $\mathcal{F}$) sont des **lois de statistiques de tests classiques**. Elle ont même parfois donné leur nom à ces tests. Dans la section 2.1 allons rappeler quelques unes de leurs propriétés et dans la section 2.2 nous étudierons les statistiques S auxquelles elles sont associées.

2.1 Loi normale, χ^2, Student, Fisher

2.1.1 Loi normale

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

— Densité

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

— $\mathcal{L}\left(\frac{X-\mu}{\sigma}\right) = \mathcal{N}(0,1)$

— $\mathcal{L}\left(aX+b\right) = \mathcal{N}(a\mu+b, a^2\sigma^2)$

— Si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ sont indépendantes alors

$$X_1 + X_2 \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$$

— Si $X_i \sim \mathcal{N}(\mu, \sigma^2)$ iid, la moyenne empirique a pour loi $\overline{X}_n = \frac{X_1+\cdots+X_n}{n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ et

$$\boxed{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)}$$

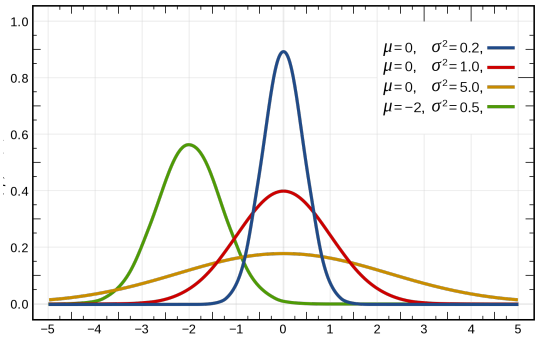


FIGURE 2.1 – Illustration de la densité de la loi normale en fonction des paramètres μ et σ

2.1.2 Loi du χ^2

— Loi du χ^2 à n degrés de liberté

$$K \sim \chi^2_n$$

— Si $X_1, \cdots, X_n$ sont des v.a. iid de loi $\mathcal{N}(0,1)$ alors

$$\boxed{K = X_1^2 + \cdots + X_n^2 \sim \chi^2_n}$$

— Densité

$$f_K(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$

— Espérance $\mathbb{E}(K) = n$ Variance $V(K) = 2n$

— Si $K_1 \sim \chi^2_n$ et $K_2 \sim \chi^2_m$ sont indépendante alors

$$K_1 + K_2 \sim \chi^2_{n+m}$$

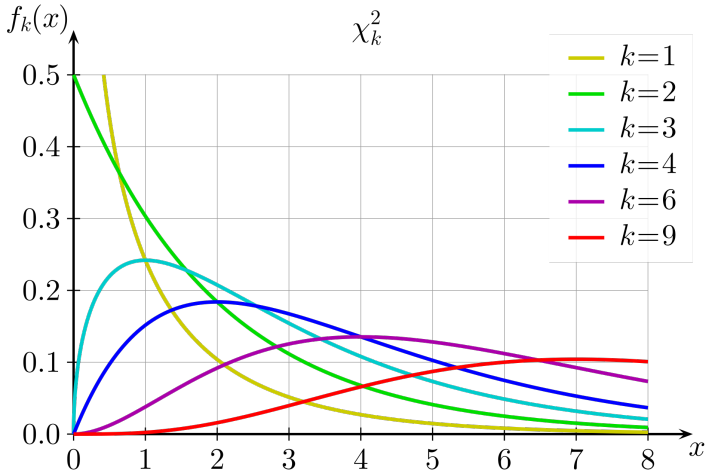


FIGURE 2.2 – Illustration de la densité de la loi du chi-deux en fonction du degré de liberté

2.1.3 Loi de Student

— Loi de Student à n degrés de liberté

$$T \sim \mathcal{T}_n$$

— Si $X \sim \mathcal{N}(0,1)$ et $K \sim \chi^2_n$ sont indépendantes alors

$$\boxed{T = \frac{X}{\sqrt{K/n}} \sim \mathcal{T}_n}$$

— Densité

$$f_T(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} (1 + \frac{x^2}{n})^{\frac{n+1}{2}}$$

— Espérance $\mathbb{E}(T) = 0$ Variance $V(T) = \frac{n}{n-2}$

— Quand $n \to \infty$, $\mathcal{T}_n$ tend en loi vers $\mathcal{N}(0,1)$

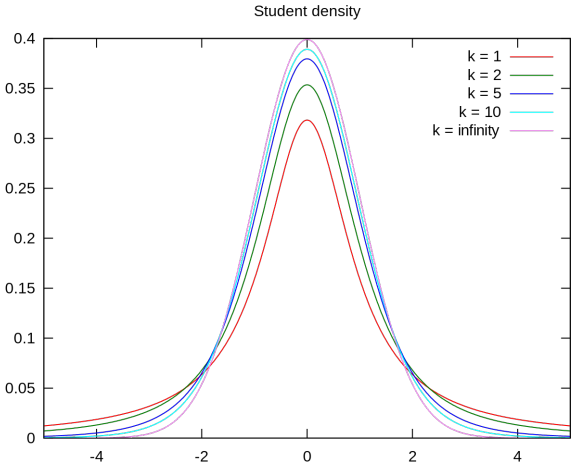


FIGURE 2.3 – Illustration de la densité de la loi de Student en fonction du degré de liberté

2.1.4 Loi de Fisher-Snedecor

— Loi de Fisher-Snedecor de paramètres m et n

$$F \sim \mathcal{F}_{m,n}$$

— Si $K_1 \sim \chi^2_m$ et $K_2 \sim \chi^2_n$ sont indépendantes alors

$$\boxed{F = \frac{K_1/m}{K_2/n} \sim \mathcal{F}_{m,n}}$$

— Densité

$$f_F(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{m/2} n^{n/2} \frac{x^{m/2-1}}{(mx+n)^{\frac{m+n}{2}}}$$

— Espérance $\mathbb{E}(F) = \frac{n}{n-2}$

— Variance $V(F) = \frac{2n^2(n+m-2)}{m(n-2)^2(n-4)}$

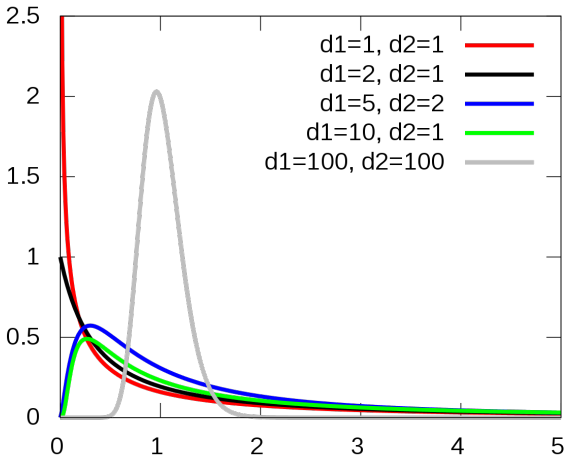


FIGURE 2.4 – Illustration de la densité de la loi de Fisher-Snedecor en fonction des paramètres

2.2 Statistiques classiques pour la moyenne et la variance

Moyenne et variance empirique Soient $(X_1, \cdots, X_n)$ un n-échantillon. De manière classique on défini la **moyenne empirique** par

$$\overline{X}_n = \frac{X_1 + \cdots X_n}{n}$$

et la **variance empirique** par

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

Proposition 2.1 (Théorème de Cochran)**.** *Soient $X_1, \cdots, X_n$ des v.a. iid de loi $\mathcal{N}(\mu, \sigma)$. Les variables aléatoires $\overline{X}_n$ et S_n^2 sont indépendantes et*

$$\boxed{n\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}.}$$

Remarque. Cette proposition est en fait une application d'une version simplifiée du théorème de Cochran.

Exercice (Preuve de la Proposition 2.1)**.** 1. Se ramener à des lois normales centrées réduites.
On pose ensuite $L_1 = \sqrt{\frac{1}{n}}(1, \cdots, 1)$ un vecteur ligne unitaire.
On complète ce vecteur en base orthonormale $L_1, L_2, \cdots, L_n$.
On note A la matrice orthogonale dont les lignes sont les L_i.
On a $AA^\top = A^\top A = I_n$ et on pose $Y = AX$ où $X = (X_j)_{j=1,n}$ et $Y = (Y_j)_{j=1,n}$ sont des vecteurs colonnes.

2. Montrer que $\overline{X}_n = \frac{Y_1}{\sqrt{n}}$ et que $S_n^2 = \frac{1}{n}\sum_{i=2}^{n} Y_i^2$.
3. Montrer que Y est un vecteur Gaussien et conclure.

Démonstration de la Proposition 2.1 - Solution de l'exercice. 1. On pose $X_i' = \frac{X_i - \mu}{\sigma}$ et on note $\overline{X}_n'$ et ${S_n'}^2$ les moyennes et variances empiriques associées qui vérifient ainsi

$$\overline{X}_n = \sigma\overline{X}_n' + \mu \quad \text{et} \quad S_n^2 = \sigma^2 {S_n'}^2.$$

Grâce à ces deux égalités, on peut donc supporter pour la suite de la preuve, que $\mu = 0$ et $\sigma = 1$.

2. — On a $Y_1 = L_1 X = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) = \sqrt{n}\overline{X}_n$ et donc $\overline{X}_n = \frac{Y_1}{\sqrt{n}}$.
   — $Y_1^2 + \cdots + Y_n^2 = Y^\top Y = (AX)^\top (AX) = X^\top A^\top AX = X^\top X = X_1^2 + \cdots + X_n^2$.
   — On déduit que $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\frac{Y_1}{\sqrt{n}})^2 = \frac{1}{n}\sum_{i=2}^{n} Y_i^2$.
3. Y est un vecteur Gaussien de loi $\mathcal{N}(0, AA^\top)$ car $Y = AX$ et X est une vecteur gaussien de loi $\mathcal{N}(0, I_n)$ (puisque les X_i sont des variables aléatoires iid de loi $\mathcal{N}(0, 1)$). Cela implique donc que les Y_i sont indépendants et en particulier, Y_1 est indépendant de $(Y_2, \cdots, Y_n)$ et donc $\overline{X}_n$ et S_n^2 sont indépendantes. Cela implique aussi que $nS_n^2 = \sum_{i=2}^{n} Y_i^2$ suit une loi du χ^2_{n-1}. (Pour rappel on s'était ramené au cas où $\sigma = 1$).

□

Lois de statistiques classiques On suppose encore des échantillons gaussiens pour les propositions suivantes.

Proposition 2.2 (Statistique de loi normale - Test de conformité de la moyenne, variance connue)**.**

$$\boxed{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)}$$

Proposition 2.3 (Statistique de loi de Student - Test de conformité de la moyenne, variance inconnue)**.**

$$\boxed{\sqrt{n-1}\frac{\overline{X}_n - \mu}{S_n} \sim \mathcal{T}_{n-1}}$$

Proposition 2.4 (Statistique de loi du χ^2 - Test de conformité de la variance)**.**

$$\boxed{n\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}}$$

Proposition 2.5 (Statistique de loi de Fisher - Test de comparaison de deux variances)**.**

$$\boxed{\frac{\widetilde{S}^2_{n,1}}{\widetilde{S}^2_{m,2}} \sim \mathcal{F}_{n-1,m-1}}$$

où $\widetilde{S}^2_{n,1} = \frac{n}{n-1}S^2_{n,1}$ et $\widetilde{S}^2_{n,2} = \frac{n}{n-1}S^2_{n,2}$ sont les variances corrigées de deux échantillons différents.

Les preuves des trois dernières propositions découlent directement de la proposition 2.1.

Grands échantillons Dans les propositions précédentes on supposait les v.a. gaussiennes. Pour des **grands échantillons** (typiquement $n > 30$) et des **v.a. non gaussiennes**, on a d'après le **théorème central limite** l'approximation :

$$\boxed{\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \approx \mathcal{N}(0,1)}$$

et on admet qu'on a aussi l'approximation :

$$\boxed{\sqrt{n-1}\frac{\overline{X}_n - \mu}{S_n} \approx \mathcal{T}_{n-1} \approx \mathcal{N}(0,1).}$$

Chapitre 3

Tests paramétriques

Dans ce chapitre, nous passerons en revue les tests suivants.

Test à un échantillon

- Conformité de la moyenne théorique μ à un standard μ_0
  - Variance connue
  - Variance inconnue
- Conformité de la variance théorique σ^2 à un standard σ_0^2
  - Moyenne connue
  - Moyenne inconnue
- Conformité d'une proportion théorique p à un standard p_0

Test à deux échantillons

- Comparaison de deux moyennes théoriques
  - Échantillons appariés
  - Échantillons non appariés
    - Variances connues
    - Variances inconnues égales
    - Variances inconnues différentes
- Comparaison de deux variances théoriques

3.1 Test à un échantillon

Pour un n échantillon $(X_1, \cdots, X_n)$, nous noterons la moyenne empirique

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

la variance empirique

$$S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$$

et la variance empirique corrigée

$$\widetilde{S}_n^2 = \frac{n}{n-1}S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2.$$

Lorsqu'on connait la moyenne théorique $\mathbb{E}[X_i] = \mu_0$ on pose

$$\hat{S}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_0)^2.$$

3.1.1 Conformité de la moyenne, variance connue ($\mathcal{N}(0,1)$)

— **Test de conformité** de la moyenne μ à un standard μ_0, variance $\sigma = \sigma_0$ connue

— **Loi de l'échantillon** $X_i \sim \mathcal{N}(\mu, \sigma_0^2)$ (paramètre μ inconnu)

— **Hypothèses**

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

ou, selon les cas, $H_1 : \mu > \mu_0$ ou encore $H_1 : \mu < \mu_0$.

— **Statistique et loi de la statistique sous** H_0

$$\boxed{S = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0} \sim \mathcal{N}(0,1) \quad \text{sous } H_0}$$

Si $n > 30$ ce test robuste : on peut se passer de l'hypothèse de normalité grâce au TCL.

— **Niveau** $\alpha = \mathbb{P}(S \in \mathcal{R}_\alpha | H_0)$

— **Région de rejet** $\mathcal{R}_\alpha$ et **Seuil** s_α
On définit le **quantile** de la loi normale q_x par $\mathbb{P}(N < q_x) = x$ pour $N \sim \mathcal{N}(0,1)$.

- Pour $H_1 : \mu \neq \mu_0$
  Zone bilatérale $\mathcal{R}_\alpha =]-\infty, -s_\alpha] \cup [s_\alpha, \infty[$ et on a $s_\alpha = q_{1-\alpha/2}$
- Pour $H_1 : \mu > \mu_0$
  Zone unilatérale à droite $\mathcal{R}_\alpha = [s_\alpha, \infty[$ et on a $s_\alpha = q_{1-\alpha}$.
- Pour $H_1 : \mu < \mu_0$
  Zone unilatérale à gauche $\mathcal{R}_\alpha =]-\infty, s_\alpha]$ et on a $s_\alpha = q_\alpha$.

3.1.2 Conformité de la moyenne, variance inconnue (Test de Student)

— **Test de conformité** de la moyenne μ à un standard μ_0, variance σ inconnue

— **Loi de l'échantillon** $X_i \sim \mathcal{N}(\mu, \sigma^2)$ (paramètre μ et σ inconnues)

— **Hypothèses**

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

ou, selon les cas, $H_1 : \mu > \mu_0$ ou encore $H_1 : \mu < \mu_0$.

— **Statistique et loi de la statistique sous** H_0

$$\boxed{T = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\widetilde{S}_n} = \sqrt{n-1}\frac{\overline{X}_n - \mu_0}{S_n} \sim \mathcal{T}_{n-1} \quad \text{sous } H_0}$$

Si $n > 30$ ce test robuste : on peut se passer de l'hypothèse de normalité et $\mathcal{T}_{n-1} \approx \mathcal{N}(0,1)$.

— **Niveau** $\alpha = \mathbb{P}(T \in \mathcal{R}_\alpha | H_0)$

— **Région de rejet** $\mathcal{R}_\alpha$ et **Seuil** s_α
On définit le **quantile** de la loi de Student $t_{n-1,x}$ par $\mathbb{P}(T < t_{n-1,x}) = x$ pour $T \sim \mathcal{T}_{n-1}$.

- Pour $H_1 : \mu \neq \mu_0$
  Zone bilatérale $\mathcal{R}_\alpha =]-\infty, -s_\alpha] \cup [s_\alpha, \infty[$ et on a $s_\alpha = t_{n-1,1-\alpha/2}$
- Pour $H_1 : \mu > \mu_0$
  Zone unilatérale à droite $\mathcal{R}_\alpha = [s_\alpha, \infty[$ et on a $s_\alpha = t_{n-1,1-\alpha}$.
- Pour $H_1 : \mu < \mu_0$
  Zone unilatérale à gauche $\mathcal{R}_\alpha =]-\infty, s_\alpha]$ et on a $t_\alpha = t_{n-1,\alpha}$.

3.1.3 Conformité de la variance, moyenne connue (χ^2_n)

— **Test de conformité** de la variance σ à un standard σ_0, moyenne $\mu = \mu_0$ connue

— **Loi de l'échantillon** $X_i \sim \mathcal{N}(\mu_0, \sigma^2)$ (paramètre σ inconnu)

— **Hypothèses**

$$H_0 : \sigma = \sigma_0 \text{ versus } H_1 : \sigma \neq \sigma_0$$

ou, selon les cas, $H_1 : \sigma > \sigma_0$ ou encore $H_1 : \sigma < \sigma_0$.

— **Statistique et loi de la statistique sous** H_0

$$\boxed{K = n\frac{\hat{S}_n^2}{\sigma_0^2} = \sum_{k=1}^{n} \frac{(X_k - \mu_0)^2}{\sigma_0^2} \sim \chi^2_n \quad \text{sous } H_0}$$

— **Niveau** $\alpha = \mathbb{P}(K \in \mathcal{R}_\alpha | H_0)$

— **Région de rejet** $\mathcal{R}_\alpha$ et **Seuil** s_α
On définit le **quantile** de la loi de Student $k_{n,x}$ par $\mathbb{P}(K < k_{n,x}) = x$ pour $K \sim \chi^2_n$.

- Pour $H_1 : \sigma \neq \sigma_0$
  Zone bilatérale $\mathcal{R}_\alpha =]-\infty, -s_\alpha] \cup [s_\alpha, \infty[$ et on a $s_\alpha = k_{n,1-\alpha/2}$
- Pour $H_1 : \sigma > \sigma_0$
  Zone unilatérale à droite $\mathcal{R}_\alpha = [s_\alpha, \infty[$ et on a $s_\alpha = k_{n,1-\alpha}$
- Pour $H_1 : \sigma < \sigma_0$
  Zone unilatérale à gauche $\mathcal{R}_\alpha =]-\infty, s_\alpha]$ et on a $t_\alpha = k_{n,\alpha}$.

3.1.4 Conformité de la variance, moyenne inconnue (χ^2_{n-1})

— **Test de conformité** de la variance σ à un standard σ_0, moyenne μ inconnue

— **Loi de l'échantillon** $X_i \sim \mathcal{N}(\mu, \sigma^2)$ (paramètre μ et σ inconnues)

— **Hypothèses**

$$H_0 : \sigma = \sigma_0 \text{ versus } H_1 : \sigma \neq \sigma_0$$

ou, selon les cas, $H_1 : \sigma > \sigma_0$ ou encore $H_1 : \sigma < \sigma_0$.

— **Statistique et loi de la statistique sous** H_0

$$\boxed{K = n\frac{S_n^2}{\sigma_0^2} = (n-1)\frac{\widetilde{S}_n^2}{\sigma_0^2} \sim \chi^2_{n-1} \quad \text{sous } H_0}$$

— **Niveau** $\alpha = \mathbb{P}(K \in \mathcal{R}_\alpha | H_0)$

— **Région de rejet** $\mathcal{R}_\alpha$ et **Seuil** s_α
On définit le **quantile** de la loi de Student $k_{n-1,x}$ par $\mathbb{P}(K < k_{n-1,x}) = x$ pour $K \sim \chi^2_{n-1}$.

- Pour $H_1 : \sigma \neq \sigma_0$
  Zone bilatérale $\mathcal{R}_\alpha =]-\infty, -s_\alpha] \cup [s_\alpha, \infty[$ et on a $s_\alpha = k_{n-1,1-\alpha/2}$
- Pour $H_1 : \sigma > \sigma_0$
  Zone unilatérale à droite $\mathcal{R}_\alpha = [s_\alpha, \infty[$ et on a $s_\alpha = k_{n-1,1-\alpha}$
- Pour $H_1 : \sigma < \sigma_0$
  Zone unilatérale à gauche $\mathcal{R}_\alpha =]-\infty, s_\alpha]$ et on a $s_\alpha = k_{n-1,\alpha}$.

3.1.5 Conformité d'une proportion ($\mathcal{N}(0,1)$)

Test semblable à un test de conformité de la moyenne. On suppose que $n > 30$, $np > 5$ et $n(1-p) > 5$.

— **Test de conformité** d'une proportion p à un standard p_0

— **Loi de l'échantillon** $X_i \sim \mathcal{B}(1,p)$ et $X_1 + \cdots + X_n \sim \mathcal{B}(n,p) \approx \mathcal{N}(np, np(1-p))$ (TCL).

— **Hypothèses**

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

ou, selon les cas, $H_1 : p > p_0$ ou encore $H_1 : p < p_0$.

— **Statistique et loi de la statistique sous** H_0

$$\boxed{S = \sqrt{n}\frac{\overline{X}_n - p_0}{\sqrt{p_0(1-p_0)}} \approx \mathcal{N}(0,1) \quad \text{sous } H_0}$$

car on suppose que $n > 30$, $np > 5$ et $n(1-p) > 5$.

— **Niveau** $\alpha = \mathbb{P}(S \in \mathcal{R}_\alpha | H_0)$

— **Région de rejet** $\mathcal{R}_\alpha$ et **Seuil** s_α
On définit le **quantile** de la loi normale q_x par $\mathbb{P}(N < q_x) = x$ pour $N \sim \mathcal{N}(0,1)$.

- Pour $H_1 : p \neq p_0$
  Zone bilatérale $\mathcal{R}_\alpha =]-\infty, -s_\alpha] \cup [s_\alpha, \infty[$ et on a $s_\alpha = q_{1-\alpha/2}$
- Pour $H_1 : p > p_0$
  Zone unilatérale à droite $\mathcal{R}_\alpha = [s_\alpha, \infty[$ et on a $s_\alpha = q_{1-\alpha}$.
- Pour $H_1 : p < p_0$
  Zone unilatérale à gauche $\mathcal{R}_\alpha =]-\infty, s_\alpha]$ et on a $s_\alpha = q_\alpha$.

3.2 Test à deux échantillons

On considère deux échantillons $(X_1, \cdots, X_n)$ et $(Y_1, \cdots, Y_m)$ tels que $\mu_1 = \mathbb{E}[X_i]$, $\mu_2 = \mathbb{E}[Y_i]$, $\sigma_1^2 = Var[X_i]$, $\sigma_2^2 = Var[Y_i]$. On note $\widetilde{S}^2_{X,n}$ et $\widetilde{S}^2_{Y,m}$ les variances empiriques corrigées.

3.2.1 Comparaison des moyennes avec échantillons appariés

Les échantillons $(X_1, \cdots, X_n)$ et $(Y_1, \cdots, Y_n)$ (de même taille) sont dits appariés si pour tout i, X_i et Y_i sont issues d'une mesure effectuée à deux instants différents sur un même individu. Les éléments des échantillons vont par paires.

Exemple (Taux d'insuline)**.** On veut tester l'effet d'un traitement sur l'insuline.

— Le taux d'insuline est mesuré sur 30 patients avant et après le traitement médical. Les données sont donc organisées par paires (chaque patient est associé à deux mesures). Dans cette situation, il s'agit d'échantillons appariés.

— Le taux d'insuline est mesuré sur 30 patients recevant un placebo et 30 autres patients recevant un traitement médical. Dans ce cas, toutes les mesures sont indépendantes (chaque patient n'est associé qu'à une mesure unique). Il s'agit d'échantillons non appariés.

Test Dans le cas apparié, on travaillera sur l'échantillon (unique) composé des variables aléatoires indépendantes $(X_1 - Y_1, \cdots, X_n - Y_n)$. Le **test de comparaison** de μ_1 et μ_2 **se ramène alors à un test de conformité** de la moyenne $\mu_1 - \mu_2$ avec la valeur 0.

3.2.2 Comparaison des moyennes avec échantillons non appariés

Variances connues

— **Test de comparaison** des deux moyennes μ_1 et μ_2 avec les variances σ_1^2 et σ_2^2 connues.

— **Loi des échantillons** $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$ (paramètres μ_1 et μ_2 inconnus)

— **Hypothèses**

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2$$

ou, selon les cas, $H_1 : \mu_1 > \mu_2$ ou encore $H_1 : \mu_1 < \mu_2$.

— **Statistique et loi de la statistique sous** H_0

$$\boxed{S = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0,1) \quad \text{sous } H_0}$$

Si $m > 30$ et $n > 30$ ce test robuste : on peut se passer de l'hypothèse de normalité.

— La zone de rejet et la valeur seuil se détermine de manière classique avec la loi normale.

Variances inconnues égales

— **Test de comparaison** des deux moyennes μ_1 et μ_2 avec les variances $\sigma_1^2 = \sigma_2^2$ inconnue.

— **Loi des échantillons** $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$ (paramètres μ_1 et μ_2 inconnus)

— **Hypothèses**

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2$$

ou, selon les cas, $H_1 : \mu_1 > \mu_2$ ou encore $H_1 : \mu_1 < \mu_2$.

— **Statistique et loi de la statistique sous** H_0

$$\boxed{S = \frac{\overline{X}_n - \overline{Y}_m}{\widetilde{S}_{n,m}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathcal{T}_{n+m-2} \quad \text{sous } H_0}$$

où

$$\widetilde{S}_{n,m}^2 = \frac{(n-1)\widetilde{S}_{X,n}^2 + (m-1)\widetilde{S}_{Y,m}^2}{n+m-2}.$$

Remarque. Si $m > 30$ et $n > 30$ ce test robuste, on peut se passer de l'hypothèse de normalité et on considère que $S \approx \mathcal{N}(0,1)$ sous H_0.

— La zone de rejet et la valeur seuil se détermine de manière classique avec la loi normale.

Variances inconnues inégales (Test de Welch)

— **Test de comparaison** des deux moyennes μ_1 et μ_2 avec les variances $\sigma_1^2 \neq \sigma_2^2$ inconnues.

— **Loi des échantillons** $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$ (paramètres μ_1 et μ_2 inconnus)

— **Hypothèses**

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2$$

ou, selon les cas, $H_1 : \mu_1 > \mu_2$ ou encore $H_1 : \mu_1 < \mu_2$.

— **Statistique et loi de la statistique sous H_0**

$$\boxed{S = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\widetilde{S}^2_{X,n}}{n} + \frac{\widetilde{S}^2_{Y,m}}{m}}} \sim \mathcal{T}_\nu \quad \text{sous } H_0}$$

où ν est l'entier le plus de proche de

$$\nu = \frac{(\frac{\widetilde{S}^2_{X,n}}{n} + \frac{\widetilde{S}^2_{Y,m}}{m})^2}{\frac{\widetilde{S}^4_{X,n}}{n^2(n-1)} + \frac{\widetilde{S}^4_{Y,m}}{m^2(m-1)}}.$$

Remarque. Si $m > 30$ et $n > 30$ ce test robuste, on peut se passer de l'hypothèse de normalité et on considère que $S \approx \mathcal{N}(0,1)$ sous H_0.

— La zone de rejet et la valeur seuil se détermine de manière classique avec la loi normale.

3.2.3 Comparaison des variances (Test de Fisher)

— **Test de comparaison** des deux variances σ_1 et σ_2 (avec les moyennes μ_1 et μ_2 inconnues)

— **Loi des échantillons** $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$

— **Hypothèses**

$$H_0 : \sigma_1 = \sigma_2 \text{ versus } H_1 : \sigma_1 \neq \sigma_2$$

ou, selon les cas, $H_1 : \mu_1 > \mu_2$ ou encore $H_1 : \mu_1 < \mu_2$.

— **Statistique et loi de la statistique sous H_0**

$$\boxed{S = \frac{\widetilde{S}^2_{1,n}}{\widetilde{S}^2_{2,m}} \sim \mathcal{F}_{n-1,m-1} \quad \text{sous } H_0}$$

— La zone de rejet et la valeur seuil se détermine de manière classique avec la loi de Fisher.

Chapitre 4

Tests non paramétriques

4.1 Généralités sur le test du χ^2

Test inventé par Karl Pearson.

Son **objectif** est de déterminer si des effectifs empiriques sont conformes à des effectifs théoriques sous l'hypothèse nulle.

Il y a trois principaux types de tests :

— **Adéquation** (ou ajustement) à une loi donnée
ex : Est-ce que les généotypes observés sur une population suivent les lois de Mendel sur l'hérédité ?

— **Homogénéité** (ou comparaison) entre deux lois
ex : Est-ce que la distribution des groupes sanguins sont les même dans deux pays différents ?

— **Indépendance**
ex : Est-ce qu'il y a indépendance entre la couleur des yeux et la couleur des cheveux ?

Remarque. Le test d'homogénéité est en fait un cas particulier du test d'indépendance.

Construction du test du χ^2 La statistique du test χ^2 mesure l'écart entre les distributions des **effectifs théoriques** t_i et des **effectifs observés** n_i :

Définition 4.1 (Statistique de Pearson/ Distance du χ^2)**.**

$$s_{obs} = \chi^2_{obs} = \sum_i \frac{(n_i - t_i)^2}{t_i} = n \sum_i \frac{(\hat{p}_i - p_i)^2}{p_i}$$

où on a noté

— $\boxed{p_i = t_i/n}$ proportion théorique

— $\hat{p}_i = n_i/n$ proportion empirique

— $n = \sum_i n_i$ effectif total

Remarque (Conditions de validité du test)**.** Le test est valable si

— $n > 50$

— $t_i \geqslant 5 \quad \forall i$ (sinon il faut **regrouper les classes adjacentes**)

4.2 Tests du χ^2

4.2.1 Adéquation à une loi donnée

Soit X une variable aléatoire à valeur dans d classes $\{x_1, \cdots, x_d\}$.
On considère la loi de probabilité $p = (p_1, \cdots, p_d)$ fixée et connue.
On considère les hypothèses suivantes

— $H_0 : \mathcal{L}(X) = p$ (i.e. $\mathbb{P}(X = x_i) = p_i$)
La distribution empirique conforme à distribution théorique

— $H_1 : \mathcal{L}(X) \neq p$
La distribution observée ne s'ajuste pas à la distribution théorique

Soit $(X_1, \cdots, X_n)$ un n-échantillon de loi $X_i \sim X$. On définit

— $N_i = \sum_{k=1}^{n} \mathbb{1}_{X_k = x_i}$ effectif empirique,
— $\hat{p}_i = \frac{N_i}{n}$ la proportion empirique (estimateur de p_i)

Proposition 4.2 (Théorème de Pearson)**.**

$$\sum_{i=1}^{d} \frac{(N_i - np_i)^2}{np_i} = n \sum_{i=1}^{d} \frac{(\hat{p}_i - p_i)^2}{p_i} \xrightarrow[n\to\infty]{(loi)} \chi^2(d-1) \ \boxed{sous\ H_0}$$

$$\xrightarrow[n\to\infty]{} \infty \ \boxed{sous\ H_1}$$

Idée de démonstration du théorème de Pearson. On a $\mathbb{E}[N_i] = np_i$ et le TCL implique

$$Y_i = \frac{N_i - \mathbb{E}[N_i]}{\sqrt{n}} \xrightarrow[n\to\infty]{(loi)} \mathcal{N}$$

par le **TCL** et donc $\sum_{i=1}^{d} Y_i^2 \xrightarrow[n\to\infty]{(loi)} \chi^2(d-1)$

La perte d'un degré de liberté vient du fait que les Y_i ne sont pas indépendants car $\sum_{i=1}^{d} N_i = n$ (ce qui donne **une contrainte**). La démonstration utilise le théorème de Cochran. □

Test d'adéquation à une loi donnée

Mise en œuvre du test

— **Calculer** la statistique

$$\boxed{\chi^2_{obs} = \sum_{i=1}^{d} \frac{(n_i - np_i)^2}{np_i}}$$

— **Fixer** : seuil α, valeur seuil s_α, région de rejet unilatérale à droite $\mathcal{R}_\alpha = [s_\alpha, \infty)$ associé au quantile du chi-deux du bon degré de liberté.
— **Décider** :
  — On rejette H_0 si $\chi^2_{obs} \geqslant s_\alpha$
  — On conserve H_0 si $\chi^2_{obs} < s_\alpha$

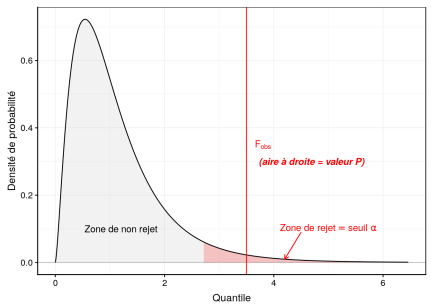


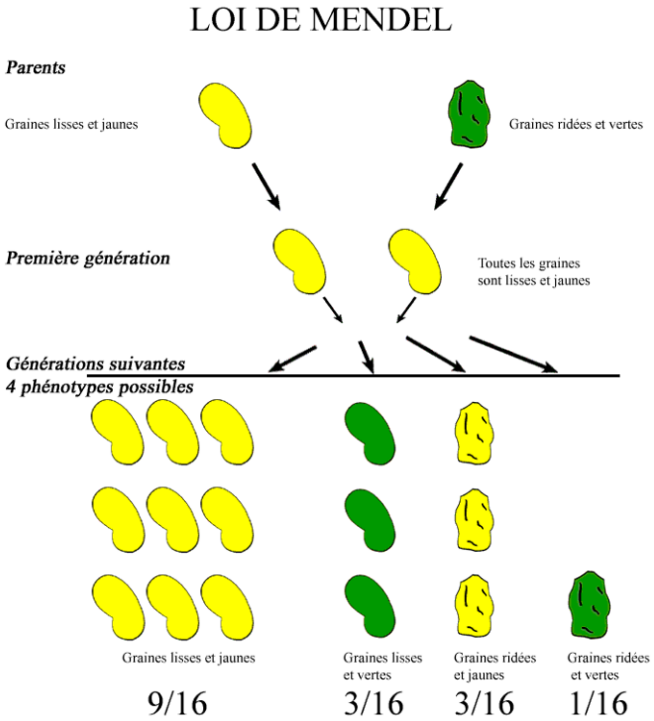
Tableau Lorsqu'on réalise un test du chi-deux, il faut toujours remplir le tableau ci dessous afin de calculer simplement χ^2_{obs}.

Classes / Effectifs	x_1	x_2	$\cdots$	x_d	Total
Empirique	n_1	n_2	$\cdots$	n_d	n
Théorique	np_1	np_2	$\cdots$	np_d	n

Exercice (Loi d'hérédité de Mendel)**.** On observe à la deuxième génération une population de pois avec :

- 76 lisses et jaunes
- 29 lisses et vertes
- 20 ridés jaunes
- 10 ridés vertes

La loi de Mendel (cf schéma ci dessous) est-elle vérifiée ?



Test d'adéquation à une famille de loi On cherche à montrer la distribution empirique est conforme à une famille de loi indexée par $\theta \in \mathbb{R}^k$.

Soit $\hat{\theta}_n$ un estimateur du paramètre de dimension k (c'est à dire qu'il y a k paramètre dans $\mathbb{R}$ et k estimateurs).

Proposition 4.3 (Statistique du chi-deux avec estimation de paramètre)**.**

$$\sum_{i=1}^{d} \frac{(N_i - np_i(\hat{\theta}_n))^2}{np_i(\hat{\theta}_n)} \xrightarrow[n\to\infty]{(loi)} \chi^2(d-1-k) \quad \boxed{sous\ H_0}$$

Remarque. Il y a k contraintes en plus donc $-k$ degrés de liberté (ddl).

4.2.2 Homogénéité

Soient q échantillons

$$\begin{cases} X_1 = (X_{11}, \cdots, X_{1n_1}) & n_1\text{-échantillon} \\ \vdots & \\ X_q = (X_{q1}, \cdots, X_{qn_q}) & n_q\text{-échantillon} \end{cases}$$

à valeurs dans r classes $\{x_1, \cdots, x_r\}$ et tels que $n_1 + \cdots + n_q = n$.

On cherche à tester l'hypothèse que les échantillons ont la même loi.

— $H_0 : \mathcal{L}(X_1) = \cdots = \mathcal{L}(X_q)$

— H_1 : un échantillon (au moins) n'a pas la même loi que les autres

On pose pour tout $i = 1, \cdots, q$,

$$(N_{i1}, \cdots, N_{ir})$$

tel que pour $j = 1, \cdots, r$,

$$N_{ij} = \sum_{k=1}^{n_i} \mathbb{1}_{X_{ik}=C_j}$$

est le nombre d'éléments dans la classe j du $i$ème échantillon. On note n_{ij} la réalisation de N_{ij} et on remarque que $n_i = n_{i1} + \cdots + n_{ir}$. On note

$$\widetilde{n}_j = n_{1j} + \cdots + n_{qj}$$

le nombre total d'élément dans la classe j pour $j = 1, \cdots, r$ et on a $\widetilde{n}_1 + \cdots + \widetilde{n}_r = n$. Sous H_0 on estime la loi théorique $p = (p_1, \cdots, p_r)$ en prenant pour $j = 1, \cdots, r$ l'estimateur

$$\hat{P}_j = \frac{1}{n}(N_{1j} + \cdots N_{qj})$$

et on note $\hat{p}_j$ sa réalisation. On estime les effectifs théoriques par

$$t_{ij} = n_i \hat{p}_j = \frac{n_i \widetilde{n}_j}{n}.$$

On a alors la statistique

$$\boxed{\chi^2_{obs} = \sum_{i=1}^{q}\sum_{j=1}^{r} \frac{(n_{ij} - t_{ij})^2}{t_{ij}} = \sum_{i=1}^{q}\sum_{j=1}^{r} \frac{(n_{ij} - \frac{n_i \widetilde{n}_j}{n})^2}{\frac{n_i \widetilde{n}_j}{n}} \sim \chi^2_{(q-1)(r-1)}.}$$

Remarque. On remarque qu'il y a $(q-1)(r-1) = qr - q - (r-1)$ degrés de liberté à la loi du chi-deux. Cela vient du fait qu'il y a q contraintes imposées par la taille de chaque échantillon et qu'on a du estimer $(r-1)$ paramètres $\hat{p_j}$ (le dernier estimateur ne compte pas car il est est imposé par $\hat{p_1} + \cdots + \hat{p_r} = 1$).

Tableaux empiriques et théoriques Lorsqu'on réalise un test du chi-deux, il faut toujours remplir les deux tableaux (empirique et théorique) ci dessous afin de calculer simplement χ^2_{obs}.

Classes / Effectifs empiriques	x_1	x_2	$\cdots$	x_r	Total
X_1	n_{11}	n_{12}	$\cdots$	n_{1r}	n_1
$\vdots$	$\vdots$	$\vdots$	n_{ij}	$\vdots$	$\vdots$
X_q	n_{q1}	n_{q2}	$\cdots$	n_{qr}	n_q
Total	$\widetilde{n}_1$	$\widetilde{n}_2$	$\cdots$	$\widetilde{n}_r$	n

Classes / Effectifs théoriques	x_1	x_2	$\cdots$	x_r	Total
X_1	$n_1\hat{p}_1$	$n_1\hat{p}_2$	$\cdots$	$n_1\hat{p}_r$	n_1
$\vdots$	$\vdots$	$\vdots$	$n_i\hat{p}_j = \frac{n_i\widetilde{n}_j}{n}$	$\vdots$	$\vdots$
X_q	$n_q\hat{p}_1$	$n_q\hat{p}_2$	$\cdots$	$n_q\hat{p}_r$	n_q
Total	$\widetilde{n}_1$	$\widetilde{n}_2$	$\cdots$	$\widetilde{n}_r$	n

4.2.3 Indépendance

Soit un échantillon de taille n

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

où les X_i sont à valeurs dans r classes $\{x_1, \cdots, x_q\}$ et les Y_i sont à valeurs dans q classes $\{y_1, \cdots, y_r\}$.

Les (X_i, Y_i) sont iid et on cherche à tester l'hypothèse que X_i est indépendant de Y_i, c'est à dire que les échantillons $(X_1, \cdots, X_n)$ et $(Y_1, \cdots, Y_n)$ sont indépendants.

— $H_0 : X \perp\!\!\!\perp Y$

— $H_1 : X \not\perp\!\!\!\perp Y$

On pose pour tout $(i, j) \in \{1, \cdots, q\} \times \{1, \cdots, r\}$,

$$N_{ij} = \sum_{k=1}^{n} \mathbb{1}_{(X_k, Y_k)=(x_i, y_j)}$$

qui est le nombre d'éléments de l'échantillon dans la classe (x_i, y_j) de l'échantillon. On note n_{ij} la réalisation de N_{ij}. On note

$$n_i = n_{i1} + \cdots + n_{ir}$$

le nombre de X_k dans la classe x_i et

$$\widetilde{n}_j = n_{1j} + \cdots + n_{qj}$$

le nombre de Y_k dans la classe y_j. On a $n_1 + \cdots + n_q = \widetilde{n}_1 + \cdots + \widetilde{n}_r = n$. On note $p_i = \mathbb{P}(X_k = x_i)$ et $p'_j = \mathbb{P}(Y_k = y_j)$. Sous H_0 on a alors

$$\mathbb{P}((X_k, Y_k) = (x_i, y_j)) = p_i p'_j.$$

On estime sous H_0 la loi théorique en prenant pour $i = 1, \cdots, q$ et $j = 1, \cdots, r$ les estimateurs

$$\hat{p}_i = \frac{1}{n}(n_{i1} + \cdots n_{ir}) = \frac{n_i}{n} \quad \text{et} \quad \hat{p'_j} = \frac{1}{n}(n_{1j} + \cdots n_{qj}) = \frac{\widetilde{n}_j}{n}.$$

On estime les effectifs théoriques par

$$t_{ij} = n\hat{p}_i\hat{p}'_j = \frac{n_i\widetilde{n}_j}{n}.$$

On a alors la statistique

$$\boxed{\chi^2_{obs} = \sum_{i=1}^{q}\sum_{j=1}^{r} \frac{(n_{ij} - t_{ij})^2}{t_{ij}} = \sum_{i=1}^{q}\sum_{j=1}^{r} \frac{(n_{ij} - \frac{n_i\widetilde{n}_j}{n})^2}{\frac{n_i\widetilde{n}_j}{n}} \sim \chi^2_{(q-1)(r-1)}.}$$

Remarque. On remarque qu'il y a $(q-1)(r-1) = qr - 1 - (q-1) - (r-1)$ degrés de liberté à la loi du chi-deux. Cela vient du fait qu'il y a une contrainte sur la taille de l'échantillon (n) et qu'on a estimé $q-1$ paramètres $\hat{p}_i$ et $r-1$ paramètres $\hat{p'_j}$ (il y en a un de moins à chaque fois à cause des relations $\hat{p_1} + \cdots + \hat{p_q} = \hat{p'_1} + \cdots + \hat{p'_r} = 1$).

Tableaux empiriques et théoriques Lorsqu'on réalise un test du chi-deux, il faut toujours remplir les deux tableaux (empirique et théorique) ci dessous afin de calculer simplement χ^2_{obs}.

Empirique / Classes	y_1	y_2	$\cdots$	y_r	Total
x_1	n_{11}	n_{12}	$\cdots$	n_{1r}	n_1
$\vdots$	$\vdots$	$\vdots$	n_{ij}	$\vdots$	$\vdots$
x_q	n_{q1}	n_{q2}	$\cdots$	n_{qr}	n_q
Total	$\widetilde{n}_1$	$\widetilde{n}_2$	$\cdots$	$\widetilde{n}_r$	n

Théorique / Classes	y_1	y_2	$\cdots$	y_r	Total
x_1	$n\hat{p}_1\hat{p}'_1$	$n\hat{p}_1\hat{p}'_2$	$\cdots$	$n\hat{p}_1\hat{p}'_r$	n_1
$\vdots$	$\vdots$	$\vdots$	$n_i\hat{p}_i\hat{p}'_j = \frac{n_i\widetilde{n}_j}{n}$	$\vdots$	$\vdots$
x_q	$n\hat{p}_q\hat{p}'_1$	$n\hat{p}_q\hat{'}p_2$	$\cdots$	$n\hat{p}_q\hat{p}'_r$	n_q
Total	$\widetilde{n}_1$	$\widetilde{n}_2$	$\cdots$	$\widetilde{n}_r$	n

Remarque (Cas particulier du test d'homogénéité)**.** En fait, tester l'homogénéité de $(X_1, \cdots, X_{n_1})$ et $(Y_1, \cdots, Y_{n_2})$ revient à tester l'indépendance de l'échantillon

$$((X_1, 0), \cdots, (X_{n_1}, 0), (Y_1, 1), \cdots, (Y_{n_2}, 1)).$$

En effet, les deux coordonnées seront indépendantes si et seulement si X et Y ont même loi.

4.3 Autres tests non paramétriques

Il existe de nombreux autres tests paramétriques comme

— Test de normalité de Shapiro-Wilk ou celui de Shapiro-Francia

— Test d'ajustement à une loi de Kolmogorov (utilise des fonctions de répartition et le théorème de Glivenko-Cantelli)

— Test d'homogénéité de Kolmogorov-Smirnov

Intervalles de confiance

Définition (Intervalle de confiance)**.** Soit $(X_1, \cdots, X_n)$ un échantillon dont la loi dépend d'un paramètre θ. Un intervalle de confiance de θ au niveau $1-\alpha$ est un intervalle **aléatoire** $I_\alpha(X_1, \cdots, X_n)$ qui dépend de l'échantillon et tel que

$$\forall \theta \quad \mathbb{P}_\theta(\theta \in I_\alpha) \geqslant 1-\alpha.$$

On dit que $I_{n,\alpha}$ est un intervalle de confiance asymptotique lorsque

$$\forall \theta \quad \lim_{n\to\infty} \mathbb{P}_\theta(\theta \in I_{n,\alpha}) \geqslant 1-\alpha.$$

Remarque (Rappels)**.** Soit $(X_1, \cdots, X_n)$ un échantillon de loi $\mathcal{N}(\mu, \sigma^2)$. Alors

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

or $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$ donc

$$\frac{\overline{X}_n - \mu}{\sqrt{\mathrm{Var}(\overline{X}_n)}} \sim \mathcal{N}(0,1).$$

Et grâce au théorème central limite, même sans l'hypothèse de normalité on a donc $\frac{\overline{X}_n - \mu}{\sqrt{\mathrm{Var}(\overline{X}_n)}}$ qui converge en loi vers $\mathcal{N}(0,1)$.

Exemple (Intervalle de confiance pour la moyenne μ)**.** Soit un échantillon $(X_1, \cdots, X_n)$ de moyenne μ et de variance σ^2. On note le quantile de la loi normale q_x où $\mathbb{P}(\mathcal{N}(0,1) \leqslant q_x) = x$ et de même on note t_x le quantile de la loi de Student.

1. σ^2 connue
   - Hypothèse de normalité, échantillon de loi $\mathcal{N}(\mu, \sigma^2)$.
     Grâce au rappel ci dessus on remarque que

$$\mathbb{P}\left(\left|\frac{\overline{X}_n - \mu}{\sqrt{\mathrm{Var}(\overline{X}_n)}}\right| \leqslant q_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

     et donc

$$\mathbb{P}(\mu \in [\overline{X}_n - q_{1-\frac{\alpha}{2}}\sqrt{\mathrm{Var}(\overline{X}_n)}, \overline{X}_n + q_{1-\frac{\alpha}{2}}\sqrt{\mathrm{Var}(\overline{X}_n)}]) = 1-\alpha$$

     Or $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$. On a ainsi l'intervalle de confiance

$$\boxed{I_\alpha = [\overline{X}_n - q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]}$$

— Sans hypothèse de normalité on obtient grâce au théorème central limite que le même intervalle $I_\alpha = [\overline{X}_n - q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$ est un intervalle de confiance asymptotique.

2. σ^2 inconnue

— Hypothèse de normalité, échantillon de loi $\mathcal{N}(\mu, \sigma^2)$
On rappelle qu'on a démontré dans la partie sur les tests que

$$\sqrt{n-1}\frac{\overline{X}_n - \mu}{S_n} \sim \mathcal{T}_{n-1}.$$

On obtient ainsi l'intervalle de confiance

$$\boxed{I_\alpha = [\overline{X}_n - t_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n-1}}, \overline{X}_n + t_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n-1}}]}$$

— Sans hypothèse de normalité, on admet comme dans la partie sur les tests que pour n grand on a

$$\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \approx \mathcal{N}(0,1)$$

et on obtient ainsi un intervalle de confiance asymptotique

$$\boxed{I_\alpha = [\overline{X}_n - q_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}, \overline{X}_n + q_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}]}$$

Remarque (Autres intervalles de confiance)**.** D'autres types d'intervalles de confiance existent fondés sur des inégalités comme celle de Bienaymé-Tchebychev ou celle Hoeffding.

Remarque (Tests et intervalles de confiance)**.** Il y a une dualité entre tests et intervalles de confiance car avec tout intervalle de confiance on peut construire un test de conformité ou d'ajustement à un paramètre et réciproquement.

— Si I_α est un intervalle de confiance de niveau $1-\alpha$, il suffit de poser $\phi_{\theta_0}(X) = \mathbb{1}_{\theta_0 \notin I_\alpha}$ qui est un test de niveau α pour $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$.

— Réciproquement, si on dispose d'un test ϕ_{θ_0} de niveau α de $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ alors $I_\alpha = \{\theta_0 : \phi_{\theta_0}(X) = 0\}$ est un intervalle de confiance de niveau $1-\alpha$.

Deuxième partie

Sondages

Les notes qui suivent sont issues du polycopié de Mathieu Ribatet dans le cadre d'un cours intitulé *Sondages et Enquêtes* du Master MIND à l'Université de Montpellier 2. Elles sont fortement inspirées du livre de Yves Tillé *Théorie des Sondages : Échantillonnage et estimation en populations finies*. Ces notes seront utiles en complément du cours dispensé au tableau en classe.

Liste des symboles

$1_{\{x\in A\}}$	Variable indicatrice
$\hat{t}_{y,\pi}, \hat{\mu}_{y,\pi}$	π–estimateur du total t_y et de la moyenne μ_y
$\mathbb{E}(X)$	Espérance de la variable aléatoire X
$\mathcal{U}$	Population
$\mathcal{U}_h$	Strate de la population
$\mathcal{U}_i$	Grappe de la population
$\mathscr{L}$	Lagrangien du problème d'optimisation sous contraintes
$\mathscr{S}, \tilde{\mathscr{S}}$	Ensemble des échantillons non ordonnés sans remise et ordonnés avec remise
$\pi_k, \pi_{k\ell}$	Probabilités d'inclusion d'ordre un et deux
σ_y^2	Variance du caractère y sur la population
$\mathrm{Var}(X)$	Variance de la variable aléatoire X
k	Unité de la population
N	Taille de la population
N_h	Taille de la strate $\mathcal{U}_h$
N_i	Taille de la grappe $\mathcal{U}_i$
n_S, n	Taille de l'échantillon S
$p(\cdot)$	Plan de sondage
S	Échantillon
S_y^2	Variance corrigée du caractère y sur la population
t_y, μ_y	Total et moyenne du caractère y sur la population
$t_{y,h}, \mu_{y,h}$	Total et moyenne du caractère y sur la strate/grappe $\mathcal{U}_h$
y	Caractère défini sur la population

Chapitre 1

Formalisation mathématique d'un sondage

Ce chapitre pose les bases de la théorie des sondage en introduisant le vocabulaire, la notion d'aléatoire spécificique aux sondages et les estimateurs principaux.

1.1 Population, Caractère et Fonction d'intérêt

En sondage on s'intéresse à une **population** (ou **univers**) finie $\mathcal{U}$ constituée de N **unités** (ou **individus**) notées $u_1, \ldots, u_N$. On supposera que ces unités sont **identifiables** si chacune d'entre elle peut se voir attribuer un numéro d'identification **unique**. Ainsi par abus de notations, on écrira indifféremment

$$\mathcal{U} = \{u_1, \ldots, u_N\}, \qquad \mathcal{U} = \{1, \ldots, N\}.$$

Remarque. La définition de la population $\mathcal{U}$ est souvent problématique. Par exemple, pour l'étude des *habitants de plus de 18 ans d'un pays* la population n'est pas parfaitement identifiée si l'on ne suppose pas une date de référence pour l'âge et si l'on ne précise pas certains critères comme : France métropolitaine, Résidants ou Nationalité, ...

L'objectif d'un sondage ne porte pas sur les unités elles mêmes mais plutôt sur un **caractère** y qui est mesuré sur chaque unité de $\mathcal{U}$. Ainsi la valeur prise par le caractère y sur la $k$ème unité est notée y_k.

Remarque. Les valeurs prises par le caractère **ne sont pas aléatoires**. C'est d'ailleurs pour cela que l'on parle de *caractère* plutôt que de *variable* ; cette dernière ayant une connotation aléatoire.

Dans un monde idéal, on aimerait donc connaître le **vecteur paramètre** $\mathbf{y}_N = (y_1, \ldots, y_N)$; mais il est clair que ceci relève de l'impossible. Comment connaître ces N valeurs à partir de n observations ($n \ll N$) ? Souvent on visera seulement (et c'est déjà bien suffisant) un résumé du vecteur paramètre $\mathbf{y}_N$ comme par exemple la moyenne, une proportion, ... Plus formellement, on souhaite estimer une fonction θ de $\mathbf{y}_N$

$$\theta = \theta(y_k \colon k \in \mathcal{U}).$$

Exemple 1.1.1. La fonction d'intérêt θ peut être
- un total : $t_y = \sum_{k\in\mathcal{U}} y_k$;
- une moyenne : $\mu_y = N^{-1} \sum_{k\in\mathcal{U}} y_k$;
- un ratio : $R = t_y/t_x$ où x est un deuxième caractère d'intérêt.

1.2 Échantillon

Dans ce cours nous allons croiser essentiellement deux types d'échantillons : avec remise et ordonné et sans remise ni ordre.

Exemple 1.2.1. Soit une population $\mathcal{U} = \{1, 2\}$. L'ensemble des échantillons ordonnés avec remise est

$$\tilde{\mathscr{S}} = \{(1), (2), (1,1), (1,2), (2,1), (2,2), (1,1,1), (1,1,2), \ldots\}.$$

En particulier puisqu'il y a remise, la taille de l'échantillon peut être supérieure à la taille de la population !

Exemple 1.2.2. Soit une population $\mathcal{U} = \{1, 2, 3\}$. L'ensemble des échantillons non ordonnés et sans remise est

$$\mathscr{S} = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}.$$

Il est commode de se représenter un échantillon non ordonné et sans remise comme un **sous ensemble non vide** de $\mathcal{U}$. En effet un ensemble est par définition non ordonné et sans répétition. Ainsi l'ensemble des échantillons non ordonnés et sans remise est l'ensemble des parties non vides de $\mathcal{U}$, i.e.,

$$\mathscr{S} = \{s \colon s \subset \mathcal{U}\} \setminus \emptyset.$$

Par conséquent la taille de l'échantillon est au plus égale à la taille de la population et $|\mathscr{S}| = 2^N - 1$.

Remarque. Clairement il est possible de passer de $\tilde{\mathscr{S}}$ à $\mathscr{S}$ en supprimant l'information sur l'ordre et la multiplicité à l'aide d'une **fonction de réduction** $r \colon \tilde{\mathscr{S}} \mapsto \mathscr{S}$.

Exemple 1.2.3. Pour $\mathcal{U} = \{1, 2, 3\}$, on a

$$r\{(1,1,2)\} = r\{(1,2)\} = r\{(2,1)\} = r\{(1,2,2)\} = \{1,2\}.$$

1.3 Plan de sondage

Définition 1.3.1. Un plan de sondage non ordonné et sans remise p est une **loi de probabilité** sur $\mathscr{S}$, i.e,

$$p(s) \geq 0, \qquad s \in \mathscr{S},$$

et

$$\sum_{s \in \mathscr{S}} p(s) = 1.$$

De même on définit un plan de sondage ordonné avec remise $\tilde{p}$ comme une loi de probabilité sur $\tilde{\mathscr{S}}$.

Clairement la fonction de réduction r permet de définir un plan de sondage sur $\mathscr{S}$ à l'aide d'un plan de sondage sur $\tilde{\mathscr{S}}$, i.e.,

$$p(s) = \sum_{\tilde{s} \in \tilde{\mathscr{S}}} \tilde{p}(\tilde{s}) 1_{\{r(\tilde{s}) = s\}}, \qquad s \in \mathscr{S}.$$

Exemple 1.3.1. Pour $\mathcal{U} = \{1, 2, 3\}$, on considère le plan de sondage consistant à sélectionner 2 unités avec remise et probabilités égales.

— Le plan de sondage sur $\tilde{\mathscr{S}}$ est alors

$$\begin{aligned}
&\tilde{p}\{(1,1)\} = 1/9, && \tilde{p}\{(1,2)\} = 1/9, && \tilde{p}\{(1,3)\} = 1/9,\\
&\tilde{p}\{(2,1)\} = 1/9, && \tilde{p}\{(2,2)\} = 1/9, && \tilde{p}\{(2,3)\} = 1/9,\\
&\tilde{p}\{(3,1)\} = 1/9, && \tilde{p}\{(3,2)\} = 1/9, && \tilde{p}\{(3,3)\} = 1/9,
\end{aligned}$$

— et celui sur $\mathscr{S}$ est

$$\begin{aligned}
&p(\{1\}) = 1/9, && p(\{1,2\}) = 2/9, && p(\{1,3\}) = 2/9\\
&p(\{2\}) = 1/9, && p(\{2,3\}) = 2/9, && p(\{3\}) = 1/9.
\end{aligned}$$

Notez que la taille de l'échantillon pour le plan de sondage sur $\mathscr{S}$ est aléatoire.

Puisqu'un plan de sondage n'est rien d'autre qu'une loi de probabilité, nous pouvons définir des échantillons aléatoires S et $\tilde{S}$, i.e., des variables aléatoires à valeurs dans $\mathscr{S}$ et $\tilde{\mathscr{S}}$ respectivement. Les lois de S et $\tilde{S}$ sont donc données par

$$\Pr(S = s) = p(s), \quad s \in \mathscr{S}, \qquad \text{et} \qquad \Pr(\tilde{S} = \tilde{s}) = \tilde{p}(\tilde{s}), \quad \tilde{s} \in \tilde{\mathscr{S}}.$$

Remarque. Comme nous l'avons vu dans l'exemple précédent, la taille de l'échantillon notée n_S peut être aléatoire. Lorsque $\mathrm{Var}(n_S) = 0$, l'échantillon est dit de **taille fixe**.

1.4 Probabilités d'inclusion

Soit un échantillon aléatoire S, la variable aléatoire $1_{\{k\in S\}}$, $k \in \mathcal{U}$, nous sera très utile. Notons que c'est bien une variable aléatoire puisque S est aléatoire.

Définition 1.4.1. La probabilité d'inclusion de la $k$ème unité, notée π_k, correspond à la probabilité que cette $k$ème unité appartienne à l'échantillon, i.e.,

$$\pi_k = \Pr(k \in S) = \sum_{s\in\mathscr{S}} p(s) 1_{\{k\in s\}} = \sum_{s\ni k} p(s), \qquad k \in \mathcal{U}.$$

Notons également que par définition, $\pi_k = \mathbb{E}(1_{\{k\in S\}})$.

De même nous pouvons définir des probabilités d'inclusion d'ordre supérieur.

Définition 1.4.2. La probabilité d'inclusion d'ordre 2 est la probabilité que deux unités distinctes appartiennent simultanément à un échantillon, i.e.,

$$\pi_{k\ell} = \Pr(k \in S, \ell \in S) = \sum_{s\ni k,\ell} p(s), \qquad k, \ell \in \mathcal{U}, \quad k \neq \ell.$$

Notons que comme précédemment, $\pi_{k\ell} = \mathbb{E}(1_{\{k\in S\}} 1_{\{\ell\in S\}})$.

On a

$$\mathrm{Var}(1_{\{k\in S\}}) = \mathbb{E}(1^2_{\{k\in S\}}) - \mathbb{E}(1_{\{k\in S\}})^2 = \pi_k(1-\pi_k),$$

et

$$\mathrm{Cov}(1_{\{k\in S\}}, 1_{\{\ell\in S\}}) = \mathbb{E}(1_{\{k\in S\}} 1_{\{\ell\in S\}}) - \mathbb{E}(1_{\{k\in S\}})\mathbb{E}(1_{\{\ell\in S\}}) = \pi_{k\ell} - \pi_k\pi_\ell,$$

avec $k, l \in \mathcal{U}$, $k \neq \ell$.

Dans la suite on notera

$$\Delta_{k\ell} = \begin{cases} \mathrm{Cov}(1_{\{k\in S\}}, 1_{\{\ell\in S\}}), & k \neq \ell \\ \mathrm{Var}(1_{\{k\in S\}}), & k = \ell. \end{cases}$$

1.5 Plans simples et de taille fixe

La théorie des sondages revient souvent à caractériser certaines propriétés de plan de sondage donnés. Ici nous nous intéressons aux plans dit **simples** et les plans de **taille fixe**.

Définition 1.5.1. Un plan est dit **simple** si tous les échantillons de même taille ont la même probabilité d'être sélectionnés.

Définition 1.5.2. Un plan est dit de **taille fixe** si $\mathrm{Var}(|S|) = \mathrm{Var}(n_S) = 0$, où $|A|$ représente la cardinalité d'un ensemble A. On notera alors $n = n_S$ la taille de l'échantillon.

Les plans de taille fixe ont des probabilités d'inclusion bien spécifiques.

Théorème 1.5.1. *Si un plan est de taille fixe n, alors*

$$\begin{aligned}
&\sum_{k\in\mathcal{U}} \pi_k = n,\\
&\sum_{\substack{k\in\mathcal{U}\\ k\neq\ell}} \pi_{k\ell} = (n-1)\pi_\ell, \qquad \ell \in \mathcal{U},\\
&\sum_{k\in\mathcal{U}} \Delta_{k\ell} = 0, \qquad \ell \in \mathcal{U}.
\end{aligned}$$

Démonstration.

□

Définition 1.5.3. Un plan sans remise est dit **simple** si tous les échantillons de même taille ont la même probabilité d'être sélectionnés, i.e.,

$$p(s_1) = p(s_2), \qquad s_1, s_2 \in \mathscr{S},\ |s_1| = |s_2|.$$

Remarque. Clairement on a $\binom{N}{n}$ échantillons (non ordonnés) de taille n dans $\mathcal{U}$. Ainsi si le plan est simple **et** de taille fixe on a pour tout $s \in \mathscr{S}$

$$p(s) = \begin{cases} \binom{N}{n}^{-1}, & |s| = n \\ 0, & \text{sinon,} \end{cases}$$

avec $\binom{N}{n} = N!/\{n!(N-n)!\}$.

En revanche, si le plan n'est pas de taille fixe on a

$$p(s) = \binom{N}{n}^{-1} \Pr(|S| = n),$$

où $|s| = n$.

1.6 Le π–estimateur

C'est sans aucun doute l'estimateur qu'il faut à tout prix connaître lorsque l'on s'intéresse aux sondages.

1.6.1 Estimation d'un total et d'une moyenne

Horvitz et Thompson (1952) ont introduit un estimateur linéaire sans biais d'un total t_y pour tout plan de sondage

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Cet estimateur est appelé le **π–estimateur**, l'estimateur d'**Horvitz–Thompson** ou encore l'estimateur **des valeurs dilatées**.

Théorème 1.6.1. *Si $\pi_k > 0$ pour tout $k \in \mathcal{U}$, alors $\hat{t}_{y,\pi}$ estime t_y sans biais.*

Démonstration.

□

Remarque. Si certaines probabilités d'inclusion sont nulles alors l'estimateur est biaisé puisque

$$\mathbb{E}(\hat{t}_{y,\pi}) = \mathbb{E}\left(\sum_{k \in S} \frac{y_k}{\pi_k}\right) = \mathbb{E}\left(\sum_{\substack{k \in \mathcal{U} \\ \pi_k > 0}} \frac{y_k}{\pi_k} 1_{\{k \in S\}}\right) = \sum_{\substack{k \in \mathcal{U} \\ \pi_k > 0}} \frac{y_k}{\pi_k} \pi_k = t_y - \sum_{\substack{k \in \mathcal{U} \\ \pi_k = 0}} y_k.$$

Notons que, lors de la deuxième égalité, la restriction aux unités telles que $\pi_k > 0$ sous le signe de sommation est justifiée par le fait qu'une unité dont la probabilité d'inclusion d'ordre un est nulle n'appartiendra jamais à l'échantillon aléatoire S.

Nous avons introduit le π–estimateur pour estimer le total t_y mais nous pouvons également l'utiliser pour estimer la moyenne μ_y par

$$\hat{\mu}_{y,\pi} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Notons toutefois que pour utiliser cet estimateur il faut que la taille de la population N soit connue—ce n'est malheureusement pas toujours le cas...

Cela dit puisque $N = \sum_{k \in \mathcal{U}} 1$, on peut estimer N par Horvitz–Thompson, i.e.,

$$\hat{N}_\pi = \sum_{k \in S} \frac{1}{\pi_k}.$$

1.6.2 Variance du π–estimateur

Il est également possible de connaître la variance du π–estimateur.

Théorème 1.6.2. *Soit $\hat{t}_{y,\pi}$ le π–estimateur d'un total t_y. Si $\pi_k > 0$ pour tout $k \in \mathcal{U}$, alors*

$$\mathrm{Var}(\hat{t}_{y,\pi}) = \sum_{k,\ell \in \mathcal{U}} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}.$$

Démonstration.

□

1.6.3 Variance pour les plans de taille fixe

Dans le cas de plans de **taille fixe**, on peut réécrire la variance du π–estimateur sont une forme différente.

Théorème 1.6.3. *Soit $\hat{t}_{y,\pi}$ le π–estimateur d'un total t_y. Si le plan est de taille fixe et que $\pi_k > 0$ pour tout $k \in \mathcal{U}$, alors*

$$\mathrm{Var}(\hat{t}_{y,\pi}) = -\frac{1}{2} \sum_{\substack{k,\ell \in \mathcal{U} \\ k \neq \ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell}.$$

Démonstration.

□

1.6.4 Estimation de la variance du π–estimateur

L'idée de base du π–estimateur peut être naturellement étendue au contexte des fonctions de deux variables $f(\cdot,\cdot)$.

Théorème 1.6.4. *Soit $f(\cdot,\cdot)$ une fonction de deux variables quelconque. Si $\pi_{k\ell} > 0$, pour tout $k, \ell \in \mathcal{U}$ $k \neq \ell$, alors*

$$\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \frac{g(y_k, y_\ell)}{\pi_{k\ell}} 1_{\{k\in S, \ell\in S\}}$$

est un estimateur sans biais de

$$\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \frac{g(y_k, y_\ell)}{\pi_{k\ell}}.$$

Démonstration.

□

On peut donc se servir du théorème précédent afin de construire un estimateur sans biais de $\mathrm{Var}(\hat{t}_{y,\pi})$. On a donc à partir de l'expression donnée en Section 1.6.2 l'estimateur

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{t}_{y,\pi}) &= \sum_{k\in\mathcal{U}} \frac{y_k^2 \pi_k^{-2} \Delta_{kk}}{\pi_k} 1_{\{k\in S\}} + \sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \frac{y_k y_\ell \pi_k^{-1} \pi_\ell^{-1} \Delta_{k\ell}}{\pi_{k\ell}} 1_{\{k\in S, \ell\in S\}} \\ &= \sum_{k\in\mathcal{U}} \frac{y_k^2 (1-\pi_k)}{\pi_k^2} 1_{\{k\in S\}} + \sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \frac{y_k y_\ell}{\pi_k \pi_\ell \pi_{k\ell}} \Delta_{k\ell} 1_{\{k\in S, \ell\in S\}}.\end{aligned}$$

Si le plan est à **taille fixe** alors nous pouvons utiliser l'expression donnée lors de la Section 1.6.3 ; ce qui nous conduit à l'estimateur

$$\widehat{\mathrm{Var}}(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}}\left(\frac{y_k}{\pi_k}-\frac{y_\ell}{\pi_\ell}\right)^2\frac{\Delta_{k\ell}}{\pi_{k\ell}}1_{\{k\in S,\ell\in S\}}.$$

Ce dernier estimateur est appelé l'estimateur de Sen–Yates–Grundy, noms des personnes l'ayant trouvé.

Remarque. Cet estimateur est sans biais uniquement lorsque le plan est de taille fixe et il n'est pas difficile de voir que l'estimateur sera toujours positif dès lors que $\Delta_{k\ell} \leq 0$ pour tout $k, \ell \in \mathcal{U}$, $k \neq \ell$. C'est la condition de Sen–Yates–Grundy.

1.7 L'estimateur de Hájek

Bien que le π–estimateur soit très largement utilisé, il existe certaines situations où ce dernier se comporte pas très bien... Afin d'illustrer nos propos, supposons que

$$\mathrm{Var}\left(\sum_{k\in\mathcal{U}}\frac{1}{\pi_k}1_{\{k\in S\}}\right)\neq 0.$$

Remarque. Ceci est par exemple le cas lorsque la taille de l'échantillon est aléatoire.

Supposons de plus que $y_k = c$ pour tout $k \in \mathcal{U}$. Alors le π–estimateur de la moyenne μ_y est alors

$$\hat{\mu}_{y,\pi} = \frac{c}{N}\sum_{k\in\mathcal{U}}\frac{1}{\pi_k}1_{\{k\in S\}},$$

et nous concluons que $\hat{\mu}_{y,\pi}$ n'est pas égale à c mais est une variable aléatoire d'espérance c. Avouons que c'est une propriété assez embarrassante.

L'estimateur de Hájek a été introduit afin de remédier à ce problème et est donné par

$$\hat{\mu}_{y,H} = \left(\sum_{k\in\mathcal{U}}\frac{1}{\pi_k}1_{\{k\in S\}}\right)^{-1}\sum_{k\in\mathcal{U}}\frac{y_k}{\pi_k}1_{\{k\in S\}}.$$

Remarque. L'estimateur de Hájek correspond à un ratio de deux variables aléatoires. Le calcul de ses moments est alors compliqué voire impossible.

Évidemment on peut étendre cet estimateur pour l'estimation d'un total t_y en posant

$$\hat{t}_{y,H} = N\left(\sum_{k\in\mathcal{U}}\frac{1}{\pi_k}1_{\{k\in S\}}\right)^{-1}\sum_{k\in\mathcal{U}}\frac{y_k}{\pi_k}1_{\{k\in S\}},$$

dès lors que N est connu bien évidemment.

Chapitre 2

Les plans simples

Ce chapitre traite exclusivement des plans simples. Il est important de bien maîtriser ces plans car ils forment souvent la base de plans de sondage plus complexes, tel que les plans stratifiés ou par grappes. A bien connaître donc !

2.1 Plans simples sans remise

2.1.1 Plan de sondage et probabilités d'inclusion

Un plan est dit **simple** si tous les échantillons **de même taille** ont la même probabilité d'être sélectionnés. En conséquence, il n'existe qu'un seul plan simple de **taille fixe** n.

Définition 2.1.1. Un plan de **taille fixe** n est dit **simple sans remise** si

$$p(s) = \begin{cases} \binom{N}{n}^{-1}, & |s| = n, \\ 0, & \text{sinon}, \end{cases}$$

avec $n \in \{1, \ldots, N\}$.

Comme vous le savez maintenant, il est souvent utile pour nous statisticiens de connaître les probabilités d'inclusion—afin de pouvoir établir le π–estimateur et sa variance par exemple.

Ces probabilités d'inclusions se calculent facilement. En effet

$$\pi_k = \sum_{s \ni k} p(s) = \underbrace{\binom{N-1}{n-1}}_{\substack{\text{nb. d'échantillons} \\ \text{contenant } k}} \binom{N}{n}^{-1} = \frac{(N-1)!}{(n-1)!(N-n)!} \frac{n!(N-n)!}{N!} = \frac{n}{N}$$

$$\pi_{k\ell} = \sum_{s:\, k,\ell \in s} p(s) = \underbrace{\binom{N-2}{n-2}}_{\substack{\text{nb. échantillons} \\ \text{contenant } k \text{ et } \ell}} \binom{N}{n}^{-1} = \frac{(N-2)!}{(n-2)!(N-n)!} \frac{n!(N-n)!}{N!} = \frac{n(n-1)}{N(N-1)}$$

Remarque. Notons que $\pi_{k\ell} \neq \pi_k \pi_\ell$, indiquant une dépendance entre les unités choisies dû au tirage sans remise.

Des deux expressions précédentes, on en déduit

$$\Delta_{k\ell} = \begin{cases} \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n(N-n)}{N^2(N-1)}, & k \neq \ell, \\ \frac{n}{N}\left(1 - \frac{n}{N}\right) = \frac{n(N-n)}{N^2}, & k = \ell. \end{cases}$$

2.1.2 Le π–estimateur pour ces plans

A l'aide des probabilités d'inclusions de la Section 2.1.1, nous pouvons donner une version plus explicite du π–estimateur. Le π–estimateur d'une moyenne μ_y

$$\hat{\mu}_{y,\pi} = \frac{1}{N}\sum_{k\in\mathcal{U}} \frac{y_k}{\pi_k} 1_{\{k\in S\}} = \frac{N}{nN}\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}} = \overline{y},$$

et le π–estimateur du total t_y est évidemment $\hat{t}_{y,\pi} = N\overline{y}$, avec

$$\overline{y} = \frac{1}{n}\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}}.$$

Rappelons que puisque le plan est à taille fixe et que les probabilités d'inclusions des deux premiers ordres sont strictement positives, on peut utiliser la formule de la variance de $\hat{\mu}_{y,\pi}$ trouvée par Sen–Yates–Grundy, i.e.,

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{y,\pi}) &= -\frac{1}{2N^2}\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \Delta_{k\ell} = \frac{1}{2N^2}\times\frac{N-n}{n(N-1)}\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} (y_k - y_\ell)^2 \\ &= \frac{N-n}{nN} S_y^2,\end{aligned}$$

avec

$$S_y^2 = \frac{1}{N-1}\sum_{k\in\mathcal{U}} (y_k - \mu_y)^2 = \frac{1}{2N(N-1)}\sum_{\substack{k,\ell\in\mathcal{U}\\ k\neq\ell}} (y_k - y_\ell)^2.$$

Remarque. La variance précédente peut également s'écrire

$$\mathrm{Var}(\hat{\mu}_{y,\pi}) = \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}.$$

Le terme S_y^2/n correspond à la variance d'une moyenne empirique pour les statistiques inférentielles classique alors que le premier terme $(1 - n/N)$ correspond au **facteur de correction en population finie**. On appelle également le ratio $f = n/N$ le **taux de sondage**.

De l'expression précédente, on en déduit directement la variance du π–estimateur du total t_y

$$\mathrm{Var}(\hat{t}_{y,\pi}) = N(N-n)\frac{S_y^2}{n}.$$

Théorème 2.1.1. *Pour un plan de taille fixe n, simple et sans remise la variance corrigée de la population S_y^2 est estimée sans biais par*

$$\widehat{S_y^2} = \frac{1}{n-1}\sum_{k\in\mathcal{U}} (y_k - \bar{y})^2 1_{\{k\in S\}}.$$

Démonstration.

□

Au final, on peut estimer sans biais la variance de $\hat{\mu}_{y,\pi}$ pour ces plans particuliers par

$$\widehat{\mathrm{Var}}(\hat{\mu}_{y,\pi}) = \frac{N-n}{N}\frac{\widehat{S_y^2}}{n},$$

et pour le π–estimateur du total $\hat{t}_{y,\pi}$

$$\widehat{\mathrm{Var}}(\hat{t}_{y,\pi}) = N(N-n)\frac{\widehat{S_y^2}}{n}.$$

2.2 Plans simples avec remise

Le plan de taille fixe n, simple et avec remise correspond au cadre de la statistique inférentielle usuelle. En effet le plan de sondage consiste à sélectionner une unité aléatoire avec probabilités égales $1/N$ et de recommencer l'opération n fois indépendemment. On se ramène donc au cadre de variable aléatoire indépendantes et identiquement distribuées de moyenne

$$\mu_y = \frac{1}{N}\sum_{k\in\mathcal{U}} y_k,$$

et de variance

$$\sigma_y^2 = \frac{1}{N}\sum_{k\in\mathcal{U}}(y_k-\mu_y)^2.$$

Nous le savons déjà mais la moyenne sur la population μ_y est estimée sans biais par

$$\hat{\mu}_y = \frac{1}{n}\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}} = \bar{y}.$$

En effet

$$\mathbb{E}(\hat{\mu}_y) = \frac{1}{n}\sum_{k\in\mathcal{U}} y_k \frac{n}{N} = \mu_y.$$

De plus, puisque les y_k de l'échantillon sont sélectionnées indépendemment et sont de même loi,

$$\mathrm{Var}(\hat{\mu}_y) = \frac{\sigma_y^2}{n}.$$

Théorème 2.2.1. *Pour un plan de taille fixe n, simple et sans remise, la variance non corrigée de la population*

$$\sigma_y^2 = \frac{1}{N}\sum_{k\in\mathcal{U}}(y_k-\mu_y)^2,$$

est estimée sans biais par

$$\widehat{\sigma_y^2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

Démonstration.

□

Au final la variance de $\hat{\mu}_y$ est estimée sans biais par

$$\widehat{\mathrm{Var}}(\hat{\mu}_y) = \frac{\widehat{\sigma_y^2}}{n}.$$

2.3 Comparaison des plans simples avec et sans remise

Table 2.1: *Récapitulatif des résultats pour les plans simples de taille fixe n.*

	Sans remise	Avec remise
Estimateur de la moyenne	$\frac{1}{n}\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}}$	$\frac{1}{n}\sum_{k\in S} y_k$
Variance de l'estimateur de la moyenne	$\frac{N-n}{N} \times \frac{S_y^2}{n}$	$\frac{\sigma_y^2}{n}$
Estimateur de la variance de l'estimateur de la moyenne	$\frac{N-n}{N}\frac{\widehat{S_y^2}}{n}$	$\frac{\widehat{\sigma_y^2}}{n}$

Le sondage simple et sans remise est **toujours** préférable à celui avec remise. En effet si l'on appelle $\hat{\mu}_{y,\pi}$ et $\tilde{\mu}_{y,\pi}$ les π–estimateurs de la moyenne avec et sans remise, alors pour tout $n \geq 2$

$$\frac{\mathrm{Var}(\tilde{\mu}_{y,\pi})}{\mathrm{Var}(\hat{\mu}_{y,\pi})} = \frac{(N-n)}{N} \times \frac{S_y^2}{\sigma_y^2} = \frac{N-n}{N} \times \frac{N}{N-1} = \frac{N-n}{N-1} < 1.$$

Voilà pourquoi nous allons essentiellement nous concentrer sur les plan simples sans remise.

2.4 Plans simples sans remise et fonction d'intérêt

Jusqu'à présent nous avons essentiellement parlé de l'estimation d'un total t_y ou d'une moyenne μ_y. Parfois l'étude porte sur d'autres grandeurs et donc d'autres fonctions d'intérêt.

2.4.1 Estimation d'une proportion

Il est fréquent qu'une étude porte sur l'estimation d'une proportion p. Avec notre terminologie estimer une proportion revient à compter le nombre d'unités y_k, $k \in \mathcal{U}$, possédant une certaine caractéristique. À partir du caractère y_k, on introduit alors un nouveau caractère

$$z_k = \begin{cases} 1, & \text{si } y_k \text{ possède la caractéristique,} \\ 0, & \text{sinon,} \end{cases} \qquad k \in \mathcal{U},$$

ce qui nous permettra généralement de nous servir des fonctions d'intérêt déjà rencontrées :

$$\begin{aligned} \mu_z &= \frac{1}{N}\sum_{k\in\mathcal{U}} z_k = \frac{\#\{z_k \in \mathcal{U} \colon z_k = 1\}}{N} = p \\ t_z &= \#\{z_k \in \mathcal{U} \colon z_k = 1\} = Np \\ \sigma_z^2 &= \frac{1}{N}\sum_{k\in\mathcal{U}} z_k^2 - \mu_z^2 = p - p^2 = p(1-p) \\ S_z^2 &= \frac{N}{N-1}p(1-p). \end{aligned}$$

Nous voyons donc qu'estimer une proportion n'est rien d'autre qu'estimer une moyenne. En revanche, pour des proportions, les expressions pour la variance se voient considérablement simplifiées du fait que $z_k^2 = z_k$ pour tout $k \in \mathcal{U}$. Ainsi pour un plan simple sans remise, nous avons

$$\begin{aligned} \hat{p} &= \frac{1}{n}\sum_{k\in\mathcal{U}} z_k 1_{\{k\in S\}} \\ s_z^2 &= \frac{1}{n-1}\sum_{k\in\mathcal{U}} (z_k - \hat{p})^2 1_{\{k\in S\}} = \frac{n}{n-1}\hat{p}(1-\hat{p}) \\ \operatorname{Var}(\hat{p}) &= \frac{N-n}{N} \times \frac{S_p^2}{n} = \frac{N-n}{N-1} \times \frac{p(1-p)}{n} \\ \widehat{\operatorname{Var}}(\hat{p}) &= \frac{N-n}{N} \times \frac{\hat{p}(1-\hat{p})}{n-1}. \end{aligned}$$

Remarque. Une fois p estimé par $\hat{p}$, nous obtenons directement une estimation de $\operatorname{Var}(\hat{p})$. Merci les proportions !

2.4.2 Estimation d'un ratio

Considérons cette fois deux caractères y et x. On sera souvent intéressé par l'estimation du ratio

$$R = \frac{\sum_{k\in\mathcal{U}} y_k}{\sum_{k\in\mathcal{U}} x_k} = \frac{\mu_y}{\mu_x}.$$

Pour un plan simple sans remise, on estimera ce ratio par le rapport des moyennes empiriques, i.e.,

$$\hat{R} = \frac{\sum_{k\in\mathcal{U}} y_k 1_{\{k\in S\}}}{\sum_{k\in\mathcal{U}} x_k 1_{\{k\in S\}}} = \frac{\bar{y}}{\bar{x}}.$$

Toutefois l'étude des propriétés de cet estimateur, comme le biais ou l'erreur quadratique, s'avère compliquée puisque nous sommes en présence d'un rapport de deux variables aléatoires ! Lors de tels cas, une technique à retenir est la **linéarisation** du ratio.

$$\hat{R} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}} = \frac{\bar{y} - R\bar{x}}{\mu_x(1+\varepsilon)}, \qquad \varepsilon = \frac{\bar{x} - \mu_x}{\mu_x}$$

Commençons par noter que $\mathbb{E}(\varepsilon) = 0$ et que $\varepsilon \to 0$ lorsque $n \to N$. Ainsi un développement limité de $(1+\varepsilon)^{-1}$ en 0 à l'ordre 1 donne

$$(1+\varepsilon)^{-1} = 1 - \varepsilon + o(\varepsilon^2),$$

et donc

$$\begin{aligned}
\mathbb{E}(\hat{R} - R) &\approx \mathbb{E}\left\{\frac{\bar{y} - R\bar{x}}{\mu_x}(1-\varepsilon)\right\} \\
&= -\mathbb{E}\left(\frac{\bar{y} - R\bar{x}}{\mu_x}\varepsilon\right), && \text{car } \mathbb{E}\left(\frac{\bar{y} - R\bar{x}}{\mu_x}\right) = 0 \\
&= \mathbb{E}\left\{\frac{(R\bar{x} - \bar{y})(\bar{x} - \mu_x)}{\mu_x^2}\right\} \\
&= \mathbb{E}\left\{\frac{(R\bar{x} - R\mu_x + \mu_y - \bar{y})(\bar{x} - \mu_x)}{\mu_x^2}\right\}, && \text{car } \mu_y = R\mu_x \\
&= \frac{R\mathbb{E}\{(\bar{x} - \mu_x)^2\} - \mathrm{Cov}(\bar{x}, \bar{y})}{\mu_x^2}.
\end{aligned}$$

Au final on a donc que le biais de $\hat{R}$ est approximativement

$$\mathrm{Biais}(\hat{R}) \approx \frac{1}{\mu_x^2}\left(R\frac{N-n}{N} \times \frac{S_x^2}{n} - \frac{N-n}{N} \times \frac{S_{xy}}{n}\right) = \frac{1}{\mu_x^2} \times \frac{N-n}{N} \times \frac{1}{n}\left(RS_x^2 - S_{xy}\right),$$

avec

$$S_{xy} = \frac{1}{N-1}\sum_{k \in \mathcal{U}}(x_k - \mu_x)(y_k - \mu_y).$$

Remarque. Le biais est donc approximativement nul dès lors que la taille de l'échantillon est grande.

On procède de même pour approcher l'erreur quadratique moyenne de $\hat{R}$.

$$\begin{aligned}
\mathbb{E}\{(\hat{R} - R)^2\} &\approx \mathbb{E}\left\{\left(\frac{\bar{y} - R\bar{x}}{\mu_x}\right)^2\right\} \\
&= \mathbb{E}\left\{\left(\frac{\bar{y} - \mu_y + R\mu_x - R\bar{x}}{\mu_x}\right)^2\right\} \\
&= \frac{1}{\mu_x^2}\left\{\mathrm{Var}(\bar{y}) + R^2\,\mathrm{Var}(\bar{x}) - 2R\,\mathrm{Cov}(\bar{x}, \bar{y})\right\} \\
&= \frac{1}{\mu_x^2} \times \frac{N-n}{N} \times \frac{1}{n}\left(S_y^2 + R^2S_x^2 - 2RS_{xy}\right).
\end{aligned}$$

Cette erreur quadratique étant naturellement estimée par

$$\widehat{\mathbb{E}}\{(\hat{R} - R)^2\} = \frac{1}{\bar{x}^2} \times \frac{N-n}{N} \times \frac{1}{n}\left(\widehat{S_y^2} + \hat{R}^2\widehat{S_x^2} - 2\hat{R}\widehat{S_{xy}}\right),$$

avec

$$\widehat{S_{xy}} = \frac{1}{n-1}\sum_{k \in \mathcal{U}}(x_k - \bar{x})(y_k - \bar{y})1_{\{k \in S\}}.$$

2.5 Détermination de la taille de l'échantillon

Avant de commencer un sondage, il est toujours souhaitable de se poser la question des incertitudes liées à nos futures estimations. Généralement les limites budgétaires fixeront la taille de l'échantillon et on se contentera alors de répondre si le budget alloué est suffisant pour une précision donnée—et de demander une rallonge à son chef le cas échéant...

Par précision donnée nous entendons que le paramètre d'intérêt θ sera contenu dans un intervalle de confiance centré en $\hat{\theta}$ avec une probabilité d'au moins $1-\alpha$, i.e., trouver $\ell > 0$ tel que

$$\Pr\left\{\theta \in \left[\hat{\theta} - \ell, \hat{\theta} + \ell\right]\right\} \geq 1 - \alpha$$

En supposant que notre estimateur $\hat{\theta}$ suit approximativement une loi normale (ce qui sera souvent le cas), on sait que

$$\Pr\left\{\theta \in \left[\hat{\theta} - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}, \hat{\theta} + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}\right]\right\} = 1 - \alpha,$$

où $z_{1-\alpha/2}$ le quantile d'une loi normale centrée réduite de probabilité au non dépassement $1-\alpha/2$, i.e., $\Pr(Z \leq z_{1-\alpha/2}) = 1-\alpha/2$, $Z \sim N(0,1)$.

Remarque. Puisque $\widehat{\mathrm{Var}}(\hat{\theta})$ dépend de la taille de l'échantillon n, on cherchera donc la taille minimale n_0 induisant la précision requise.

Pour illustrer nos propos prenons le cas de l'estimation de la moyenne μ_y pour un plan simple sans remise. On a donc

$$\Pr\left\{\mu_y \in \left[\bar{y} - z_{1-\alpha/2}\sqrt{\frac{N-n}{nN}S_y^2}, \bar{y} + z_{1-\alpha/2}\sqrt{\frac{N-n}{nN}S_y^2}\right]\right\} = 1 - \alpha,$$

et il faut donc nécessairement

$$\begin{aligned}
\ell^2 \geq z_{1-\alpha/2}^2 \frac{N-n}{nN} S_y^2 &\iff nN\ell^2 \geq z_{1-\alpha/2}^2 (N-n) S_y^2 \\
&\iff n(N\ell^2 + z_{1-\alpha/2}^2 S_y^2) \geq N S_y^2 z_{1-\alpha/2}^2 \\
&\iff n \geq \frac{N S_y^2 z_{1-\alpha/2}^2}{N\ell^2 + z_{1-\alpha/2}^2 S_y^2}
\end{aligned}$$

Malheureusement cette expression n'est pas si utile en pratique car si notre objectif initial était d'estimer μ_y, il est fort à parier que nous connaissions la variance corrigée S_y^2... En pratique on pourra prendre par exemple une estimation de S_y^2 basée sur des études antérieures.

Lorsque notre paramètre d'intérêt est une proportion, nous pouvons tout de même déterminer la taille minimale. Dans ce contexte, nous avons alors

$$n \geq \frac{N \frac{n}{n-1}\hat{p}(1-\hat{p}) z_{1-\alpha/2}^2}{N\ell^2 + z_{1-\alpha/2}^2 \frac{n}{n-1}\hat{p}(1-\hat{p})},$$

et nous pouvons considérer le pire cas possible qui est atteint lorsque $\hat{p} = 0.5$. En effet puisque $\widehat{\mathrm{Var}}(\hat{p})$ est proportionnel à $\hat{p}(1-\hat{p})$ la variance est maximale lorsque $\hat{p} = 1/2$.

Chapitre 3

Plans à probabilités inégales

Ce chapitre explique comment nous pouvons bénéficier de la connaissance d'un caractère auxiliaire pour obtenir des estimations plus précises.

3.1 Caractère auxiliaire et probabilités d'inclusion

Soit x_k, $k \in \mathcal{U}$, les valeurs prises par le caractère auxiliaire. Notons tout de suite que cela implique donc sa connaissance sur toute la population ! Notre étude portant toujours sur une fonction d'intérêt telle que la moyenne ou le total d'un caractère y. Le principe d'un plan à probabilités inégales consiste à définir des probabilités d'inclusion du premier ordre proportionnelles aux x_k.

Rappelons que pour un plan de taille fixe, la variance du π–estimateur du total t_y est

$$\mathrm{Var}(\hat{t}_y) = \frac{1}{2} \sum_{\substack{k,\ell \in \mathcal{U} \\ k \neq \ell}} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \Delta_{k\ell}. \tag{3.1}$$

Si nous souhaitons minimiser (3.1) en jouant seulement sur les probabilités d'inclusions du premier ordre π_k, il est clair que prendre

$$\pi_k = \frac{y_k}{\sum_{\ell \in \mathcal{U}} y_\ell} \propto y_k, \qquad k \in \mathcal{U},$$

est un choix judicieux puisque $\mathrm{Var}(\hat{t}_y)$ est alors nulle. Bien évidemment cette approche est impossible puisqu'elle suppose connaître les valeurs prise par le caractère y sur toute la population $\mathcal{U}$ — inutile alors de faire un sondage !

En revanche si nous disposons d'un caractère auxiliaire x connu sur toute la population et dont on pense qu'il est approximativement proportionnel au caractère y, alors on gagnera à définir les probabilités d'inclusion du premier ordre proportionnellement aux x_k.

Remarque. Si au contraire le caractère x n'est pas du tout proportionnel à y, le plan de sondage sera alors catastrophique et il sera préférable de prendre un plan simple. A méditer donc !

Puisque pour un plan de taille fixe n, cf. Section 1.5,

$$\sum_{k \in \mathcal{U}} \pi_k = n, \tag{3.2}$$

pour obtenir des probabilités d'inclusion proportionnelles aux x_k, i.e., $\pi_k = c x_k$ avec

$$c = \frac{n}{\sum_{\ell \in \mathcal{U}} x_\ell} = \frac{n}{t_x}.$$

Attention toutefois, il n'y a aucune garantie que les $\pi_k \in [0,1]$ et il sera fréquent que certaines "probabilités d'inclusion" soient supérieure à 1. Pour de telles situations, on sélectionnera d'office les unités correspondantes, i.e., $\pi_k = 1$, et l'on recommencera la procédure avec les unités restantes en prenant soin de diminuer la taille de l'échantillon n dans (3.2).

Exemple 3.1.1. Considérons la population $\mathcal{U} = \{1, 2, \ldots, 6\}$ avec une variable auxiliaire x telle que

$$x_1 = 1, \quad x_2 = 9, \quad x_3 = 10, \quad x_4 = 70, \quad x_5 = 90, \quad x_6 = 120.$$

On a donc $t_x = 300$. Si l'on souhaite obtenir un plan de taille fixe $n = 3$, alors les "probabilités d'inclusions temporaires".

$$\pi_1 = \frac{3 \times 1}{300}, \qquad \pi_2 = \frac{3 \times 9}{300}, \qquad \pi_3 = \frac{3 \times 10}{300},$$
$$\pi_4 = \frac{3 \times 70}{300}, \qquad \pi_5 = \frac{3 \times 90}{300}, \qquad \pi_6 = \frac{3 \times 104}{300} > 1.$$

L'unité 6 est alors sélectionnée d'office, le total sans la 6ème unité est

$$\sum_{k \in \mathcal{U} \setminus \{6\}} x_k = t_x - 120 = 180,$$

et les "probabilités d'inclusions" deviennent

$$\pi_1 = \frac{(3-1) \times 1}{180}, \qquad \pi_2 = \frac{(3-1) \times 9}{180}, \qquad \pi_3 = \frac{(3-1) \times 10}{180},$$
$$\pi_4 = \frac{(3-1) \times 70}{180}, \qquad \pi_5 = \frac{(3-1) \times 90}{180}, \qquad \pi_6 = 1.$$

On arrête ici la procédure et les "vraies" probabilités d'inclusion sont

$$\pi_1 = \frac{1}{90}, \quad \pi_2 = \frac{1}{10}, \quad \pi_3 = \frac{1}{9}, \quad \pi_4 = \frac{7}{9}, \quad \pi_5 = \pi_6 = 1.$$

Les unités 5 et 6 sont donc sélectionnées d'office et il restera donc à choisir une unité parmi $\{1, 2, 3, 4\}$. Notons que

$$\sum_{k=1}^{6} \pi_k = \frac{1 + 9 + 10 + 70}{90} + 2 = 3,$$

comme souhaité.

Rappelons qu'un plan de sondage est défini par les $p(s)$ et non par les π_k. Pour avoir un plan à probabilités inégales, il faut donc définir un plan de sondage $p(\cdot)$ tel que pour tout $k \in \mathcal{U}$,

$$\sum_{\substack{s \ni k \\ s \in \mathscr{S}_n}} p(s) = \pi_k, \qquad \mathscr{S}_n = \{s \subset \mathcal{U} \colon |s| = n\}.$$

Remarque. Il existe une infinité de plans de sondage vérifiant ces conditions. Nous allons donc par la suite introduire quelques plans de sondage à probabilités inégales à taille fixe n couramment utilisés.

Algorithme 1 : Algorithme pour un plan de Poisson.

Entrée : les probabilités d'inclusions π_k, la taille de la population N
Sortie : Un échantillon s

```
s = ∅;
pour k ← 1 a N faire
    U ∼ U(0, 1);
    si U < π_k alors
        s ← s ∪ {k};
    fin
fin
retourner s;
```

3.2 Plan de Poisson†

Le plan de Poisson a de très bonnes qualités mais également un gros défaut : il n'est pas de taille fixe. Néanmoins nous allons l'introduire car il va nous servir afin d'en déduire des plans de taille fixe.

Le plan de Poisson se programme très facilement et est décrit par l'algorithme 1. Il est clair que cet algorithme n'est pas de taille fixe : impossible de connaître la taille de l'échantillon avant d'avoir terminé l'exécution de l'algorithme. Il y a même une probabilité non nulle de sélectionner un échantillon de taille nulle ! ! ! Il a cependant de bonnes qualités.

Puisque les unités sont sélectionnées indépendemment

$$\pi_{k\ell} = \Pr(k \in S, \ell \in S) = \Pr(k \in S)\Pr(\ell \in S) = \pi_k \pi_\ell,$$

et donc

$$\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell = 0, \qquad k \neq \ell.$$

Clairement le plan de sondage est donné pour tout $s \subset \mathcal{U}$

$$p(s) = \underbrace{\prod_{k\in s} \pi_k}_{\substack{\text{proba. de sélectionner} \\ \text{les unités chosies}}} \times \underbrace{\prod_{k\in\mathcal{U}\setminus\{s\}} (1-\pi_k)}_{\substack{\text{proba. de ne pas sélectionner} \\ \text{les unités non retenues}}}$$

Puisque $\Delta_{k\ell} = 0$, $k \neq \ell$, la variance du π–estimateur du total t_y est

$$\mathrm{Var}(\hat{t}_y) = \sum_{k,\ell\in\mathcal{U}} \frac{y_k y_\ell}{\pi_k\pi_\ell}\Delta_{k\ell} = \sum_{k\in\mathcal{U}} \frac{y_k^2\pi_k(1-\pi_k)}{\pi_k^2} = \sum_{k\in\mathcal{U}} \frac{y_k^2(1-\pi_k)}{\pi_k},$$

et peut être estimée par

$$\widehat{\mathrm{Var}}(\hat{t}_y) = \sum_{k\in\mathcal{U}} \frac{y_k^2(1-\pi_k)}{\pi_k^2} 1_{\{k\in S\}}.$$

Le plan de Poisson est intéressant car il est simple à mettre en oeuvre mais également car il maximise l'entropie. Nous introduisons maintenant un mesure du "désordre".

Définition 3.2.1. On appelle entropie d'un plan $p(\cdot)$ la quantité

$$I(p) = -\sum_{s\subset\mathcal{U}} p(s)\ln p(s),$$

avec la convention que $0\ln 0 = 0$.

Clairement l'entropie est toujours positive. De plus, comme mesure du désordre, plus $I(p)$ sera grand plus le plan $p(\cdot)$ sera "aléatoire". Pour des probabilités d'inclusion fixées, on cherchera donc le plan le plus aléatoire ou désordonné, i.e., celui maximisant l'entropie.

Lemme 3.2.1.

$$\sum_{s\subset\mathcal{U}}\prod_{k\in s} x_k = \prod_{k\in\mathcal{U}}(1+x_k).$$

Démonstration.

□

Théorème 3.2.2. *Étant donné des probabilités d'inclusions fixées π_k, $k \in \mathcal{U}$, le plan de Poisson est le plan d'entropie maximale sur $\mathscr{S} = \{s\colon s \subset \mathcal{U}\}$.*

Démonstration.

□

On retiendra donc que le plan de Poisson est un plan de sondage respectant les probabilités d'inclusions d'ordre un fixée a priori et étant le « plus aléatoire possible » (au sens de l'entropie). Il a toutefois l'inconvénient de ne pas être à taille fixe.

3.3 Sondage systématique à probabilités inégales

Ce plan de sondage a été introduit vers 1950 et est toujours largement utilisé puisqu'il a le mérite d'être simple et exact ! Contrairement au plan de Poisson, elle a également le bon goût d'être de taille fixe.

Comme depuis le début de ce Chapitre, on désire tirer des échantillons dont les probabilités d'inclusion d'ordre un sont fixées a priori et telles que $0 < \pi_i < 1$, $k \in \mathcal{U}$ et

$$\sum_{k \in \mathcal{U}} \pi_k = n.$$

Définissons les probabilités d'inclusion cumulées

$$C_k = \sum_{\ell=1}^{k} \pi_\ell, \qquad k \in \mathcal{U}, \qquad C_0 = 0.$$

L'approche consiste à générer $U \sim U(0,1)$ et de sélectionner les unités à partir de cette unique réalisation. La première unité sélectionnée, appelons là k_1, sera celle telle que

$$C_{k_1 - 1} \leq U < C_{k_1};$$

la deuxième unité sélectionnée, notons la k_2, sera cette fois ci

$$C_{k_2 - 1} \leq 1 + U < C_{k_2};$$

et ainsi de suite... De manière générale, la $j$ème unité sélectionnée, notée k_j, sera alors

$$C_{k_j - 1} \leq j - 1 + U < C_{k_j}.$$

Exercice 1. Prenons la situation où $N = 6$, $n = 3$, $\pi_1 = 0.2, \pi_2 = 0.7, \pi_3 = 0.8, \pi_4 = 0.5$ et $\pi_5 = \pi_6 = 0.4$. Déterminer l'échantillon sélectionné sachant que $U = 0.3658$.

Solution.

□

Nous venons de voir que cette méthode est en effet très simple. Elle a quand même quelques défauts ; notamment les probabilités d'inclusions d'ordre deux sont souvent nulles.

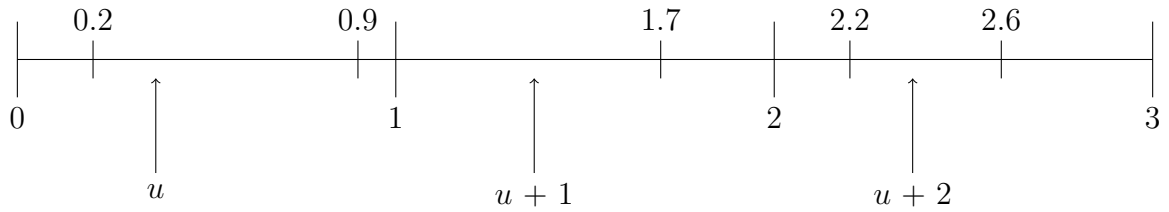


Figure 3.1: *Illustration du tirage systématique de l'exercice 1.*

Exercice 2. Montrez que la matrice $P = (\pi_{k\ell})_{k,\ell}$ des probabilités d'inclusion d'ordre deux de l'exercice 1 est

$$P = \begin{bmatrix} — & 0.0 & 0.2 & 0.2 & 0.0 & 0 \\ 0.0 & — & 0.5 & 0.2 & 0.4 & 0.3 \\ 0.2 & 0.5 & — & 0.3 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.3 & — & 0.0 & 0.3 \\ 0.0 & 0.4 & 0.4 & 0.0 & — & 0 \\ 0.0 & 0.3 & 0.2 & 0.3 & 0.0 & — \end{bmatrix}$$

Solution.

□

Chapitre 4

Stratification

La technique de stratification est largement utilisée en sondage car elle permet facilement d'introduire de l'information auxiliaire pour la construction d'un plan de sondage adéquat.

4.1 Population et strates

Supposons que la population $\mathcal{U}$ soit partitionnée en H sous-ensembles $\mathcal{U}_1, \ldots, \mathcal{U}_H$ appelés strates et tels que

$$\bigcup_{i=1}^{H} \mathcal{U}_i = \mathcal{U}, \qquad \mathcal{U}_i \cap \mathcal{U}_h = \emptyset, \quad i \neq h.$$

Chaque strate $\mathcal{U}_h$ admet une taille N_h et l'on a bien évidemment

$$\sum_{h=1}^{H} N_h = N,$$

où N est la taille de la population $\mathcal{U}$.

Remarque. Les tailles des strates N_h sont ici supposées connues et constituent l'information auxiliaire.

Notre but étant toujours d'estimer un total ou une moyenne, remarquons que le total (resp. la moyenne) s'écrit à l'aide des strates

$$t_y = \sum_{k \in \mathcal{U}} y_k = \sum_{h=1}^{H} \sum_{k \in \mathcal{U}_h} y_k = \sum_{h=1}^{H} t_{y,h},$$

où $t_{y,h}$ est le total des valeurs prises par le caractère y sur la strate $\mathcal{U}_h$, i.e.,

$$t_{y,h} = \sum_{k \in \mathcal{U}_h} y_k.$$

De même la moyenne sur la population s'écrit

$$\mu_y = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k = \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in \mathcal{U}_h} y_k = \frac{1}{N} \sum_{h=1}^{H} N_h \mu_{y,h},$$

où $\mu_{y,h}$ est la moyenne des valeurs prises par le caractère y sur la strate $\mathcal{U}_h$, i.e.,

$$\mu_{y,h} = \frac{1}{N_h} \sum_{k \in \mathcal{U}_h} y_k.$$

On définit également la variance et la variance corrigée sur une strate $\mathcal{U}_h$ par

$$\sigma_{y,h}^2 = \frac{1}{N_h} \sum_{k \in \mathcal{U}_h} (y_k - \mu_{y,h})^2,$$

et

$$S_{y,h}^2 = \frac{1}{N_h - 1} \sum_{k \in \mathcal{U}_h} (y_k - \mu_{y,h})^2.$$

Remarque. La variance sur la population (totale) σ_y^2 s'écrit

$$\begin{aligned}
\sigma_y^2 &= \frac{1}{N} \sum_{k \in \mathcal{U}} (y_k - \mu_y)^2 \\
&= \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in \mathcal{U}_h} \{(y_k - \mu_{y,h}) + (\mu_{y,h} - \mu_y)\}^2 \\
&= \frac{1}{N} \sum_{h=1}^{H} \left\{ \sum_{k \in \mathcal{U}_h} (y_k - \mu_{y,h})^2 + 2(\mu_{y,h} - \mu_y) \underbrace{\sum_{k \in \mathcal{U}_h} (y_k - \mu_{y,h})}_{=0} + N_h (\mu_{y,h} - \mu_y)^2 \right\} \\
&= \frac{1}{N} \sum_{h=1}^{H} N_h \sigma_{y,h}^2 + \frac{1}{N} \sum_{h=1}^{H} N_h (\mu_{y,h} - \mu_y)^2 \\
&= \sigma_{y,\text{intra}}^2 + \sigma_{y,\text{inter}}^2,
\end{aligned}$$

où $\sigma_{y,\text{intra}}^2$ est la variance intra–srates, i.e.,

$$\sigma_{y,\text{intra}}^2 = \frac{1}{N} \sum_{h=1}^{H} N_h \sigma_{y,h}^2,$$

et $\sigma_{y,\text{inter}}^2$ est la variance inter–srates, i.e.,

$$\sigma_{y,\text{inter}}^2 = \frac{1}{N} \sum_{h=1}^{H} N_h (\mu_{y,h} - \mu_y)^2.$$

4.2 Échantillons, probabilités d'inclusion et estimation

Définition 4.2.1. Un sondage est dit **stratifié** si, pour chaque strate, on tire un échantillon selon un sondage aléatoire simple sans remise de taille fixe n_h et que les tirages au sein de chaque strate sont mutuellement indépendant.

Soit S_h l'échantillon aléatoire tiré dans la strate $\mathcal{U}_h$ à l'aide d'un plan de sondage $p_h(\cdot)$. L'échantillon aléatoire S obtenu au final est donc

$$S = \bigcup_{h=1}^{H} S_h.$$

Le plan de sondage associé $p(\cdot)$ n'est rien d'autre que

$$p(s) = \prod_{h=1}^{H} p_h(s_h), \qquad s = \bigcup_{h=1}^{H} s_h,$$

et la taille de l'échantillon S est

$$n = \sum_{h=1}^{H} n_h.$$

Le calcul des probabilités d'inclusion pour un sondage stratifié n'est pas difficile mais il faut tout de même faire attention. Pour les probabilités d'inclusion d'ordre un et si l'unité k appartient à la strate $\mathcal{U}_h$ alors

$$\pi_k = \frac{n_h}{N_h},$$

puisqu'on a effectué un plan simple sans remise de taille n_h pour cette strate.

Pour les probabilités d'inclusion d'ordre deux, c'est un peu plus difficile et le résultat dépend du fait ou non que les unités k et ℓ appartiennent à la même strate ou non.

— Si k et ℓ appartiennent à la même strate $\mathcal{U}_h$ alors

$$\pi_{k\ell} = \frac{n_h(n_h-1)}{N_h(N_h-1)}.$$

— Si k et ℓ appartiennent à deux strates différentes $\mathcal{U}_{h_1}$ et $\mathcal{U}_{h_2}$ alors (par indépendance entre les strates)

$$\pi_{kl} = \pi_k \pi_\ell = \frac{n_{h_1}}{N_{h_1}} \frac{n_{h_2}}{N_{h_2}}.$$

En conséquence on a

$$\Delta_{k\ell} = \begin{cases} \frac{n_h}{N_h}\left(1 - \frac{n_h}{N_h}\right), & k = \ell,\ k \in \mathcal{U}_h, \\ -\frac{n_h(N_h - n_h)}{N_h^2(N_h-1)}, & k \neq \ell,\ k, \ell \in \mathcal{U}_h, \\ 0, & k \in \mathcal{U}_h,\ \ell \notin \mathcal{U}_h. \end{cases}$$

Du coup les π–estimateurs du total t_y et de la moyenne μ_y sont

$$\hat{t}_{y,\text{strat}} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \sum_{k \in S_h} \frac{N_h y_k}{n_h} = \sum_{h=1}^{H} \hat{t}_{y,h},$$

et

$$\hat{\mu}_{y,\text{strat}} = \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in S_h} y_k = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h,$$

où $\hat{t}_{y,h}$ est l'estimateur du total pour la strate h, i.e.,

$$\hat{t}_{y,h} = \frac{N_h}{n_h} \sum_{k \in S_h} y_k,$$

et $\bar{y}_h$ est la moyenne de l'échantillon prélevé sur la strate h, i.e.,

$$\bar{y}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k.$$

Puisque les strates sont indépendantes, la variance de ces estimateurs se calcule facilement

$$\text{Var}(\hat{t}_{y,\text{strat}}) \overset{\text{ind}}{=} \sum_{h=1}^{H} \text{Var}(\hat{t}_{y,h}) \overset{\text{simple}}{=} \sum_{h=1}^{H} N_h(N_h - n_h) \frac{S_{y,h}^2}{n_h},$$

variance qui s'estime sans biais par

$$\widehat{\text{Var}}(\hat{t}_{y,\text{strat}}) = \sum_{h=1}^{H} N_h(N_h - n_h) \frac{\widehat{S_{y,h}^2}}{n_h},$$

avec

$$\widehat{S_{y,h}^2} = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_h)^2, \qquad h = 1, \ldots, H.$$

4.3 Plan stratitifié et allocation proportionnelle

Définition 4.3.1. Un plan stratifié est dit à **allocation proportionnelle** si

$$\frac{n_h}{N_h} = \frac{n}{N}, \qquad h = 1, \dots, H,$$

c'est à dire que les « strates de tailles importantes » auront plus d'unité dans l'échantillon que celles de « tailles plus petites ».

Remarque. Généralement la taille d'échantillon pour chaque strate

$$n_h = n\frac{N_h}{N}$$

ne sera pas entière mais afin de simplifier les développements théoriques qui viennent nous allons tout de même le supposer... No comment !

Les π–estimateur du total et de la moyenne sont alors

$$\hat{t}_{y,\text{strat. prop.}} = \sum_{h=1}^{H} \hat{t}_{y,h} = \frac{N}{n} \sum_{k \in S} y_k,$$
$$\hat{\mu}_{y,\text{strat. prop.}} = \frac{1}{n} \sum_{k \in S} y_k.$$

La variance du total est alors

$$\begin{aligned}
\text{Var}(\hat{t}_{y,\text{strat. prop.}}) &= \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h} \\
&= \sum_{h=1}^{H} N_h \left(\frac{N}{n} - 1\right) S_{y,h}^2 \\
&= \frac{N-n}{n} \sum_{h=1}^{H} N_h S_{y,h}^2.
\end{aligned}$$

Remarque. Lorsque les tailles des strates N_h sont suffisamment grandes, alors $S_{y,h}^2 \approx \sigma_{y,h}^2$ et donc

$$\text{Var}(\hat{t}_{y,\text{strat. prop.}}) \approx \frac{N-n}{n} \sum_{h=1}^{H} N_h \sigma_{y,h}^2 = N(N-n)\frac{\sigma_{y,\text{intra}}^2}{n},$$

alors que la variance de l'estimateur du total pour un plan simple sans remise vérifie

$$\text{Var}(\hat{t}_{y,\pi}) \approx N(N-n)\frac{\sigma_y^2}{n}.$$

Les deux expressions sont quasiment identiques mais puisque

$$\sigma_y^2 = \sigma_{y,\text{intra}}^2 + \sigma_{y,\text{inter}}^2,$$

la première expression est plus petite, i.e., on obtient de meilleurs résultat avec un plan stratifié avec allocation proportionnelle qu'avec un plan simple sans remise !

Ceci est bien entendu d'autant plus vrai que la variance inter-strate sera grande, ce qui est le cas lorsque le caractère d'intérêt y dépend fortement du caractère servant à la stratification, ici les tailles N_h.

Bien entendu on estimera sans biais cette variance par

$$\widehat{\text{Var}}(\hat{t}_{y,\text{strat. prop.}}) = \frac{N-n}{n} \sum_{h=1}^{H} N_h \widehat{S_{y,h}^2},$$

avec

$$\widehat{S_{y,h}^2} = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_h)^2, \qquad h = 1, \dots, H.$$

4.4 Plan stratifié optimal pour le total

Si notre intérêt est d'estimer un total ou une moyenne alors il existe une taille optimale pour les strates. On cherche donc les tailles d'échantillon $n_1, \dots, n_h$ minimisant la variance du π–estimateur du total t_y pour une taille d'échantillon fixée n, i.e., minimiser

$$\operatorname{Var}(\hat{t}_{y,\text{strat}}) = \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h}$$

par rapport aux n_h et sous la contrainte

$$\sum_{h=1}^{H} n_h = n.$$

Le Lagrangien de ce problème de minimisation est

$$\mathscr{L}(n_1, \dots, n_H, \lambda) = \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h} + \lambda\left(\sum_{h=1}^{H} n_h - n\right).$$

On a donc

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial n_h} = 0 &\iff -\frac{N_h^2}{n_h^2}S_{y,h}^2 + \lambda = 0\\ &\iff n_h = \frac{N_h S_{y,h}}{\sqrt{\lambda}}.\end{aligned}$$

Mais puisque $\sum_h n_h = n$ on a

$$\lambda^{-1/2}\sum_{h=1}^{H} N_h S_{y,h} = n,$$

et il vient

$$n_h = n\frac{N_h S_{y,h}}{\sum_{j=1}^{H} N_j S_{y,j}}, \qquad h = 1, \dots, H.$$

Remarque. La taille optimale pour une strate $\mathcal{U}_h$ est donc proportionnelle au produit de la taille de cette strate et de l'écart-type du caractère y sur cette strate.

Bien entendu en pratique on ne connaîtra pas $S_{y,h}$ et donc la formule précédente n'est pas d'un grand intérêt. Elle est cependant assez instructive—et intuitive ! Instructive puisqu'elle indique qu'il faut surreprésenter les strates qui ont une forte variabilité ; ce qui est intuitif non ?

Remarque. En pratique les tailles $n_h \notin \mathbb{N}$ et on arrondira les résultats. De plus il peut arriver également que $n_h > N_h$ pour un $h \in \{1, \dots, H\}$. Pour de tels cas, on posera alors $n_h = N_h$ et on déterminera les tailles optimales sur les strates restantes — en itérant le procédé si nécessaire.

Supposons que nos tailles optimales soient des entiers et telles que $n_h < N_h$ pour tout h. Alors la variance du π–estimateur est alors

$$\begin{aligned}\operatorname{Var}(\hat{t}_{y,\text{opt}}) &= \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h}\\ &= \sum_{h=1}^{H} N_h^2\frac{\sum_{\ell=1}^{H} N_\ell S_{y,\ell}}{nN_h S_{y,h}}S_{y,h}^2 - \sum_{h=1}^{H} N_h S_{y,h}^2\\ &= \left(\frac{\sum_{\ell=1}^{H} N_\ell S_{y,\ell}}{n}\right)\sum_{h=1}^{H} N_h S_{y,h} - \sum_{h=1}^{H} N_h S_{y,h}^2\\ &= \frac{1}{n}\left(\sum_{h=1}^{H} N_h S_{y,h}\right)^2 - \sum_{h=1}^{H} N_h S_{y,h}^2.\end{aligned}$$

4.5 Prise en compte du coût

Faire une enquête est bien souvent coûteux de sorte que l'allocation optimale présentée dans la Section 4.4 sera bien souvent déconnectée de la réalité. Bien souvent on visera plutôt une allocation optimale pour un budget fixé C. Nous allons donc minimiser la variance de l'estimateur du total

$$\mathrm{Var}(\hat{t}_{y,\mathrm{strat}}) = \sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h},$$

sous la contrainte

$$\sum_{h=1}^{H} n_h C_h = C,$$

o C_h représente le coût d'interroger une unité dans la strate $\mathcal{U}_h$.

Exercice 3. Montrez que la taille optimale est alors

$$n_h = \frac{C N_h S_{y,h}}{\sqrt{C_h}\sum_{\ell=1}^{H} N_\ell S_{y,\ell}\sqrt{C_\ell}}, \qquad h = 1, \ldots, H.$$

Solution.

□

Remarque. De manière assez logique nous constatons que nous sélectionnons moins les strates les plus « coûteuses ».

Chapitre 5

Plans par grappes et à plusieurs degrés

Dans ce chapitre nous allons voir comment une variable auxiliaire peut être utilisée non pas pour améliorer la précision de nos estimations mais le déroulement d'une enquête !

5.1 Plans par grappes

Les plans par grappes ressemblent (aux premiers abords) fortement aux plans stratifiés. Ce n'est pourtant pas du tout le cas !!!

Supposons que la population $\mathcal{U}$ soit partitionnée en M sous-ensembles $\mathcal{U}_1, \ldots, \mathcal{U}_M$ appelés grappes et tels que

$$\bigcup_{i=1}^{M} \mathcal{U}_i = \mathcal{U}, \qquad \mathcal{U}_i \cap \mathcal{U}_j = \emptyset, \quad i \neq j.$$

Chaque grappe $\mathcal{U}_i$ admet une taille N_i et l'on a bien évidemment

$$\sum_{i=1}^{M} N_i = N,$$

où N est la taille de la population $\mathcal{U}$.

Notre but étant toujours d'estimer un total ou une moyenne, remarquons que le total (resp. la moyenne) s'écrit à l'aide des grappes

$$t_y = \sum_{k \in \mathcal{U}} y_k = \sum_{i=1}^{M} \sum_{k \in \mathcal{U}_i} y_k = \sum_{i=1}^{M} t_{y,i},$$

où $t_{y,i}$ est le total des valeurs prises par le caractère y sur la grappe $\mathcal{U}_i$, i.e.,

$$t_{y,i} = \sum_{k \in \mathcal{U}_i} y_k.$$

De même la moyenne sur la population s'écrit

$$\mu_y = \frac{1}{N} \sum_{k \in \mathcal{U}} y_k = \frac{1}{N} \sum_{i=1}^{M} \sum_{k \in \mathcal{U}_i} y_k = \frac{1}{N} \sum_{i=1}^{M} N_i \mu_{y,i},$$

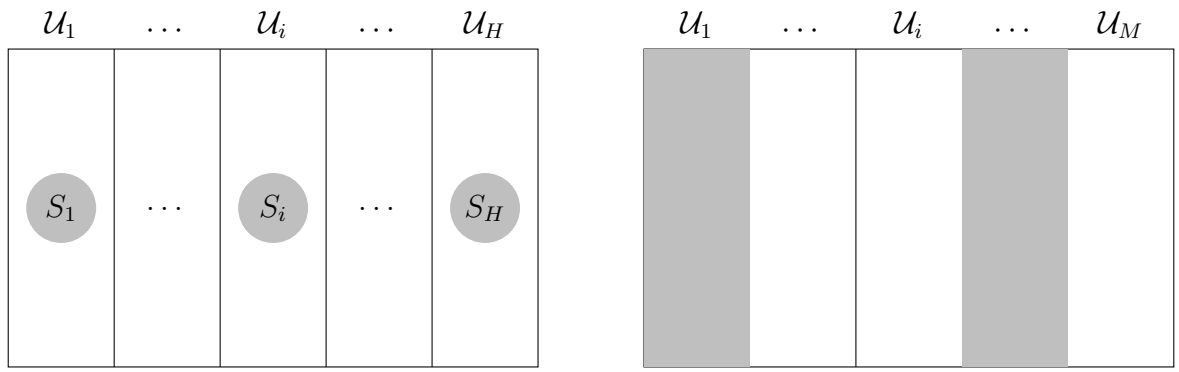


Figure 5.1: *Illustration de la différence entre un plan de sondage stratifié (gauche) et par grappes (droite). Pour l'un, un échantillon aléatoire est prélevé dans chaque strate. Pour l'autre un échantillon aléatoire sur les grappes est prélevé et chaque grappe ainsi piochée est entièrement retenue.*

où $\mu_{y,i}$ est la moyenne des valeurs prises par le caractère y sur la grappe $\mathcal{U}_i$, i.e.,

$$\mu_{y,i} = \frac{1}{N_i} \sum_{k \in \mathcal{U}_i} y_k.$$

On définit également la variance et la variance corrigée sur une grappe $\mathcal{U}_i$ par

$$\sigma^2_{y,i} = \frac{1}{N_i} \sum_{k \in \mathcal{U}_i} (y_k - \mu_{y,i})^2,$$

et

$$S^2_{y,i} = \frac{1}{N_i - 1} \sum_{k \in \mathcal{U}_i} (y_k - \mu_{y,i})^2.$$

Jusque là rien de bien nouveau par rapport à la manière dont nous avions introduit les plans stratifiés me direz vous. C'est à ce moment bien précis que les deux approches divergent !!!

Définition 5.1.1. Un plan est dit **par grappes** si l'on procède comme suit :

1. On sélectionne un échantillon aléatoire de grappes S_g selon un plan de sondage $p_g(\cdot)$ défini sur les parties non vides de $\mathcal{U}_g = \{1, \dots, M\}$;
2. Toutes les unités des grappes sélectionnées sont alors retenues.

La Figure 5.1 illustre la différence entre ces deux plans de sondages. Nous voyons clairement que le plan stratifié utilise un échantillons dans chaque strates alors que le plan par grappes sélectionne soit totalement une grappe soit pas du tout. Ainsi un échantillon aléatoire S issu d'un plan par grappes s'écrit

$$S = \bigcup_{i \in S_g} \mathcal{U}_i,$$

et sa taille n_S est

$$n_S = \sum_{i \in S_g} N_i.$$

Remarque. La taille de l'échantillon n_S sera le plus souvent aléatoire même si le plan de sondage sur les grappes $p_g(\cdot)$ est à taille fixe — les grappes n'ayant pas forcément des tailles identiques.

Les probabilités d'inclusion d'ordre un et deux découlent des probabilités de sélection des grappes (et donc du plan de sondage $p_g(\cdot)$). Ainsi si l'unité k appartient à la grappe i, on a

$$\pi_k = \sum_{\substack{s\in\mathscr{S}_g \\ i\in s}} p_g(s) \stackrel{\text{def}}{=} \pi_{g,i}, \qquad k\in\mathcal{U}_i, \quad i\in\mathcal{U}_g,$$

où $\mathscr{S}_g$ est l'ensemble des échantillons possibles de $\mathcal{U}_g$.

Les probabilités d'inclusions d'ordre deux s'écrivent de manière analogue

$$\pi_{k\ell} = \begin{cases} \pi_{g,i}, & k,\ell\in\mathcal{U}_i \\ \pi_{g,ij}, & k\in\mathcal{U}_i,\ \ell\in\mathcal{U}_j, \end{cases}$$

avec

$$\pi_{g,ij} = \sum_{\substack{s\in\mathscr{S}_g \\ i,j\in s}} p_g(s), \qquad i,j\in\mathcal{U}_g, \quad i\neq j.$$

Exercice 4. Montrez que les conditions de Sen–Yates–Grundy, i.e., $\Delta_{k\ell}<0$, ne sont pas satisfaites lorsque k et ℓ appartiennent à la même grappe.

Solution.

□

Les π–estimateurs du total et de la moyenne sont

$$\hat{t}_{y,\pi} = \sum_{i\in S_g} \frac{t_{y,i}}{\pi_{g,i}}, \qquad \hat{\mu}_{y,\pi} = \frac{1}{N}\sum_{i\in S_g}\frac{N_i\mu_{y,i}}{\pi_{g,i}}.$$

Notons toutefois que, pour les plans par grappes, il est rare que la taille de la population N soit connue. On utilisera plutôt le ratio de Hájek de la Section 1.7 pour estimer la moyenne μ_y.

La variance du π–estimateur du total t_y est, cf. Section 1.6.2,

$$\begin{aligned} \mathrm{Var}(\hat{t}_{y,\pi}) &= \sum_{k,\ell\in\mathcal{U}} \frac{y_k y_\ell}{\pi_k\pi_\ell}\Delta_{k\ell} \\ &= \sum_{i,j=1}^{M} \frac{t_{y,i}t_{y,j}}{\pi_{g,i}\pi_{g,j}}\Delta_{g,ij} \\ &= \sum_{i=1}^{M} \frac{t_{y,i}^2}{\pi_{g,i}^2}\pi_{g,i}(1-\pi_{g,i}) + \sum_{i\neq j}\frac{t_{y,i}t_{y,j}}{\pi_{g,i}\pi_{g,j}}(\pi_{g,ij}-\pi_{g,i}\pi_{g,j}), \end{aligned}$$

que l'on estimera classiquement par le π–estimateur.

Si le nombre de grappe sélectionné est fixe, alors on peut écrire cette variance sous une autre forme (cf. Section 1.6.3),

$$\mathrm{Var}(\hat{t}_{y,\pi}) = -\frac{1}{2}\sum_{\substack{i,j=1 \\ i\neq j}}^{M}\left(\frac{t_{y,i}}{\pi_{g,i}} - \frac{t_{y,j}}{\pi_{g,j}}\right)^2\Delta_{g,ij}. \tag{5.1}$$

5.2 Choix sur le plan de sondage $p_g(\cdot)$

5.2.1 Tirage des grappes à probabilités égales

La première idée venant à l'esprit pour le choix de $p_g(\cdot)$ est de faire un plan de sondage sans remise et à taille fixe m. Pour ce choix nous avons alors

$$\pi_{g,i} = \frac{m}{M}, \qquad \pi_{g,ij} = \frac{m(m-1)}{M(M-1)}, \qquad i,j = 1,\dots,M, \quad i \neq j.$$

Cela dit bien que $p_g(\cdot)$ soit de taille fixe, la taille n_S de l'échantillon S obtenue est comme nous l'avons déjà dit aléatoire et vaut en espérance

$$\mathbb{E}(n_S) = \mathbb{E}\left(\sum_{i\in S_g} N_i\right) = \sum_{i=1}^{M} N_i \mathbb{E}(1_{\{i\in S_g\}}) = \sum_{i=1}^{M} N_i \pi_{g,i} = \frac{mN}{M}.$$

Les π–estimateurs du total et de la moyenne se simplifient en

$$\hat{t}_{y,\pi} = \frac{M}{m}\sum_{i\in S_g} t_{y,i}, \qquad \hat{\mu}_{y,\pi} = \frac{M}{mN}\sum_{i\in S_g} N_i \mu_{y,i}.$$

Puisque $p_g(\cdot)$ est à taille fixe, la variance s'écrit d'après (5.1)

$$\begin{aligned}
\mathrm{Var}(\hat{t}_{y,\pi}) &= -\frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{M}\left(\frac{t_{y,i}}{\pi_{g,i}} - \frac{t_{y,j}}{\pi_{g,j}}\right)^2 \Delta_{g,ij} \\
&= -\frac{M^2}{2m^2}\sum_{i\neq j}(t_{y,i}-t_{y,j})^2\left\{\frac{m(m-1)}{M(M-1)} - \frac{m^2}{M^2}\right\} \\
&= -\frac{M^2}{2m^2}\frac{m(m-M)}{M^2(M-1)}\sum_{i\neq j}(t_{y,i}-t_{y,j})^2 \\
&= \frac{M-m}{M-1}\frac{M}{m}\sum_{i=1}^{M}\left(t_{y,i} - \frac{t_y}{M}\right)^2,
\end{aligned}$$

où on a utilisé pour la dernière équation le fait que

$$\sum_{i,j=1}^{n}(x_i - x_j)^2 = 2n\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Il est peut-être plus parlant d'écrire cette dernière expression de la manière suivante

$$\mathrm{Var}(\hat{t}_{y,\pi}) = M(M-m)\frac{\frac{1}{M-1}\sum_{i\in S_g}(t_{y,i} - t_y/M)^2}{m},$$

qui nous fait furieusement penser à l'expression vue à maintes reprises

$$\mathrm{Var}(\hat{t}_{y,\pi}) = N(N-n)\frac{S_y^2}{n},$$

mais où la variance corrigée est maintenant calculée sur les sous–totaux des grappes—ce qui est logique non ?

On estimera bien entendu cette variance par

$$\widehat{\mathrm{Var}}(\hat{t}_{y,\pi}) = \frac{M-m}{m-1}\frac{M}{m}\sum_{i\in S_g}\left(t_{y,i} - \frac{\hat{t}_y}{M}\right)^2.$$

5.2.2 Tirage proportionnel aux tailles des grappes

On peut également effectuer un plan sans remise de taille fixe m dont les probabilités de sélection sont proportionnelles à la taille de chacune des grappes comme nous l'avons vu lors de la Section 3.1.

Pour simplifier les choses on supposera que $mN_i \leq N$ pour tout $i = 1, \dots, M$ — sinon les grappes ne vérifiant pas cela seront systématiquement choisies. Les probabilités de sélection des grappes sont alors

$$\pi_{g,i} = \frac{mN_i}{N}, \qquad i = 1, \dots, M.$$

La taille n_S de l'échantillon S est toujours aléatoire et vaut en moyenne

$$\mathbb{E}(n_S) = \mathbb{E}\left(\sum_{i \in S_g} N_i\right) = \sum_{i=1}^{M} N_i \pi_{g,i} = \frac{m}{N} \sum_{i=1}^{M} N_i^2.$$

Les π–estimateurs du total et de la moyenne sont

$$\hat{t}_{y,\pi} = \frac{N}{m} \sum_{i \in S_g} N_i t_{y,i} = \frac{N}{m} \sum_{i \in S_g} \mu_{y,i}, \qquad \hat{\mu}_{y,\pi} = \frac{1}{m} \sum_{i \in S_g} \mu_{y,i},$$

et puisque le plan de sondage $p_g(\cdot)$ est à taille fixe, la variance s'écrit d'après (5.1)

$$\begin{aligned}
\mathrm{Var}(\hat{t}_{y,\pi}) &= -\frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \left(\frac{t_{y,i}}{\pi_i} - \frac{t_{y,j}}{\pi_j}\right)^2 \Delta_{g,ij} \\
&= -\frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \left(\frac{N t_{y,i}}{m N_i} - \frac{N t_{y,j}}{m N_j}\right)^2 \left(\pi_{g,ij} - \frac{mN_i}{N} \frac{mN_j}{N}\right) \\
&= -\frac{N^2}{2M^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} (\mu_{y,i} - \mu_{y,j})^2 \left(\pi_{g,ij} - \frac{m^2 N_i N_j}{N^2}\right).
\end{aligned}$$

Remarque. Nous ne pouvons pas aller plus loin dans le calcul de cette variance car de manière générale les probabilités d'inclusion d'ordre 2 ne sont pas connues pour les tirages proportionnels, cf. Chapitre 3.

5.3 Plans à deux degrés

Les plans à deux degrés portent bien leurs noms puisqu'ils consistent en un double échantillonnage :

1. sur les unités primaires ;
2. puis les unités secondaires.

En exemple valant mille mots, pour un sondage sur les ménages, les unités primaires seraient les communes alors que les unités secondaires seraient les ménages. Un plan à deux degrés consisterait donc à échantillonner les communes puis à prélever, pour chaque commune retenue, un échantillon de ménages.

Remarque. C'est un peu la stratégie de diviser pour mieux régner et cela permet parfois de réduire les coût de l'enquête. En effet pour notre exemple sur les ménages, les unités (secondaires) seront forcément proches car issues de la même commune. Imaginez la facture d'essence si l'on avait échantillonné directement sur les ménages français !

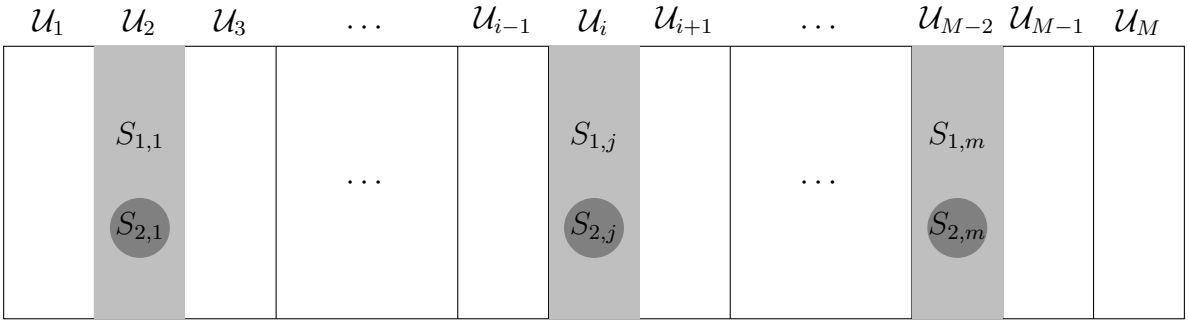


Figure 5.2: *Illustration du concept de plan à deux degrés. L'échantillon du premier degré est de taille m et est $S_1 = \cup_{j=1}^{m} S_{1,j}$. L'échantillon « final » obtenu par un plan à deux degrés est alors $S = \cup_{j \in S_1} S_{2,j}$.*

5.3.1 Population, untiés primaires et secondaires

Comme pour les sections précédentes, on supposera que la population $\mathcal{U} = \{1, \ldots, N\}$ est subdivisée en M sous–populations $\mathcal{U}_i$, $i = 1, \ldots, M$, que l'on appellera **unités primaires**. Les unités primaires sont composées de N_i **unités secondaires** et l'on a bien entendu

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

Pour effectuer un plan à deux degrés, il faut donc

— construire un échantillon S_1 d'unités primaires à partir d'un plan de sondage $p_1(\cdot)$ sur $\{1, \ldots, M\}$;

— pour chaque unité primaire sélectionnée, construire un échantillon S_2 sur les unités secondaires à partir d'un plan de sondage $p_2(\cdot)$.

Il est souhaitable que les plans à deux degrés possèdent les deux propriétés suivantes :

Invariance : le plan du second degré $p_2(\cdot)$ est indépendant du premier plan $p_1(\cdot)$, i.e., $\Pr(S_2 = s_2 \mid S_1 = s_1) = \Pr(S_2 = s_2)$;

Indépendance : les tirages du second degré sont mutuellement indépendants.

La Figure 5.2 essaye d'illustrer le principe de fonctionnement d'un plan de sondage à deux degrés. Clairement l'échantillon obtenu par de tels plan s'écrit

$$S = \bigcup_{i \in S_1} S_{2,i},$$

et sa taille (aléatoire) est

$$n_S = \sum_{i \in S_1} n_i, \qquad n_i = |S_{2,i}|.$$

Notons $\pi_{1,i}$ et $\pi_{1,ij}$ les probabilités d'inclusion d'ordre 1 et 2 pour le premier degré, i.e.,

$$\pi_{1,i} = \Pr(\mathcal{U}_i \in S_1), \qquad \pi_{1,ij} = \Pr(\mathcal{U}_i \in S_1, \mathcal{U}_j \in S_1).$$

Notons également $\pi_{k|i}$ la probabilité de sélectionner l'unité (secondaire) k sachant que l'unité (primaire) $\mathcal{U}_i$ a été choisie. De manière analogue on notera $\pi_{k\ell|i}$ la probabilité d'inclusion d'ordre 2 sachant que $\mathcal{U}_i$ a été retenue.

Avec ces notations, pour un $k \in \mathcal{U}_i$, la probabilité d'inclusion (usuelle) π_k s'écrit

$$\pi_k = \Pr(k \in S_{2,i}, i \in S_1) = \Pr(k \in S_{2,i} \mid i \in S_1) \Pr(i \in S_1) = \pi_{k|i} \pi_{1,i}.$$

Un même raisonnement nous conduit aux expressions pour les probabilités d'inclusion d'ordre 2,

$$\pi_{k\ell} = \begin{cases} \pi_{k\ell|i}\pi_{1,i}, & k,\ell \in \mathcal{U}_i, \\ \pi_{k|i}\pi_{\ell|j}\pi_{1,ij}, & k \in \mathcal{U}_i,\ \ell \in \mathcal{U}_j,\ i \neq j, \end{cases}$$

où pour le deuxième cas nous nous sommes servit de l'hypothèse d'indépendance pour le deuxième tirage.

5.3.2 Le π–estimateur

Rappelons que dans ce contexte le total t_y s'écrit

$$t_y = \sum_{k\in\mathcal{U}} y_k = \sum_{i=1}^{M} \sum_{k\in\mathcal{U}_i} y_k = \sum_{i=1}^{M} t_{y,i},$$

où $t_{y,i}$ est le total pour l'unité primaire $\mathcal{U}_i$, i.e.,

$$t_{y,i} = \sum_{k\in\mathcal{U}_i} y_k.$$

Le π–estimateur de ce total est donc

$$\hat{t}_{y,\pi} = \sum_{i\in S_1} \sum_{k\in S_{2,i}} \frac{y_k}{\pi_{k|i}\pi_{1,i}} = \sum_{i\in S_1} \frac{\hat{t}_{y,i}}{\pi_{1,i}},$$

où $\hat{t}_{y,i}$ est bien entendu le π–estimateur du (sous) total $t_{y,i}$, i.e.,

$$\hat{t}_{y,i} = \sum_{k\in S_{2,i}} \frac{y_k}{\pi_{k|i}}.$$

Remarque. On peut tout a fait calculer la variance du π–estimateur, mais nous ne le ferons pas...

Chapitre 6

Utilisation d'une information auxiliaire

Dans ce chapitre nous allons voir comment nous pouvons bénéficier de l'utilisation d'une information auxiliaire qui était non disponible lors de la mise en oeuvre du sondage. Le but étant bien entendu d'obtenir de meilleures estimation du paramètre d'intérêt.

6.1 Post-stratification

6.1.1 Notations

Lorsque l'on parle d'utilisation d'une information auxiliaire, il faut à tout prix connaître l'approche dite de **post-stratification**. Cette méthode fait office de référence et a en plus le bon goût d'être particulièrement simple !

On suppose que le caractère auxiliaire est **qualitatif** et peut prendre H valeurs distinctes disons $\{1, \ldots, H\}$. Ce caractère auxiliaire nous permet ainsi de former une partition de la population $\mathcal{U}$, i.e.,

$$\mathcal{U} = \bigcup_{h=1}^{H} \mathcal{U}_h, \qquad \mathcal{U}_h = \{i \in \mathcal{U} \colon y_i = h\}.$$

Remarque. Le terme post-stratification vient du fait que cette partition de la population $\mathcal{U}$ ressemble à s'y méprendre à la technique de stratification introduite au Chapitre 4. Puisque cette stratification intervient après le sondage, on parlera naturellement de **post**-stratification et de post-strates $\mathcal{U}_h$.

Le nombre d'unités N_h de la post-strate $\mathcal{U}_h$ est appelé la taille de la post-strate et bien entendu

$$N = \sum_{h=1}^{H} N_h.$$

Notons que nous supposons que les N_h sont connus et constituent notre fameuse information auxiliaire.

Comme pour le sondage stratifié, le total et la moyenne s'écrivent

$$t_y = \sum_{k \in \mathcal{U}} y_k = \sum_{h=1}^{H} \sum_{k \in \mathcal{U}_h} y_k = \sum_{h=1}^{H} N_h \mu_h$$

$$\mu_y = \frac{1}{N} \sum_{h=1}^{H} N_h \mu_h,$$

où μ_h est la moyenne sur la post-strate $\mathcal{U}_h$, i.e.,

$$\mu_h = \frac{1}{N_h} \sum_{k \in \mathcal{U}_h} y_k, \qquad h = 1, \ldots, H.$$

Nous pouvons également s'intéresser à la variance (corrigée) pour chaque post-strate

$$\sigma^2_{y,h} = \frac{1}{N_h} \sum_{k \in \mathcal{U}_h} (y_k - \mu_h)^2, \qquad S^2_{y,h} = \frac{1}{N_h - 1} \sum_{k \in \mathcal{U}_h} (y_k - \mu_h)^2, \qquad h = 1, \ldots, h.$$

Exercice 5. Montrez que l'on peut décomposer la variance totale σ^2_y à l'aide des variances des post-strates, i.e.,

$$\sigma^2_y = \frac{1}{N} \sum_{h=1}^{H} N_h \sigma^2_{y,h} + \frac{1}{N} \sum_{h=1}^{H} N_h (\mu_{y,h} - \mu_y)^2.$$

Solution.

□

6.1.2 L'estimateur post-stratifié

Supposons qu'un échantillon alétoire S de taille n ait été tiré au sein d'une population $\mathcal{U}$ de taille N à l'aide d'un plan simple sans remise. Le π–estimateur du total t_y est donc

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{y_k}{n/N} = \frac{N}{n} \sum_{k \in S} y_k = \frac{N}{n} \sum_{\substack{h=1 \\ n_h > 0}}^{H} n_h \hat{\mu}_{y,h},$$

où n_h est la taille des post-strates et

$$\hat{\mu}_{y,h} = \frac{1}{n_h} \sum_{k \in S_h} y_k.$$

L'estimateur post-stratifié s'écrit alors

$$\hat{t}_{y,\text{post}} = \sum_{\substack{h=1 \\ n_h>0}}^{H} N_h \hat{\mu}_{y,h}.$$

Remarque. La connaissance des tailles N_h est nécessaire afin d'utiliser cet estimateur.

6.1.3 Propriété de l'estimateur

Le calcul de l'espérance de $\hat{t}_{y,\text{post}}$ est quelque peu compliqué du fait que les tailles n_h des échantillons des post-strates sont aléatoires. Commençons par calculer cette espérance sachant les tailles n_h. Nous avons

$$\begin{aligned}
\mathbb{E}(\hat{t}_{y,\text{post}} \mid n_1, \ldots, n_H) &= \sum_{\substack{h=1 \\ n_h>0}}^{H} N_h \mathbb{E}(\hat{\mu}_{y,h} \mid n_1, \ldots, n_H) \\
&= \sum_{\substack{h=1 \\ n_h>0}}^{H} t_{y,h} \\
&= t_y - \sum_{\substack{h=1 \\ n_h=0}}^{H} t_{y,h}.
\end{aligned}$$

Puisque $\mathbb{E}\{\mathbb{E}(X \mid Y)\} = \mathbb{E}(X)$, nous avons

$$\mathbb{E}(\hat{t}_{y,\text{post}}) = t_y - \sum_{h=1}^{H} t_{y,h} \Pr(n_h = 0).$$

Or puisque

$$\Pr(n_h = 0) = \Pr(\nexists k \in S \colon k \in \mathcal{U}_h) = \frac{\binom{N-N_h}{n}}{\binom{N}{n}} = \frac{(N-N_h)!(N-n)!}{(N-N_h-n)!N!},$$

on a donc

$$\mathbb{E}(\hat{t}_{y,\text{post}}) - t_y = \sum_{h=1}^{H} t_{y,h} \frac{(N-N_h)!(N-n)!}{(N-N_h-n)!N!}.$$

Remarque. L'estimateur post-stratifié n'est donc pas sans biais mais est approximativement sans biais dès lors que $\Pr(n_h = 0)$ est suffisamment faible pour tout $h = 1, \ldots, H$.

Une règle de pouce consiste à ce que les post-strates soient suffisamment grandes, i.e., que les tailles N_h des post-strates vérifient

$$n\frac{N_h}{N} \geq 30, \qquad h = 1, \ldots, H.$$

On peut également calculer la variance de l'estimateur post-stratifié en utilisant la célèbre formule

$$\text{Var}(\hat{t}_{y,\text{post}}) = \text{Var}\{\mathbb{E}(\hat{t}_{y,\text{post}} \mid n_1, \ldots, n_H)\} + \mathbb{E}\{\text{Var}(\hat{t}_{y,\text{post}} \mid n_1, \ldots, n_H)\}.$$

Mais puisque nous avons montré que

$$\mathbb{E}(\hat{t}_{y,\text{post}} \mid n_1, \ldots, n_H) = t_y - \sum_{\substack{h=1 \\ n_h=0}}^{H} t_{y,h},$$

cela implique que $\mathrm{Var}\{\mathbb{E}(\hat{t}_{y,\mathrm{post}} \mid n_1, \ldots, n_H)\} \approx 0$ dès lors que $\Pr(n_h = 0)$ est suffisamment faible. On a donc

$$\begin{aligned}\mathrm{Var}(\hat{t}_{y,\mathrm{post}}) &= \mathbb{E}\{\mathrm{Var}(\hat{t}_{y,\mathrm{post}} \mid n_1, \ldots, n_H)\} \\ &= \mathbb{E}\left\{ \sum_{\substack{h=1 \\ n_h>0}}^{H} N_h(N_h - n_h)\frac{S_{y,h}^2}{n_h} \right\} \\ &\approx \sum_{h=1}^{H} N_h \left\{ N_h \mathbb{E}(n_h^{-1}) - 1 \right\} S_{y,h}^2.\end{aligned}$$

Il reste donc à calculer $\mathbb{E}(n_h^{-1})$ ce qui n'est pas évident. En fait on calculera une approximation de cette espérance en ayant recours à la linéarisation. Ceci est un peu long mais reste tout à fait faisable...

6.2 Caractère auxiliaire quantitatif

Dans la section précédente, nous avons introduit la technique de post-stratification ; mais cette dernière supposait que l'information auxiliaire était qualitative. Parfois cette information auxiliaire sera **quantitative**.

Soit x le caractère auxiliaire (qui est quantitatif rappelons le encore une fois) dont le total

$$t_x = \sum_{k \in \mathcal{U}} x_k$$

est supposé connu.

Si l'on soupçonne que le caractère x soit lié au caractère d'intérêt y, alors on aimerait bien bénéficier de la connaissance de x pour estimer une fonction d'intérêt sur y. Dans cette section, nous allons voir différentes approches de ce type et nous supposerons qu'un **plan de sondage simple est réalisé**. Avant d'introduire ces différentes techniques, posons quelques notations.

6.2.1 Notations

Comme d'habitude on appelera

$$\mu_x = \frac{1}{N}\sum_{k \in \mathcal{U}} x_k, \qquad \mu_y = \frac{1}{N}\sum_{k \in \mathcal{U}} y_k,$$

les moyennes des caractères x et y sur la population et

$$S_x^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu_x)^2, \qquad S_y^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \mu_y)^2,$$

les variances corrigées des caractères x et y sur la population. On introduit également la nouvelle notation

$$S_{xy} = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu_x)(y_k - \mu_y),$$

i.e., la covariance entre le caractère x et le caractère y sur la population.

En ce qui concerne les quantités échantillonnées, on notera

$$\widehat{S_x^2} = \frac{1}{n-1}\sum_{k \in S}(x_k - \hat{\mu}_x)^2, \quad \widehat{S_y^2} = \frac{1}{n-1}\sum_{k \in S}(y_k - \hat{\mu}_y)^2, \quad \widehat{S_{xy}} = \frac{1}{n-1}\sum_{k \in S}(x_k - \hat{\mu}_x)(y_k - \hat{\mu}_y),$$

les variances et la covariance calculées à partir de l'échantillon S de taille n.

6.2.2 Estimation par la différence

L'estimateur par la différence du total t_y, noté $\hat{t}_{y,D}$, est

$$\hat{t}_{y,D} = \hat{t}_{y,\pi} + t_x - \hat{t}_{x,\pi},$$

où $\hat{t}_{x,\pi}$ et $\hat{t}_{y,\pi}$ sont les π–estimateurs des totaux t_x et t_y.

En quelque sorte l'idée de cet estimateur est de reporter l'erreur du π–estimateur commise sur l'estimation de t_x sur l'estimation de t_y.

Exercice 6. Montrez que cet estimateur est sans biais.

Solution.

□

La variance (et donc l'erreur quadratique puisque c'est un estimateur sans biais) se calcule également aisément :

$$\begin{aligned}\mathrm{Var}(\hat{t}_{y,D}) &= \mathrm{Var}(\hat{t}_{y,\pi}) + \mathrm{Var}(\hat{t}_{x,\pi}) - 2\mathrm{Cov}(\hat{t}_{x,\pi}, \hat{t}_{y,\pi}) \\ &= \frac{N(N-n)}{n}\left(S_y^2 + S_x^2 - 2S_{xy}\right).\end{aligned}$$

Cette variance sera bien entendue estimée par

$$\widehat{\mathrm{Var}}(\hat{t}_{y,D}) = \frac{N(N-n)}{n}\left(\widehat{S_y^2} + \widehat{S_x^2} - 2\widehat{S_{xy}}\right).$$

6.2.3 Estimation par le quotient

L'estimateur par le quotient du total t_y, noté $\hat{t}_{y,Q}$, est

$$\hat{t}_{y,Q} = \frac{\hat{t}_{y,\pi}}{\hat{t}_{x,\pi}} t_x.$$

En quelque sorte l'idée de cet estimateur est similaire à celle de l'estimateur par la différence mais cette fois ci l'erreur est reportée de manière multiplicative plutôt qu'additive.

Le biais de cet estimateur n'est pas calculable de manière explicite du fait de la présence d'un quotient. On aura donc recours comme d'habitude à la technique de **linéarisation**.

Puisque

$$\hat{t}_{y,Q} - t_y = \frac{\hat{t}_{y,\pi} - R\hat{t}_{x,\pi}}{\hat{t}_{x,\pi}} t_x = \frac{\hat{t}_{y,\pi} - R\hat{t}_{x,\pi}}{1+\varepsilon},$$

avec $R = t_y/t_x$ et

$$\varepsilon = \frac{\hat{t}_{x,\pi} - t_x}{t_x}.$$

A l'aide d'un développement limité de $(1+\varepsilon)^{-1}$ en $\varepsilon = 0$ et d'ordre 1, on obtient

$$\hat{t}_{y,Q} - t_y \approx (\hat{t}_{y,\pi} - R\hat{t}_{x,\pi})(1-\varepsilon).$$

Au final on peut donc avoir une approximation du biais

$$\begin{aligned}\mathbb{E}(\hat{t}_{y,Q} - t_y) &\approx -\mathbb{E}\left\{(\hat{t}_{y,\pi} - R\hat{t}_{x,\pi})\varepsilon\right\} \\ &= -\frac{\mathbb{E}(\hat{t}_{x,\pi}\hat{t}_{y,\pi}) - t_x t_y - R\mathbb{E}(\hat{t}_{x,\pi}^2) + Rt_x^2}{t_x} \\ &= \frac{R\,\mathrm{Var}(\hat{t}_{x,\pi}) - \mathrm{Cov}(\hat{t}_{x,\pi}, \hat{t}_{y,\pi})}{t_x} \\ &= \frac{N(N-n)}{n}\frac{RS_x^2 - S_{xy}}{t_x}.\end{aligned}$$

Remarque. Le biais devient négligeable dès lors que n est grand.

Exercice 7. Calculez une approximation de l'erreur quadratique de l'estimateur par quotient.

Solution.

□

6.2.4 Estimation par la régression

L'estimateur du total t_y par la régression est

$$\hat{t}_{y,R} = \hat{t}_{y,\pi} + \hat{a}(t_x - \hat{t}_{x,\pi}), \qquad \hat{a} = \frac{\widehat{S_{xy}}}{\widehat{S_x^2}}.$$

L'idée de cet estimateur est de supposer qu'il existe une relation linéaire de la forme $y = ax + b$ entre les caractères x et y et donc que

$$t_y \approx \hat{a}t_x + \hat{b}, \qquad \hat{t}_{y,\pi} \approx \hat{a}\hat{t}_{x,\pi} + \hat{b}.$$

On estime alors le total par

$$\hat{t}_{y,\pi} + (t_y - \hat{t}_{y,\pi}) = \hat{t}_{y,\pi} + \hat{a}(t_x - \hat{t}_{x,\pi}).$$

Comme pour les estimateurs précédents, le calcul de l'espérance de $\hat{t}_{y,R}$ ne peut être qu'approché. Puisque

$$\hat{t}_{y,R} = \hat{t}_{y,\pi} + a(t_x - \hat{t}_{x,\pi}) + (\hat{a} - a)(t_x - \hat{t}_{x,\pi}), \qquad a = \frac{S_{xy}}{S_x^2},$$

Table 6.1: *Récapitulatif des différentes méthodes de redressement à l'aide d'une variable quantitative.*

Estimateur	Définition	$\left\{\frac{N(N-n)}{n}\right\}^{-1} \times$ EQM
π–estimateur	$\hat{t}_{y,\pi} = n^{-1}N\sum_{k\in S} y_k$	S_y^2
par la différence	$\hat{t}_{y,D} = \hat{t}_{y,\pi} + t_x - \hat{t}_{x,\pi}$	$S_y^2 + S_x^2 - 2S_{xy}$
par le quotient	$\hat{t}_{y,Q} = \hat{t}_{y,\pi} t_x / \hat{t}_{x,\pi}$	$S_y^2 + R^2 S_x^2 - 2RS_{xy}$
par la régression	$\hat{t}_{y,R} = \hat{t}_{y,\pi} + \hat{a}(t_x - \hat{t}_{x,\pi})$	$S_y^2(1-\rho^2)$

et où l'on peut montrer (admis) que le dernier terme est négligeable, on a donc

$$\mathbb{E}(\hat{t}_{y,R}) \approx \mathbb{E}\{\hat{t}_{y,\pi} + a(t_x - \hat{t}_{x,\pi})\} = t_y.$$

L'erreur quadratique est approchée par

$$\begin{aligned}
\mathrm{EQM}(\hat{t}_{y,R}) &\approx \mathrm{Var}(\hat{t}_{y,\pi}) + a^2\,\mathrm{Var}(\hat{t}_{x,\pi}) - 2a\mathrm{Cov}(\hat{t}_{x,\pi}, \hat{t}_{y,\pi}) \\
&= \frac{N(N-n)}{n}\left(S_y^2 + a^2 S_x^2 - 2aS_{xy}\right) \\
&= \frac{N(N-n)}{n}\left(S_y^2 + \frac{S_{xy}^2}{S_x^2} - 2\frac{S_{xy}^2}{S_x^2}\right) \\
&= \frac{N(N-n)}{n}\left(S_y^2 - \frac{S_{xy}^2}{S_x^2}\right) \\
&= \frac{N(N-n)}{n} S_y^2\left(1-\rho^2\right), \qquad \rho = \frac{S_{xy}}{S_x S_y}.
\end{aligned}$$

On estimera cette dernière par

$$\frac{N(N-n)}{n}\widehat{S_y^2}\left(1-\hat{\rho}^2\right), \qquad \hat{\rho} = \frac{\widehat{S_{xy}}}{\widehat{S_x}\widehat{S_y}}.$$

6.2.5 Comparaison

Le Tableau 6.1 donne l'expression des erreurs quadratiques moyennes pour les différents estimateurs par redressement introduit précédemment ainsi, qu'à titre de référence, celle du π–estimateur. Nous allons donc maintenant comparer ces estimateurs deux à deux afin d'établir une "règle de décision" afin de choisir le meilleur estimateur — au sens de l'erreur quadratique bien entendu.

— Estimateur par la différence vs. π–estimateur :

$$\begin{aligned}
\mathrm{EQM}(\hat{t}_{y,\pi}) - \mathrm{EQM}(\hat{t}_{y,D}) &= \frac{N(N-n)}{n} S_y^2 - \frac{N(N-n)}{n}(S_y^2 + S_x^2 - 2S_{xy}) \\
&= \frac{N(N-n)}{n}\left(2S_{xy} - S_x^2\right).
\end{aligned}$$

L'estimateur par la différence est donc meilleur lorsque

$$2S_{xy} - S_x^2 > 0 \Longleftrightarrow a > \frac{1}{2}.$$

— Estimateur par quotient vs. π–estimateur :

$$\begin{aligned}
\mathrm{EQM}(\hat{t}_{y,\pi}) - \mathrm{EQM}(\hat{t}_{y,Q}) &\approx \frac{N(N-n)}{n} S_y^2 - \frac{N(N-n)}{n}\left(S_y^2 + R^2 S_x^2 - 2RS_{xy}\right) \\
&= \frac{N(N-n)}{n}\left(2RS_{xy} - R^2 S_x^2\right).
\end{aligned}$$

L'estimateur par le quotient est donc (approximativement !!!) meilleur lorsque

$$2RS_{xy} - R^2S_x^2 > 0 \iff \begin{cases} a > \frac{R}{2}, & R > 0, \\ a < \frac{R}{2}, & R \leq 0. \end{cases}$$

— Estimateur par le quotient vs. estimateur par la différence :

$$\begin{aligned} \mathrm{EQM}(\hat{t}_{y,D}) - \mathrm{EQM}(\hat{t}_{y,Q}) &\approx \frac{N(N-n)}{n}\left(S_y^2 + S_x^2 - 2S_{xy}\right) - \\ &\quad \frac{N(N-n)}{n}\left(S_y^2 + R^2S_x^2 - 2RS_{xy}\right) \\ &= \frac{N(N-n)}{n}\left\{(1-R^2)S_x^2 + 2(1-R)S_{xy}\right\}. \end{aligned}$$

L'estimateur par le quotient est donc (approximativement !!!) meilleur lorsque

$$(1-R^2)S_x^2 + 2(1-R)S_{xy} > 0 \iff 2(1-R)a > 1 - R^2.$$

— Estimateur par régression vs. "les autres" : Cet estimateur est (approximativement !!!) le meilleurs de tous. En effet

$$\begin{aligned} \mathrm{EQM}(\hat{t}_{y,\pi}) - \mathrm{EQM}(\hat{t}_{y,R}) &\approx \frac{N(N-n)}{n}S_y^2\rho^2 \\ &= \rho^2\mathrm{EQM}\left(\hat{t}_{y,\pi}\right) \geq 0 \\ \mathrm{EQM}(\hat{t}_{y,D}) - \mathrm{EQM}(\hat{t}_{y,R}) &\approx \frac{N(N-n)}{n}\left(\rho^2S_y^2 + S_x^2 - 2S_{xy}\right) \\ &= \frac{N(N-n)}{n}\left(\frac{S_{xy}^2}{S_x^2} + S_x^2 - 2S_{xy}\right) \\ &= \frac{N(N-n)}{n}\left(\frac{S_{xy}^2}{S_x^2} - S_x\right)^2 \geq 0 \\ \mathrm{EQM}(\hat{t}_{y,Q}) - \mathrm{EQM}(\hat{t}_{y,R}) &\approx \frac{N(N-n)}{n}\left(\rho^2S_y^2 + R^2S_x^2 - 2RS_{xy}\right) \\ &= \frac{N(N-n)}{n}\left(\frac{S_{xy}^2}{S_x^2} + R^2S_x^2 - 2RS_{xy}\right) \\ &= \frac{N(N-n)}{n}\left(\frac{S_{xy}}{S_x} - RS_x\right)^2 \geq 0 \end{aligned}$$

Remarque. Il faut tout de même nuancer le fait que l'estimateur par régression soit toujours meilleur que les autres estimateurs, puisque ce n'est que du calcul approché. De plus l'estimateur par régression requiert l'estimation de la "pente" a ; et la variabilité de l'estimation de a n'a pas été prise en compte dans nos calculs.

Troisième partie

Exercices

Tests d'hypothèses

Tests paramétriques à un échantillon

Exercice 1 (Gaz nocif)**.** Dans l'atmosphère, le taux d'un gaz nocif, pour un volume donné, suit une loi normale d'espérance μ et de variance σ^2. On effectue n prélèvements.

1. On effectue aléatoirement $n = 20$ prélèvements. La moyenne et l'écart-type calculés sur cet échantillon sont respectivement : $m = 8,57$ et $s = 0,6$. On veut savoir si le seuil d'alerte fixé à $\mu_0 = 8,30$ est atteint.
   (a) Quel type de test faut-il effectuer ?
   (b) Que peut-on en conclure aux seuils $\alpha = 5\%, 1\%$?
2. Quelle serait la conclusion avec un échantillon de taille $n = 100$ et les mêmes valeurs observées ?
3. Une étude précédente a montré que $\sigma = 0,6$. Afin de savoir si le seuil d'alerte est atteint, on effectue à nouveau $n = 20$ prélèvements aléatoirement et la moyenne calculée sur cet échantillon est $m = 8,57$.
   (a) Quel type de test faut-il effectuer ?
   (b) Que peut-on en conclure aux mêmes seuils que précédemment
4. Quelles serait la conclusion avec un échantillon de taille n=100 et les mêmes valeurs observées ?

Exercice 2 (Moteurs d'automobiles)**.** Les pièces des moteurs d'automobiles de dernière génération sont usinées avec une très grande précision. L'écart-type des dimensions des pièces produites ne peut pas dépasser $\sigma_0 = 10\mu m$ (les dimensions sont représentées par un variable aléatoire normale de moyenne inconnue). On prélève au hasard, sur une unité de production, 25 pièces pour lesquelles on observe un écart-type $s = 13,5\mu m$.

1. Peut-on en conclure, au seuil de 1%, que l'écart-type théorique est supérieur à la valeur tolérée ?
2. Que peut-on dire de l'unité de production ?

Exercice 3 (Bébés prématurés)**.** Les recherches en psychologie du développement montrent que « normalement », 50% des bébés âgés de 12 mois savent marcher. Dans le cadre d'une étude sur les retards de développement des bébés prématurés, on veut savoir si ces derniers apprennent à marcher au même rythme que les autres. On observe donc un échantillon de 80 bébés prématurés âgés de 12 mois dont 35 marchent.

1. Faut-il réaliser un test unilatéral ou bilatéral ?
2. Quelle est la p-value de ce test ?
3. Comment les chercheurs peuvent-ils interpréter ce résultat ?

Tests paramétriques à deux échantillons

Exercice 4 (Elongation de pièce en acier)**.** Des mesures du pourcentage d'élongation ont été effectués sur 10 pièces d'acier (on suppose que ces mesures suivent une loi normale). 5 de ces pièces ont été traitées avec le produit A (Aluminium seulement) et les 5 autres avec le produit B (Aluminium et Calcium). Les mesures effectuées sont les suivantes :

Pourcentage d'élongation Traitement A	28	29	25	23	30
Pourcentage d'élongation Traitement B	34	27	30	26	33

1. Ces échantillons sont-ils appariés ?
2. Tester l'hypothèse d'égalité des variances au seuil de 5%.
3. Peut-on conclure, au seuil de 5%, que les deux traitements ont le même effet sur la moyenne ?

Exercice 5 (Usure d'un pneu)**.** Plusieurs méthodes existent pour déterminer l'usure d'un pneu. L'une de ces méthodes consiste à mesurer la profondeur de pénétration d'un instrument à des points fixes de la semelle du pneu. On pense que cette méthode produit de grandes variations lorsqu'elle est exécutée par des individus différents. Les données suivantes représentent les mesures, sur un lot de 12 pneus différents, effectuées par deux individus A et B.

A	121	121	126	130	127	131	127	124	125	119	126	123
B	120	129	128	136	117	112	138	124	119	136	135	134

Existe-t-il, en moyenne, une différence entre les mesures effectuées par A et B dans les deux cas suivants :

1. Les pneus ont été présentés à A et B dans le même ordre.
2. Les pneus ont été présentés à A et B dans un ordre quelconque.

Tests non paramétriques

Pour les exercices suivant on effectuera des tests au seuil de 5%.

Exercice 6 (M&M's)**.** On souhaite tester l'hypothèse suivant laquelle la répartition des couleurs de M&M's dans un paquet est bien en moyenne celle annoncée officiellement :

— Jaune : 15%

— Rouge : 12%

— Orange : 23%

— Bleu : 23%

— Marron : 12%

— Vert : 15%

Compter le nombre de M&M's de chaque couleur dans les paquets distribués et conclure sur l'observation de cet échantillon.

Exercice 7 (Péage)**.** Une distribution expérimentale du nombre de voiture qui arrivent à un péage durant une période de 10 secondes est donnée par le tableau suivant :

Nbre de voitures observées en 10sec	0	1	2	3	4	5	6	7	8
Effectifs observés	52	151	130	102	45	12	5	1	2

Peut-on affirmer que cette distribution est celle d'une loi de Poisson ?

Exercice 8 (Vaccin)**.** On veut savoir si l'efficacité d'un vaccin contre la grippe est indépendante du fait qu'on l'administre à des patients de moins de 55 ans ou à des patients strictement plus âgés. Sur 120 personnes de moins de 55 ans vaccinées, 38 ont contracté le virus, alors que sur les 180 personnes ayant strictement plus de 55 ans qui ont été vaccinées, 72 ont été malades.

Peut-on conclure à l'indépendance entre l'efficacité du vaccin et le fait que la personne vaccinée ait plus ou moins de 55 ans ?

Exercice 9 (Pneus)**.** On veut savoir si l'adhérence des types de pneus A et B sont équivalentes. Pour cela le fabriquant interroge 1000 clients auxquels il a vendu ces pneus. Les résultats de l'enquête sont présentés dans le tableau ci-dessous. Que peut-on en conclure ?

	Adhérence toujours bonne	Adhérence médiocre sur certains revêtements	Adhérence médiocre par temps de pluie
Pneus A	300	150	90
Pneus B	400	50	10

Exercice 10 (Contrôle qualité)**.** Le département de contrôle de qualité d'une entreprise a accumulé des données sur la qualité des produits semi-finis de trois fournisseurs A, B et C. Le tableau suivant représente la classification obtenue en fonction des défectuosités et des fournisseurs.

Nouvellement embauché par ce département, vous avez pour mission d'interpréter ces résultats afin d'en tirer le maximum d'information grâce à des méthodes statistiques dont vous détaillerez, avec rigueur, les développements.

	Fournisseur A	Fournisseur B	Fournisseur C
Défectuosité critique	35	25	40
Défectuosité majeure	250	150	300
Défectuosité mineure	325	275	600
Défectuosité aucune	800	1250	950

Sondages

Intervalles de confiance

Exercice 11 (Billes)**.** Une machine fabrique des billes dont le poids suit une loi normale. On prélève 10 billes et on obtient les poids (exprimés en grammes) :

19.6, 20, 20.2, 20.1, 20, 19.9, 20, 20.3, 20.1 et 19.8.

1. Donner un intervalle de confiance à 95% pour le poids moyen.
2. On suppose que l'écart-type est connu et égal à 0.2. Répondre à la question 1.

Exercice 12 (Partiel)**.** Voulant évaluer rapidement les résultats obtenus par ses deux cents étudiants lors d'un partiel, un professeur décide de corriger quelques copies tirées au hasard. Il admet par ailleurs que les notes de ses étudiants suivent une loi normale de variance 4.

1. Il corrige un échantillon de sept copies et trouve une moyenne de 11. Quel est l'intervalle de confiance à 95% de la moyenne des 200 copies ?
2. Combien de copies doit-il corriger s'il veut situer la moyenne générale de ses étudiants dans un intervalle de confiance d'amplitude 2, avec un risque de 5% ?
3. En trouvant une moyenne égale à 11, combien de copies devrait-il corriger pour pouvoir dire, avec un risque de 1%, que la moyenne de tous les étudiants est supérieure à 10 ?

Plan de sondage aléatoire simple

Exercice 13 (Culture et étudiants)**.** Dans un groupe de 80 étudiants on tire au hasard à probabilités égales et sans remise un échantillon de taille n. Nous prendrons n = 4 puis n = 40.

1. Nous observons dans l'échantillon la variable aléatoire « dépense hebdomadaire pour la culture ». Nous trouvons $\hat{\mu}_n(obs) = 12$ euros et $s_n = 6$ euros. Donner dans les deux cas une estimation de la dépense moyenne dans le groupe et la précision de cette estimation (i.e. donner un intervalle de confiance de niveau 95%).
2. Nous observons dans l'échantillon 75% de femmes. Donner dans les deux cas une estimation de la proportion de femmes dans le groupe et la précision de cette estimation.
3. Commenter les résultats obtenus.

Exercice 14 (Maladie professionnelle)**.** Nous nous intéressons à la proportion d'hommes P atteints par une maladie professionnelle dans une entreprise de 1500 travailleurs. Nous savons par ailleurs que trois travailleurs sur dix sont ordinairement touchés par cette maladie dans des entreprises du même type. Nous nous proposons de sélectionner un échantillon au ,moyen d'un sondage aléatoire simple.

1. Quelle taille d'échantillon faut-il sélectionner pour que la longueur totale d'un intervalle de confiance avec un niveau de confiance de 0,95 soit inférieure à 0,02 pour les plans simples avec et sans remise ?
2. Que faire si nous ne connaissons pas la proportion d'hommes habituellement touchés par la maladie pour le cas du plan sans remise ?

Plan de sondage stratifié

Exercice 15 (Éléphants)**.** Un directeur de cirque possède 100 éléphants classés en deux catégories : « mâles et femelles ». Le directeur veut estimer le poids total de son troupeau car il veut traverser un fleuve en bateau. L'année précédente, ce même directeur avait fait peser tous les éléphants de son troupeau et il avait obtenu les résultats suivants (l'unité est la tonne) :

	Effectifs	Moyennes	Variances corrigées
Mâles	60	6	4
Femelles	40	4	2,25

1. Calculer la variance dans la population de la variable « poids de l'éléphant » pour l'année précédente puis la variance corrigée.
2. Le directeur suppose désormais que les dispersions de poids n'évoluent pas sensiblement d'une année sur l'autre. Si le directeur procède à un PSAS à PESR de dix éléphants, quelle est la variance de l'estimateur du poids total ?
3. Si le directeur procède à un tirage stratifié avec allocation proportionnelle de dix éléphants, quels sont les effectifs de l'échantillon dans chacune des deux strates et quelle est la variance de l'estimateur du poids total ?
4. Si le directeur procède à un tirage stratifié optimal de dix éléphants, quels sont les effectifs de l'échantillon dans chacune des deux strates et quelle est la variance de l'estimateur du poids total ?

Exercice 16 (Revenus)**.** Sur les 7500 employés d'une entreprise, on souhaite connaître la proportion de ceux qui possèdent au moins un véhicule. On dispose de la valeur du revenu de chaque individu de la base et on décide de constituer 3 strates dans la population :

— revenu faible (strate 1)

— revenu moyen (strate 2)

— revenu élevé (strate 3)

On note :

— N_h : la taille de la strate h ;

— n_h : la taille de l'échantillon dans la strate h (tirage aléatoire simple sans remise) ;

— p_h : l'estimateur, dans la strate h, de la proportion du nombre d'individus possédant au moins un véhicule.

On obtient les résultats suivants :

	$h=1$	$h=2$	$h=3$
N_h	3500	2000	2000
n_h	500	300	200
p_h	0,23	0,45	0,50

1. Quel estimateur $\hat{P}$ de P proposez-vous ? Que peut-on dire de son biais ?
2. Donner un intervalle de confiance à 95% pour P.
3. En bon statisticien, le critère de stratification vous paraît-il justifié ?

Exercice 17 (Plan de sondage)**.** Soit la population $U = \{1,2,3,4,5,6,7,8,9\}$ et le plan de sondage suivant :

$$p(\{1,2\})) = \frac{1}{6}, \quad p(\{1,3\})) = \frac{1}{6}, \quad p(\{2,3\})) = \frac{1}{6},$$

$$p(\{4,5\}) = \frac{1}{12}, \quad p(\{4,6\}) = \frac{1}{12}, \quad p(\{5,6\}) = \frac{1}{12},$$
$$p(\{7,8\}) = \frac{1}{12}, \quad p(\{7,9\}) = \frac{1}{12}, \quad p(\{8,9\}) = \frac{1}{12}.$$

1. Donner les probabilités d'inclusion de premier degré.
2. Ce plan de sondage est-il simple, stratifié, en grappes, à deux degrés ou aucun de ces plans en particulier ?

Redressement à postériori

Exercice 18 (Parking)**.** Avant d'envisager la construction d'un nouveau parking dans une université, on fait un sondage pour estimer la proportion d'étudiants utilisant un véhicule personnel pour venir sur le campus. Comme on connaît l'origine géographique des étudiants (ville, reste du département, autre), on utilisera cette information pour faire éventuellement un redressement.

1. On a tiré un PSASPESR un échantillon de 150 étudiants. Les résultats sont donnés dans le tableau ci-dessous :

	Véhicule personnel	
Origine	Oui	Non
Ville	15	45
Département	25	25
Autre	30	10

   (a) Donner une estimation de la proportion d'étudiants utilisant un véhicule personnel.

   (b) Quelle est la variance de cette estimation ? (N=10000)

2. Sachant que sur ce campus la répartition de étudiants est la suivante :

Ville	5000
Département	3000
Autre	2000

   (a) Faire un redressement de l'estimation.

   (b) Donner une variance de cette nouvelle estimation.

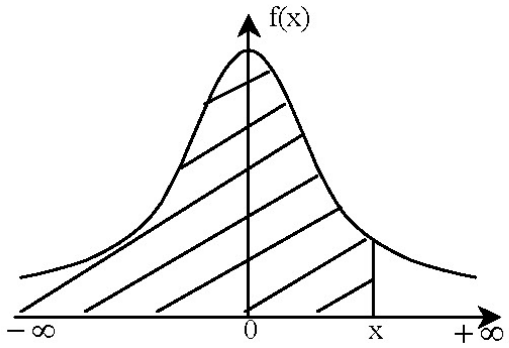
   (c) Ce redressement est-il justifié ?

Quatrième partie

Tables statistiques

Loi Normale centrée réduite

Probabilité de trouver une valeur inférieure à x.



$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du$$

X	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	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Table pour les grandes valeurs de x :

x	3	3,2	3,4	3,6	3,8	4	4,2	4,4	4,6	4,8
F(x)	0,99865003	0,99931280	0,99966302	0,99984085	0,99992763	0,99996831	0,99998665	0,99999458	0,99999789	0,99999921

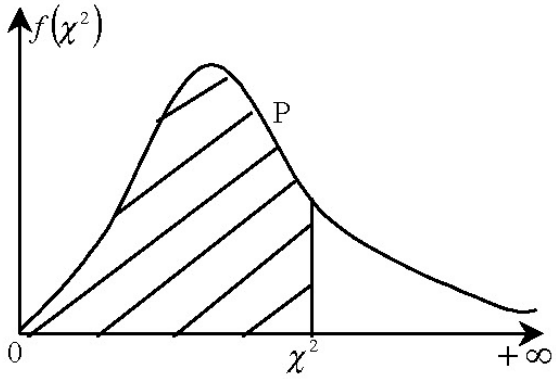
Table de Student

Table des valeurs critiques positives **bilatérales** de la loi de **Student** en fonction du seuil α et du degré de liberté dl. Cette table a été construite avec le logiciel SAS.

n \ α	0,350	0,300	0,250	0,200	0,150	0,100	0,050	0,020	0,010	0,001
1	1,63185	1,96261	2,41421	3,07768	4,16530	6,31375	12,7062	31,8205	63,6567	636,619
2	1,20963	1,38621	1,60357	1,88562	2,28193	2,91999	4,3027	6,9646	9,9248	31,599
3	1,10452	1,24978	1,42263	1,63774	1,92432	2,35336	3,1824	4,5407	5,8409	12,924
4	1,05730	1,18957	1,34440	1,53321	1,77819	2,13185	2,7764	3,7469	4,6041	8,610
5	1,03055	1,15577	1,30095	1,47588	1,69936	2,01505	2,5706	3,3649	4,0321	6,869
6	1,01335	1,13416	1,27335	1,43976	1,65017	1,94318	2,4469	3,1427	3,7074	5,959
7	1,00137	1,11916	1,25428	1,41492	1,61659	1,89458	2,3646	2,9980	3,4995	5,408
8	0,99254	1,10815	1,24032	1,39682	1,59222	1,85955	2,3060	2,8965	3,3554	5,041
9	0,98578	1,09972	1,22966	1,38303	1,57374	1,83311	2,2622	2,8214	3,2498	4,781
10	0,98043	1,09306	1,22126	1,37218	1,55924	1,81246	2,2281	2,7638	3,1693	4,587
11	0,97608	1,08767	1,21446	1,36343	1,54756	1,79588	2,2010	2,7181	3,1058	4,437
12	0,97249	1,08321	1,20885	1,35622	1,53796	1,78229	2,1788	2,6810	3,0545	4,318
13	0,96948	1,07947	1,20415	1,35017	1,52992	1,77093	2,1604	2,6503	3,0123	4,221
14	0,96690	1,07628	1,20014	1,34503	1,52310	1,76131	2,1448	2,6245	2,9768	4,140
15	0,96468	1,07353	1,19669	1,34061	1,51723	1,75305	2,1314	2,6025	2,9467	4,073
16	0,96275	1,07114	1,19369	1,33676	1,51213	1,74588	2,1199	2,5835	2,9208	4,015
17	0,96105	1,06903	1,19105	1,33338	1,50766	1,73961	2,1098	2,5669	2,8982	3,965
18	0,95954	1,06717	1,18871	1,33039	1,50371	1,73406	2,1009	2,5524	2,8784	3,922
19	0,95819	1,06551	1,18663	1,32773	1,50019	1,72913	2,0930	2,5395	2,8609	3,883
20	0,95699	1,06402	1,18476	1,32534	1,49704	1,72472	2,0860	2,5280	2,8453	3,850
21	0,95590	1,06267	1,18308	1,32319	1,49419	1,72074	2,0796	2,5176	2,8314	3,819
22	0,95491	1,06145	1,18155	1,32124	1,49162	1,71714	2,0739	2,5083	2,8188	3,792
23	0,95401	1,06034	1,18016	1,31946	1,48928	1,71387	2,0687	2,4999	2,8073	3,768
24	0,95318	1,05932	1,17888	1,31784	1,48714	1,71088	2,0639	2,4922	2,7969	3,745
25	0,95242	1,05838	1,17772	1,31635	1,48517	1,70814	2,0595	2,4851	2,7874	3,725
26	0,95173	1,05752	1,17664	1,31497	1,48336	1,70562	2,0555	2,4786	2,7787	3,707
27	0,95108	1,05673	1,17564	1,31370	1,48169	1,70329	2,0518	2,4727	2,7707	3,690
28	0,95048	1,05599	1,17472	1,31253	1,48014	1,70113	2,0484	2,4671	2,7633	3,674
29	0,94993	1,05530	1,17386	1,31143	1,47870	1,69913	2,0452	2,4620	2,7564	3,659
30	0,94941	1,05466	1,17306	1,31042	1,47736	1,69726	2,0423	2,4573	2,7500	3,646
31	0,94892	1,05406	1,17232	1,30946	1,47611	1,69552	2,0395	2,4528	2,7440	3,633
32	0,94847	1,05350	1,17162	1,30857	1,47494	1,69389	2,0369	2,4487	2,7385	3,622
33	0,94804	1,05298	1,17096	1,30774	1,47384	1,69236	2,0345	2,4448	2,7333	3,611
34	0,94764	1,05248	1,17035	1,30695	1,47281	1,69092	2,0322	2,4411	2,7284	3,601
35	0,94726	1,05202	1,16976	1,30621	1,47184	1,68957	2,0301	2,4377	2,7238	3,591
36	0,94691	1,05158	1,16922	1,30551	1,47092	1,68830	2,0281	2,4345	2,7195	3,582
37	0,94657	1,05117	1,16870	1,30485	1,47005	1,68709	2,0262	2,4314	2,7154	3,574
38	0,94625	1,05077	1,16821	1,30423	1,46923	1,68595	2,0244	2,4286	2,7116	3,566
39	0,94595	1,05040	1,16774	1,30364	1,46846	1,68488	2,0227	2,4258	2,7079	3,558
40	0,94566	1,05005	1,16730	1,30308	1,46772	1,68385	2,0211	2,4233	2,7045	3,551
50	0,94343	1,04729	1,16387	1,29871	1,46199	1,67591	2,0086	2,4033	2,6778	3,496
60	0,94194	1,04547	1,16160	1,29582	1,45820	1,67065	2,0003	2,3901	2,6603	3,460
70	0,94088	1,04417	1,15998	1,29376	1,45550	1,66691	1,9944	2,3808	2,6479	3,435
80	0,94009	1,04320	1,15876	1,29222	1,45349	1,66412	1,9901	2,3739	2,6387	3,416
90	0,93948	1,04244	1,15782	1,29103	1,45192	1,66196	1,9867	2,3685	2,6316	3,402
100	0,93899	1,04184	1,15707	1,29007	1,45067	1,66023	1,9840	2,3642	2,6259	3,390

Loi du χ^2

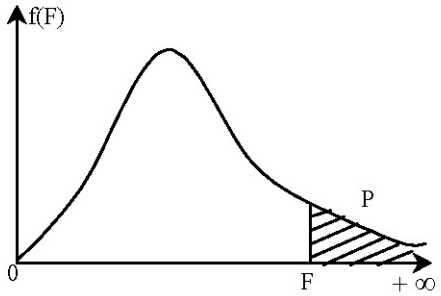
Valeur de χ^2 ayant la probabilité P d'être dépassée.



ddl/P	0,5%	1,0%	2,5%	5,0%	10,0%	50,0%	90,0%	95,0%	97,5%	99,0%	99,5%
1	0,000	0,000	0,001	0,004	0,016	0,455	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	1,386	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	2,366	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	3,357	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	4,351	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	5,348	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	6,346	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	7,344	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	8,343	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	9,342	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	10,341	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	11,340	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,041	12,340	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	13,339	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	14,339	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	15,338	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	16,338	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	17,338	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	18,338	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	19,337	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	20,337	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	21,337	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	22,337	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	23,337	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	24,337	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	25,336	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	26,336	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	27,336	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	28,336	39,087	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	20,599	29,336	40,256	43,773	46,979	50,892	53,672
31	14,458	15,655	17,539	19,281	21,434	30,336	41,422	44,985	48,232	52,191	55,002
32	15,134	16,362	18,291	20,072	22,271	31,336	42,585	46,194	49,480	53,486	56,328
33	15,815	17,073	19,047	20,867	23,110	32,336	43,745	47,400	50,725	54,775	57,648
34	16,501	17,789	19,806	21,664	23,952	33,336	44,903	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	24,797	34,336	46,059	49,802	53,203	57,342	60,275

Lorsque $\nu > 30$ on peut admettre que la quantité $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ suit une loi normale centrée réduite.

Loi de Fisher-Snedecor



Valeurs de F ayant 5% de chances d'être dépassées.

ν_2 \ ν_1	1	2	3	4	5	6	8	10	12	18	24	30	50	60	120
1	161,446	199,499	215,707	224,583	230,160	233,988	238,884	241,882	243,905	247,324	249,052	250,096	251,774	252,196	253,254
2	18,513	19,000	19,164	19,247	19,296	19,329	19,371	19,396	19,412	19,440	19,454	19,463	19,476	19,479	19,487
3	10,128	9,552	9,277	9,117	9,013	8,941	8,845	8,785	8,745	8,675	8,638	8,617	8,581	8,572	8,549
4	7,709	6,944	6,591	6,388	6,256	6,163	6,041	5,964	5,912	5,821	5,774	5,746	5,699	5,688	5,658
5	6,608	5,786	5,409	5,192	5,050	4,950	4,818	4,735	4,678	4,579	4,527	4,496	4,444	4,431	4,398
6	5,987	5,143	4,757	4,534	4,387	4,284	4,147	4,060	4,000	3,896	3,841	3,808	3,754	3,740	3,705
7	5,591	4,737	4,347	4,120	3,972	3,866	3,726	3,637	3,575	3,467	3,410	3,376	3,319	3,304	3,267
8	5,318	4,459	4,066	3,838	3,688	3,581	3,438	3,347	3,284	3,173	3,115	3,079	3,020	3,005	2,967
9	5,117	4,256	3,863	3,633	3,482	3,374	3,230	3,137	3,073	2,960	2,900	2,864	2,803	2,787	2,748
10	4,965	4,103	3,708	3,478	3,326	3,217	3,072	2,978	2,913	2,798	2,737	2,700	2,637	2,621	2,580
11	4,844	3,982	3,587	3,357	3,204	3,095	2,948	2,854	2,788	2,671	2,609	2,570	2,507	2,490	2,448
12	4,747	3,885	3,490	3,259	3,106	2,996	2,849	2,753	2,687	2,568	2,505	2,466	2,401	2,384	2,341
13	4,667	3,806	3,411	3,179	3,025	2,915	2,767	2,671	2,604	2,484	2,420	2,380	2,314	2,297	2,252
14	4,600	3,739	3,344	3,112	2,958	2,848	2,699	2,602	2,534	2,413	2,349	2,308	2,241	2,223	2,178
15	4,543	3,682	3,287	3,056	2,901	2,790	2,641	2,544	2,475	2,353	2,288	2,247	2,178	2,160	2,114
16	4,494	3,634	3,239	3,007	2,852	2,741	2,591	2,494	2,425	2,302	2,235	2,194	2,124	2,106	2,059
17	4,451	3,592	3,197	2,965	2,810	2,699	2,548	2,450	2,381	2,257	2,190	2,148	2,077	2,058	2,011
18	4,414	3,555	3,160	2,928	2,773	2,661	2,510	2,412	2,342	2,217	2,150	2,107	2,035	2,017	1,968
19	4,381	3,522	3,127	2,895	2,740	2,628	2,477	2,378	2,308	2,182	2,114	2,071	1,999	1,980	1,930
20	4,351	3,493	3,098	2,866	2,711	2,599	2,447	2,348	2,278	2,151	2,082	2,039	1,966	1,946	1,896
21	4,325	3,467	3,072	2,840	2,685	2,573	2,420	2,321	2,250	2,123	2,054	2,010	1,936	1,916	1,866
22	4,301	3,443	3,049	2,817	2,661	2,549	2,397	2,297	2,226	2,098	2,028	1,984	1,909	1,889	1,838
23	4,279	3,422	3,028	2,796	2,640	2,528	2,375	2,275	2,204	2,075	2,005	1,961	1,885	1,865	1,813
24	4,260	3,403	3,009	2,776	2,621	2,508	2,355	2,255	2,183	2,054	1,984	1,939	1,863	1,842	1,790
25	4,242	3,385	2,991	2,759	2,603	2,490	2,337	2,236	2,165	2,035	1,964	1,919	1,842	1,822	1,768
26	4,225	3,369	2,975	2,743	2,587	2,474	2,321	2,220	2,148	2,018	1,946	1,901	1,823	1,803	1,749
27	4,210	3,354	2,960	2,728	2,572	2,459	2,305	2,204	2,132	2,002	1,930	1,884	1,806	1,785	1,731
28	4,196	3,340	2,947	2,714	2,558	2,445	2,291	2,190	2,118	1,987	1,915	1,869	1,790	1,769	1,714
29	4,183	3,328	2,934	2,701	2,545	2,432	2,278	2,177	2,104	1,973	1,901	1,854	1,775	1,754	1,698
30	4,171	3,316	2,922	2,690	2,534	2,421	2,266	2,165	2,092	1,960	1,887	1,841	1,761	1,740	1,683
31	4,160	3,305	2,911	2,679	2,523	2,409	2,255	2,153	2,080	1,948	1,875	1,828	1,748	1,726	1,670
32	4,149	3,295	2,901	2,668	2,512	2,399	2,244	2,142	2,070	1,937	1,864	1,817	1,736	1,714	1,657
33	4,139	3,285	2,892	2,659	2,503	2,389	2,235	2,133	2,060	1,926	1,853	1,806	1,724	1,702	1,645
34	4,130	3,276	2,883	2,650	2,494	2,380	2,225	2,123	2,050	1,917	1,843	1,795	1,713	1,691	1,633
35	4,121	3,267	2,874	2,641	2,485	2,372	2,217	2,114	2,041	1,907	1,833	1,786	1,703	1,681	1,623
40	4,085	3,232	2,839	2,606	2,449	2,336	2,180	2,077	2,003	1,868	1,793	1,744	1,660	1,637	1,577
50	4,034	3,183	2,790	2,557	2,400	2,286	2,130	2,026	1,952	1,814	1,737	1,687	1,599	1,576	1,511
80	3,960	3,111	2,719	2,486	2,329	2,214	2,056	1,951	1,875	1,734	1,654	1,602	1,508	1,482	1,411
100	3,936	3,087	2,696	2,463	2,305	2,191	2,032	1,927	1,850	1,708	1,627	1,573	1,477	1,450	1,376
120	3,920	3,072	2,680	2,447	2,290	2,175	2,016	1,910	1,834	1,690	1,608	1,554	1,457	1,429	1,352

Valeurs de F ayant 2,5% de chances d'être dépassées.

ν_2 \ ν_1	1	2	3	4	5	6	8	10	12	18	24	30	50	60	120
1	647,793	799,482	864,151	899,599	921,835	937,114	956,643	968,634	976,725	990,345	997,272	1001,405	1008,098	1009,787	1014,036
2	38,506	39,000	39,166	39,248	39,298	39,331	39,373	39,398	39,415	39,442	39,457	39,465	39,478	39,481	39,489
3	17,443	16,044	15,439	15,101	14,885	14,735	14,540	14,419	14,337	14,196	14,124	14,081	14,010	13,992	13,947
4	12,218	10,649	9,979	9,604	9,364	9,197	8,980	8,844	8,751	8,592	8,511	8,461	8,381	8,360	8,309
5	10,007	8,434	7,764	7,388	7,146	6,978	6,757	6,619	6,525	6,362	6,278	6,227	6,144	6,123	6,069
6	8,813	7,260	6,599	6,227	5,988	5,820	5,600	5,461	5,366	5,202	5,117	5,065	4,980	4,959	4,904
7	8,073	6,542	5,890	5,523	5,285	5,119	4,899	4,761	4,666	4,501	4,415	4,362	4,276	4,254	4,199
8	7,571	6,059	5,416	5,053	4,817	4,652	4,433	4,295	4,200	4,034	3,947	3,894	3,807	3,784	3,728
9	7,209	5,715	5,078	4,718	4,484	4,320	4,102	3,964	3,868	3,701	3,614	3,560	3,472	3,449	3,392
10	6,937	5,456	4,826	4,468	4,236	4,072	3,855	3,717	3,621	3,453	3,365	3,311	3,221	3,198	3,140
11	6,724	5,256	4,630	4,275	4,044	3,881	3,664	3,526	3,430	3,261	3,173	3,118	3,027	3,004	2,944
12	6,554	5,096	4,474	4,121	3,891	3,728	3,512	3,374	3,277	3,108	3,019	2,963	2,871	2,848	2,787
13	6,414	4,965	4,347	3,996	3,767	3,604	3,388	3,250	3,153	2,983	2,893	2,837	2,744	2,720	2,659
14	6,298	4,857	4,242	3,892	3,663	3,501	3,285	3,147	3,050	2,879	2,789	2,732	2,638	2,614	2,552
15	6,200	4,765	4,153	3,804	3,576	3,415	3,199	3,060	2,963	2,792	2,701	2,644	2,549	2,524	2,461
16	6,115	4,687	4,077	3,729	3,502	3,341	3,125	2,986	2,889	2,717	2,625	2,568	2,472	2,447	2,383
17	6,042	4,619	4,011	3,665	3,438	3,277	3,061	2,922	2,825	2,652	2,560	2,502	2,405	2,380	2,315
18	5,978	4,560	3,954	3,608	3,382	3,221	3,005	2,866	2,769	2,596	2,503	2,445	2,347	2,321	2,256
19	5,922	4,508	3,903	3,559	3,333	3,172	2,956	2,817	2,720	2,546	2,452	2,394	2,295	2,270	2,203
20	5,871	4,461	3,859	3,515	3,289	3,128	2,913	2,774	2,676	2,501	2,408	2,349	2,249	2,223	2,156
21	5,827	4,420	3,819	3,475	3,250	3,090	2,874	2,735	2,637	2,462	2,368	2,308	2,208	2,182	2,114
22	5,786	4,383	3,783	3,440	3,215	3,055	2,839	2,700	2,602	2,426	2,332	2,272	2,171	2,145	2,076
23	5,750	4,349	3,750	3,408	3,183	3,023	2,808	2,668	2,570	2,394	2,299	2,239	2,137	2,111	2,041
24	5,717	4,319	3,721	3,379	3,155	2,995	2,779	2,640	2,541	2,365	2,269	2,209	2,107	2,080	2,010
25	5,686	4,291	3,694	3,353	3,129	2,969	2,753	2,613	2,515	2,338	2,242	2,182	2,079	2,052	1,981
26	5,659	4,265	3,670	3,329	3,105	2,945	2,729	2,590	2,491	2,314	2,217	2,157	2,053	2,026	1,954
27	5,633	4,242	3,647	3,307	3,083	2,923	2,707	2,568	2,469	2,291	2,195	2,133	2,029	2,002	1,930
28	5,610	4,221	3,626	3,286	3,063	2,903	2,687	2,547	2,448	2,270	2,174	2,112	2,007	1,980	1,907
29	5,588	4,201	3,607	3,267	3,044	2,884	2,669	2,529	2,430	2,251	2,154	2,092	1,987	1,959	1,886
30	5,568	4,182	3,589	3,250	3,026	2,867	2,651	2,511	2,412	2,233	2,136	2,074	1,968	1,940	1,866
31	5,549	4,165	3,573	3,234	3,010	2,851	2,635	2,495	2,396	2,217	2,119	2,057	1,950	1,922	1,848
32	5,531	4,149	3,557	3,218	2,995	2,836	2,620	2,480	2,381	2,201	2,103	2,041	1,934	1,905	1,831
33	5,515	4,134	3,543	3,204	2,981	2,822	2,606	2,466	2,366	2,187	2,088	2,026	1,918	1,890	1,815
34	5,499	4,120	3,529	3,191	2,968	2,808	2,593	2,453	2,353	2,173	2,075	2,012	1,904	1,875	1,799
35	5,485	4,106	3,517	3,179	2,956	2,796	2,581	2,440	2,341	2,160	2,062	1,999	1,890	1,861	1,785
40	5,424	4,051	3,463	3,126	2,904	2,744	2,529	2,388	2,288	2,107	2,007	1,943	1,832	1,803	1,724
50	5,340	3,975	3,390	3,054	2,833	2,674	2,458	2,317	2,216	2,033	1,931	1,866	1,752	1,721	1,639
80	5,218	3,864	3,284	2,950	2,730	2,571	2,355	2,213	2,111	1,925	1,820	1,752	1,632	1,599	1,508
100	5,179	3,828	3,250	2,917	2,696	2,537	2,321	2,179	2,077	1,890	1,784	1,715	1,592	1,558	1,463
120	5,152	3,805	3,227	2,894	2,674	2,515	2,299	2,157	2,055	1,866	1,760	1,690	1,565	1,530	1,433

Valeurs de F ayant 1% de chances d'être dépassées.

ν_2 \ ν_1	1	2	3	4	5	6	8	10	12	18	24	30	50	60	120
1	4052,185	4999,340	5403,534	5624,257	5763,955	5858,950	5980,954	6055,925	6106,682	6191,432	6234,273	6260,350	6302,260	6312,970	6339,513
2	98,502	99,000	99,164	99,251	99,302	99,331	99,375	99,397	99,419	99,444	99,455	99,466	99,477	99,484	99,491
3	34,116	30,816	29,457	28,710	28,237	27,911	27,489	27,228	27,052	26,751	26,597	26,504	26,354	26,316	26,221
4	21,198	18,000	16,694	15,977	15,522	15,207	14,799	14,546	14,374	14,079	13,929	13,838	13,690	13,652	13,558
5	16,258	13,274	12,060	11,392	10,967	10,672	10,289	10,051	9,888	9,609	9,466	9,379	9,238	9,202	9,112
6	13,745	10,925	9,780	9,148	8,746	8,466	8,102	7,874	7,718	7,451	7,313	7,229	7,091	7,057	6,969
7	12,246	9,547	8,451	7,847	7,460	7,191	6,840	6,620	6,469	6,209	6,074	5,992	5,858	5,824	5,737
8	11,259	8,649	7,591	7,006	6,632	6,371	6,029	5,814	5,667	5,412	5,279	5,198	5,065	5,032	4,946
9	10,562	8,022	6,992	6,422	6,057	5,802	5,467	5,257	5,111	4,860	4,729	4,649	4,517	4,483	4,398
10	10,044	7,559	6,552	5,994	5,636	5,386	5,057	4,849	4,706	4,457	4,327	4,247	4,115	4,082	3,996
11	9,646	7,206	6,217	5,668	5,316	5,069	4,744	4,539	4,397	4,150	4,021	3,941	3,810	3,776	3,690
12	9,330	6,927	5,953	5,412	5,064	4,821	4,499	4,296	4,155	3,910	3,780	3,701	3,569	3,535	3,449
13	9,074	6,701	5,739	5,205	4,862	4,620	4,302	4,100	3,960	3,716	3,587	3,507	3,375	3,341	3,255
14	8,862	6,515	5,564	5,035	4,695	4,456	4,140	3,939	3,800	3,556	3,427	3,348	3,215	3,181	3,094
15	8,683	6,359	5,417	4,893	4,556	4,318	4,004	3,805	3,666	3,423	3,294	3,214	3,081	3,047	2,959
16	8,531	6,226	5,292	4,773	4,437	4,202	3,890	3,691	3,553	3,310	3,181	3,101	2,967	2,933	2,845
17	8,400	6,112	5,185	4,669	4,336	4,101	3,791	3,593	3,455	3,212	3,083	3,003	2,869	2,835	2,746
18	8,285	6,013	5,092	4,579	4,248	4,015	3,705	3,508	3,371	3,128	2,999	2,919	2,784	2,749	2,660
19	8,185	5,926	5,010	4,500	4,171	3,939	3,631	3,434	3,297	3,054	2,925	2,844	2,709	2,674	2,584
20	8,096	5,849	4,938	4,431	4,103	3,871	3,564	3,368	3,231	2,989	2,859	2,778	2,643	2,608	2,517
21	8,017	5,780	4,874	4,369	4,042	3,812	3,506	3,310	3,173	2,931	2,801	2,720	2,584	2,548	2,457
22	7,945	5,719	4,817	4,313	3,988	3,758	3,453	3,258	3,121	2,879	2,749	2,667	2,531	2,495	2,403
23	7,881	5,664	4,765	4,264	3,939	3,710	3,406	3,211	3,074	2,832	2,702	2,620	2,483	2,447	2,354
24	7,823	5,614	4,718	4,218	3,895	3,667	3,363	3,168	3,032	2,789	2,659	2,577	2,440	2,403	2,310
25	7,770	5,568	4,675	4,177	3,855	3,627	3,324	3,129	2,993	2,751	2,620	2,538	2,400	2,364	2,270
26	7,721	5,526	4,637	4,140	3,818	3,591	3,288	3,094	2,958	2,715	2,585	2,503	2,364	2,327	2,233
27	7,677	5,488	4,601	4,106	3,785	3,558	3,256	3,062	2,926	2,683	2,552	2,470	2,330	2,294	2,198
28	7,636	5,453	4,568	4,074	3,754	3,528	3,226	3,032	2,896	2,653	2,522	2,440	2,300	2,263	2,167
29	7,598	5,420	4,538	4,045	3,725	3,499	3,198	3,005	2,868	2,626	2,495	2,412	2,271	2,234	2,138
30	7,562	5,390	4,510	4,018	3,699	3,473	3,173	2,979	2,843	2,600	2,469	2,386	2,245	2,208	2,111
31	7,530	5,362	4,484	3,993	3,675	3,449	3,149	2,955	2,820	2,577	2,445	2,362	2,221	2,183	2,086
32	7,499	5,336	4,459	3,969	3,652	3,427	3,127	2,934	2,798	2,555	2,423	2,340	2,198	2,160	2,062
33	7,471	5,312	4,437	3,948	3,630	3,406	3,106	2,913	2,777	2,534	2,402	2,319	2,176	2,139	2,040
34	7,444	5,289	4,416	3,927	3,611	3,386	3,087	2,894	2,758	2,515	2,383	2,299	2,156	2,118	2,019
35	7,419	5,268	4,396	3,908	3,592	3,368	3,069	2,876	2,740	2,497	2,364	2,281	2,137	2,099	2,000
40	7,314	5,178	4,313	3,828	3,514	3,291	2,993	2,801	2,665	2,421	2,288	2,203	2,058	2,019	1,917
50	7,171	5,057	4,199	3,720	3,408	3,186	2,890	2,698	2,563	2,318	2,183	2,098	1,949	1,909	1,803
80	6,963	4,881	4,036	3,563	3,255	3,036	2,742	2,551	2,415	2,169	2,032	1,944	1,788	1,746	1,630
100	6,895	4,824	3,984	3,513	3,206	2,988	2,694	2,503	2,368	2,120	1,983	1,893	1,735	1,692	1,572
120	6,851	4,787	3,949	3,480	3,174	2,956	2,663	2,472	2,336	2,089	1,950	1,860	1,700	1,656	1,533

Valeurs de F ayant 0,5% de chances d'être dépassées.

ν_2 \ ν_1	1	2	3	4	5	6	8	10	12	18	24	30	50	60	120
1	16212,463	19997,358	21614,134	22500,753	23055,822	23439,527	23923,814	24221,838	24426,728	24765,730	24937,093	25041,401	25212,765	25253,743	25358,051
2	198,503	199,012	199,158	199,245	199,303	199,332	199,376	199,390	199,419	199,449	199,449	199,478	199,478	199,478	199,492
3	55,552	49,800	47,468	46,195	45,391	44,838	44,125	43,685	43,387	42,881	42,623	42,466	42,211	42,150	41,990
4	31,332	26,284	24,260	23,154	22,456	21,975	21,352	20,967	20,705	20,258	20,030	19,892	19,667	19,611	19,469
5	22,785	18,314	16,530	15,556	14,939	14,513	13,961	13,618	13,385	12,985	12,780	12,656	12,454	12,402	12,274
6	18,635	14,544	12,917	12,028	11,464	11,073	10,566	10,250	10,034	9,664	9,474	9,358	9,170	9,122	9,001
7	16,235	12,404	10,883	10,050	9,522	9,155	8,678	8,380	8,176	7,826	7,645	7,534	7,354	7,309	7,193
8	14,688	11,043	9,597	8,805	8,302	7,952	7,496	7,211	7,015	6,678	6,503	6,396	6,222	6,177	6,065
9	13,614	10,107	8,717	7,956	7,471	7,134	6,693	6,417	6,227	5,899	5,729	5,625	5,454	5,410	5,300
10	12,827	9,427	8,081	7,343	6,872	6,545	6,116	5,847	5,661	5,340	5,173	5,071	4,902	4,859	4,750
11	12,226	8,912	7,600	6,881	6,422	6,102	5,682	5,418	5,236	4,921	4,756	4,654	4,488	4,445	4,337
12	11,754	8,510	7,226	6,521	6,071	5,757	5,345	5,085	4,906	4,595	4,431	4,331	4,165	4,123	4,015
13	11,374	8,186	6,926	6,233	5,791	5,482	5,076	4,820	4,643	4,334	4,173	4,073	3,908	3,866	3,758
14	11,060	7,922	6,680	5,998	5,562	5,257	4,857	4,603	4,428	4,122	3,961	3,862	3,697	3,655	3,547
15	10,798	7,701	6,476	5,803	5,372	5,071	4,674	4,424	4,250	3,946	3,786	3,687	3,523	3,480	3,372
16	10,576	7,514	6,303	5,638	5,212	4,913	4,521	4,272	4,099	3,797	3,638	3,539	3,375	3,332	3,224
17	10,384	7,354	6,156	5,497	5,075	4,779	4,389	4,142	3,971	3,670	3,511	3,412	3,248	3,206	3,097
18	10,218	7,215	6,028	5,375	4,956	4,663	4,276	4,030	3,860	3,560	3,402	3,303	3,139	3,096	2,987
19	10,073	7,093	5,916	5,268	4,853	4,561	4,177	3,933	3,763	3,464	3,306	3,208	3,043	3,000	2,891
20	9,944	6,987	5,818	5,174	4,762	4,472	4,090	3,847	3,678	3,380	3,222	3,123	2,959	2,916	2,806
21	9,829	6,891	5,730	5,091	4,681	4,393	4,013	3,771	3,602	3,305	3,147	3,049	2,884	2,841	2,730
22	9,727	6,806	5,652	5,017	4,609	4,322	3,944	3,703	3,535	3,239	3,081	2,982	2,817	2,774	2,663
23	9,635	6,730	5,582	4,950	4,544	4,259	3,882	3,642	3,474	3,179	3,021	2,922	2,756	2,713	2,602
24	9,551	6,661	5,519	4,890	4,486	4,202	3,826	3,587	3,420	3,125	2,967	2,868	2,702	2,658	2,546
25	9,475	6,598	5,462	4,835	4,433	4,150	3,776	3,537	3,370	3,075	2,918	2,819	2,652	2,609	2,496
26	9,406	6,541	5,409	4,785	4,384	4,103	3,730	3,492	3,325	3,031	2,873	2,774	2,607	2,563	2,450
27	9,342	6,489	5,361	4,740	4,340	4,059	3,687	3,450	3,284	2,990	2,832	2,733	2,565	2,522	2,408
28	9,284	6,440	5,317	4,698	4,300	4,020	3,649	3,412	3,246	2,952	2,794	2,695	2,527	2,483	2,369
29	9,230	6,396	5,276	4,659	4,262	3,983	3,613	3,376	3,211	2,917	2,759	2,660	2,492	2,448	2,333
30	9,180	6,355	5,239	4,623	4,228	3,949	3,580	3,344	3,179	2,885	2,727	2,628	2,459	2,415	2,300
31	9,133	6,316	5,204	4,590	4,195	3,918	3,549	3,314	3,149	2,855	2,697	2,598	2,429	2,385	2,269
32	9,090	6,281	5,172	4,559	4,166	3,889	3,521	3,286	3,121	2,828	2,670	2,570	2,401	2,356	2,240
33	9,049	6,248	5,141	4,531	4,138	3,861	3,495	3,260	3,095	2,802	2,644	2,544	2,374	2,330	2,213
34	9,012	6,217	5,113	4,504	4,112	3,836	3,470	3,235	3,071	2,778	2,620	2,520	2,350	2,305	2,188
35	8,976	6,188	5,086	4,479	4,088	3,812	3,447	3,212	3,048	2,755	2,597	2,497	2,327	2,282	2,164
40	8,828	6,066	4,976	4,374	3,986	3,713	3,350	3,117	2,953	2,661	2,502	2,401	2,230	2,184	2,064
50	8,626	5,902	4,826	4,232	3,849	3,579	3,219	2,988	2,825	2,533	2,373	2,272	2,097	2,050	1,925
80	8,335	5,665	4,611	4,028	3,652	3,387	3,032	2,803	2,641	2,349	2,188	2,084	1,903	1,854	1,720
100	8,241	5,589	4,542	3,963	3,589	3,325	2,972	2,744	2,583	2,290	2,128	2,024	1,840	1,790	1,652
120	8,179	5,539	4,497	3,921	3,548	3,285	2,933	2,705	2,544	2,251	2,089	1,984	1,798	1,747	1,606

Pour les grands échantillons, $\dfrac{s_1 - s_2}{s\sqrt{\dfrac{1}{2n_1} + \dfrac{1}{2n_2}}} \to \mathrm{N}(0,1)$ avec $s = \sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$.

Table de Shapiro - Francia

Tables des valeurs critiques du test de Normalité de Shapiro - Francia, pour trois seuils α différents, en fonction de la taille n de l'échantillon.

n \ α	0,10	0,05	0,01	n \ α	0,10	0,05	0,01	n \ α	0,10	0,05	0,01
5	0,9045	0,8866	0,8336	61	0,9838	0,9810	0,9718	117	0,9909	0,9887	0,9840
6	0,9119	0,8942	0,8485	62	0,9839	0,9811	0,9722	118	0,9910	0,9888	0,9842
7	0,9193	0,9018	0,8488	63	0,9842	0,9812	0,9722	119	0,9911	0,9888	0,9843
8	0,9248	0,9023	0,8676	64	0,9843	0,9813	0,9728	120	0,9913	0,9890	0,9844
9	0,9302	0,9113	0,8715	65	0,9844	0,9813	0,9732	125	0,9915	0,9894	0,9853
10	0,9351	0,9186	0,8724	66	0,9849	0,9817	0,9734	130	0,9918	0,9897	0,9859
11	0,9371	0,9224	0,8889	67	0,9852	0,9820	0,9736	135	0,9920	0,9900	0,9863
12	0,9409	0,9261	0,8917	68	0,9852	0,9821	0,9739	140	0,9922	0,9904	0,9865
13	0,9505	0,9288	0,8919	69	0,9854	0,9821	0,9744	145	0,9924	0,9906	0,9871
14	0,9514	0,9377	0,9063	70	0,9855	0,9822	0,9755	150	0,9925	0,9908	0,9875
15	0,9523	0,9398	0,9075	71	0,9856	0,9827	0,9757	155	0,9927	0,9911	0,9880
16	0,9532	0,9400	0,9141	72	0,9857	0,9828	0,9759	160	0,9931	0,9919	0,9887
17	0,9558	0,9431	0,9143	73	0,9859	0,9830	0,9762	165	0,9932	0,9920	0,9890
18	0,9561	0,9439	0,9161	74	0,9862	0,9832	0,9764	170	0,9934	0,9921	0,9892
19	0,9590	0,9493	0,9235	75	0,9864	0,9834	0,9765	175	0,9935	0,9922	0,9896
20	0,9595	0,9500	0,9283	76	0,9866	0,9834	0,9771	180	0,9936	0,9923	0,9899
21	0,9617	0,9542	0,9291	77	0,9867	0,9837	0,9774	185	0,9937	0,9925	0,9902
22	0,9626	0,9544	0,9311	78	0,9868	0,9839	0,9776	190	0,9939	0,9926	0,9903
23	0,9655	0,9554	0,9329	79	0,9869	0,9841	0,9779	195	0,9940	0,9929	0,9904
24	0,9662	0,9557	0,9388	80	0,9871	0,9841	0,9780	200	0,9941	0,9931	0,9908
25	0,9669	0,9572	0,9432	81	0,9872	0,9845	0,9782	210	0,9945	0,9934	0,9910
26	0,9675	0,9576	0,9437	82	0,9873	0,9847	0,9785	220	0,9946	0,9935	0,9913
27	0,9688	0,9599	0,9459	83	0,9874	0,9848	0,9786	230	0,9948	0,9937	0,9917
28	0,9701	0,9637	0,9466	84	0,9876	0,9848	0,9787	240	0,9951	0,9940	0,9920
29	0,9712	0,9641	0,9490	85	0,9877	0,9849	0,9788	250	0,9953	0,9944	0,9923
30	0,9723	0,9651	0,9506	86	0,9877	0,9851	0,9790	260	0,9955	0,9946	0,9924
31	0,9734	0,9654	0,9524	87	0,9878	0,9853	0,9791	270	0,9956	0,9947	0,9928
32	0,9745	0,9669	0,9538	88	0,9880	0,9854	0,9792	280	0,9957	0,9949	0,9930
33	0,9756	0,9678	0,9546	89	0,9880	0,9855	0,9793	290	0,9959	0,9950	0,9932
34	0,9758	0,9686	0,9548	90	0,9882	0,9857	0,9794	300	0,9959	0,9951	0,9934
35	0,9760	0,9691	0,9551	91	0,9883	0,9859	0,9796	310	0,9960	0,9952	0,9935
36	0,9762	0,9695	0,9551	92	0,9887	0,9859	0,9800	320	0,9962	0,9955	0,9937
37	0,9763	0,9698	0,9556	93	0,9888	0,9861	0,9802	330	0,9964	0,9956	0,9940
38	0,9765	0,9709	0,9560	94	0,9888	0,9862	0,9805	340	0,9964	0,9957	0,9941
39	0,9770	0,9712	0,9567	95	0,9889	0,9862	0,9807	350	0,9965	0,9958	0,9943
40	0,9775	0,9724	0,9576	96	0,9889	0,9863	0,9811	360	0,9965	0,9958	0,9943
41	0,9779	0,9726	0,9604	97	0,9890	0,9867	0,9812	370	0,9967	0,9960	0,9945
42	0,9780	0,9730	0,9610	98	0,9893	0,9868	0,9814	380	0,9968	0,9961	0,9946
43	0,9782	0,9732	0,9611	99	0,9894	0,9872	0,9818	390	0,9969	0,9961	0,9947
44	0,9785	0,9734	0,9631	100	0,9896	0,9873	0,9820	400	0,9969	0,9962	0,9951
45	0,9789	0,9743	0,9636	101	0,9896	0,9873	0,9821	420	0,9970	0,9963	0,9953
46	0,9793	0,9758	0,9643	102	0,9897	0,9873	0,9821	440	0,9972	0,9967	0,9955
47	0,9801	0,9764	0,9646	103	0,9897	0,9875	0,9822	460	0,9973	0,9967	0,9958
48	0,9806	0,9767	0,9653	104	0,9897	0,9876	0,9822	480	0,9974	0,9969	0,9959
49	0,9807	0,9768	0,9655	105	0,9898	0,9876	0,9823	500	0,9975	0,9971	0,9959
50	0,9810	0,9769	0,9667	106	0,9899	0,9877	0,9823	550	0,9978	0,9974	0,9961
51	0,9815	0,9772	0,9669	107	0,9899	0,9878	0,9825	600	0,9979	0,9975	0,9963
52	0,9818	0,9773	0,9673	108	0,9899	0,9878	0,9826	650	0,9979	0,9978	0,9967
53	0,9822	0,9777	0,9685	109	0,9900	0,9878	0,9828	700	0,9981	0,9979	0,9972
54	0,9824	0,9780	0,9694	110	0,9901	0,9879	0,9832	750	0,9983	0,9979	0,9973
55	0,9825	0,9788	0,9695	111	0,9902	0,9880	0,9834	800	0,9984	0,9980	0,9974
56	0,9826	0,9792	0,9697	112	0,9902	0,9882	0,9835	850	0,9985	0,9981	0,9975
57	0,9828	0,9796	0,9704	113	0,9906	0,9885	0,9837	900	0,9986	0,9983	0,9976
58	0,9830	0,9797	0,9705	114	0,9906	0,9886	0,9837	950	0,9987	0,9984	0,9978
59	0,9833	0,9800	0,9707	115	0,9907	0,9887	0,9839	999	0,9988	0,9985	0,9979
60	0,9836	0,9801	0,9710	116	0,9909	0,9887	0,9840	1000	0,9988	0,9985	0,9979

Fonction de répartition de la loi de Poisson

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197
3	1	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810
4	1	1	1	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963
5	1	1	1	1	1	1	0.9999	0.9998	0.9997	0.9994
6	1	1	1	1	1	1	1	1	1	0.9999
7	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.6990	0.6626	0.6268	0.5918	0.5578	0.5249	0.4932	0.4628	0.4337	0.4060
2	0.9004	0.8795	0.8571	0.8335	0.8088	0.7834	0.7572	0.7306	0.7037	0.6767
3	0.9743	0.9662	0.9569	0.9463	0.9344	0.9212	0.9068	0.8913	0.8747	0.8571
4	0.9946	0.9923	0.9893	0.9857	0.9814	0.9763	0.9704	0.9636	0.9559	0.9473
5	0.9990	0.9985	0.9978	0.9968	0.9955	0.9940	0.9920	0.9896	0.9868	0.9834
6	0.9999	0.9997	0.9996	0.9994	0.9991	0.9987	0.9981	0.9974	0.9966	0.9955
7	1	1	0.9999	0.9999	0.9998	0.9997	0.9996	0.9994	0.9992	0.9989
8	1	1	1	1	1	1	0.9999	0.9999	0.9998	0.9998
9	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	0.1225	0.1108	0.1003	0.0907	0.0821	0.0743	0.0672	0.0608	0.0550	0.0498
1	0.3796	0.3546	0.3309	0.3084	0.2873	0.2674	0.2487	0.2311	0.2146	0.1991
2	0.6496	0.6227	0.5960	0.5697	0.5438	0.5184	0.4936	0.4695	0.4460	0.4232
3	0.8386	0.8194	0.7993	0.7787	0.7576	0.7360	0.7141	0.6919	0.6696	0.6472
4	0.9379	0.9275	0.9162	0.9041	0.8912	0.8774	0.8629	0.8477	0.8318	0.8153
5	0.9796	0.9751	0.9700	0.9643	0.9580	0.9510	0.9433	0.9349	0.9258	0.9161
6	0.9941	0.9925	0.9906	0.9884	0.9858	0.9828	0.9794	0.9756	0.9713	0.9665
7	0.9985	0.9980	0.9974	0.9967	0.9958	0.9947	0.9934	0.9919	0.9901	0.9881
8	0.9997	0.9995	0.9994	0.9991	0.9989	0.9985	0.9981	0.9976	0.9969	0.9962
9	0.9999	0.9999	0.9999	0.9998	0.9997	0.9996	0.9995	0.9993	0.9991	0.9989
10	1	1	1	1	0.9999	0.9999	0.9999	0.9998	0.9998	0.9997
11	1	1	1	1	1	1	1	1	0.9999	0.9999
12	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	0.0450	0.0408	0.0369	0.0334	0.0302	0.0273	0.0247	0.0224	0.0202	0.0183
1	0.1847	0.1712	0.1586	0.1468	0.1359	0.1257	0.1162	0.1074	0.0992	0.0916
2	0.4012	0.3799	0.3594	0.3397	0.3208	0.3027	0.2854	0.2689	0.2531	0.2381
3	0.6248	0.6025	0.5803	0.5584	0.5366	0.5152	0.4942	0.4735	0.4532	0.4335
4	0.7982	0.7806	0.7626	0.7442	0.7254	0.7064	0.6872	0.6678	0.6484	0.6288
5	0.9057	0.8946	0.8829	0.8705	0.8576	0.8441	0.8301	0.8156	0.8006	0.7851
6	0.9612	0.9554	0.9490	0.9421	0.9347	0.9267	0.9182	0.9091	0.8995	0.8893
7	0.9858	0.9832	0.9802	0.9769	0.9733	0.9692	0.9648	0.9599	0.9546	0.9489
8	0.9953	0.9943	0.9931	0.9917	0.9901	0.9883	0.9863	0.9840	0.9815	0.9786
9	0.9986	0.9982	0.9978	0.9973	0.9967	0.9960	0.9952	0.9942	0.9931	0.9919
10	0.9996	0.9995	0.9994	0.9992	0.9990	0.9987	0.9984	0.9981	0.9977	0.9972
11	0.9999	0.9999	0.9998	0.9998	0.9997	0.9996	0.9995	0.9994	0.9993	0.9991
12	1	1	1	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	0.9997
13	1	1	1	1	1	1	1	1	0.9999	0.9999
14	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	0.0166	0.0150	0.0136	0.0123	0.0111	0.0101	0.0091	0.0082	0.0074	0.0067
1	0.0845	0.0780	0.0719	0.0663	0.0611	0.0563	0.0518	0.0477	0.0439	0.0404
2	0.2238	0.2102	0.1974	0.1851	0.1736	0.1626	0.1523	0.1425	0.1333	0.1247
3	0.4142	0.3954	0.3772	0.3594	0.3423	0.3257	0.3097	0.2942	0.2793	0.2650
4	0.6093	0.5898	0.5704	0.5512	0.5321	0.5132	0.4946	0.4763	0.4582	0.4405
5	0.7693	0.7531	0.7367	0.7199	0.7029	0.6858	0.6684	0.6510	0.6335	0.6160
6	0.8786	0.8675	0.8558	0.8436	0.8311	0.8180	0.8046	0.7908	0.7767	0.7622
7	0.9427	0.9361	0.9290	0.9214	0.9134	0.9049	0.8960	0.8867	0.8769	0.8666
8	0.9755	0.9721	0.9683	0.9642	0.9597	0.9549	0.9497	0.9442	0.9382	0.9319
9	0.9905	0.9889	0.9871	0.9851	0.9829	0.9805	0.9778	0.9749	0.9717	0.9682
10	0.9966	0.9959	0.9952	0.9943	0.9933	0.9922	0.9910	0.9896	0.9880	0.9863
11	0.9989	0.9986	0.9983	0.9980	0.9976	0.9971	0.9966	0.9960	0.9953	0.9945
12	0.9997	0.9996	0.9995	0.9993	0.9992	0.9990	0.9988	0.9986	0.9983	0.9980
13	0.9999	0.9999	0.9998	0.9998	0.9997	0.9997	0.9996	0.9995	0.9994	0.9993
14	1	1	1	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998
15	1	1	1	1	1	1	1	1	0.9999	0.9999
16	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	0.0061	0.0055	0.0050	0.0045	0.0041	0.0037	0.0033	0.0030	0.0027	0.0025
1	0.0372	0.0342	0.0314	0.0289	0.0266	0.0244	0.0224	0.0206	0.0189	0.0174
2	0.1165	0.1088	0.1016	0.0948	0.0884	0.0824	0.0768	0.0715	0.0666	0.0620
3	0.2513	0.2381	0.2254	0.2133	0.2017	0.1906	0.1800	0.1700	0.1604	0.1512
4	0.4231	0.4061	0.3895	0.3733	0.3575	0.3422	0.3272	0.3127	0.2987	0.2851
5	0.5984	0.5809	0.5635	0.5461	0.5289	0.5119	0.4950	0.4783	0.4619	0.4457
6	0.7474	0.7324	0.7171	0.7017	0.6860	0.6703	0.6544	0.6384	0.6224	0.6063
7	0.8560	0.8449	0.8335	0.8217	0.8095	0.7970	0.7841	0.7710	0.7576	0.7440
8	0.9252	0.9181	0.9106	0.9027	0.8944	0.8857	0.8766	0.8672	0.8574	0.8472
9	0.9644	0.9603	0.9559	0.9512	0.9462	0.9409	0.9352	0.9292	0.9228	0.9161
10	0.9844	0.9823	0.9800	0.9775	0.9747	0.9718	0.9686	0.9651	0.9614	0.9574
11	0.9937	0.9927	0.9916	0.9904	0.9890	0.9875	0.9859	0.9841	0.9821	0.9799
12	0.9976	0.9972	0.9967	0.9962	0.9955	0.9949	0.9941	0.9932	0.9922	0.9912
13	0.9992	0.9990	0.9988	0.9986	0.9983	0.9980	0.9977	0.9973	0.9969	0.9964
14	0.9997	0.9997	0.9996	0.9995	0.9994	0.9993	0.9991	0.9990	0.9988	0.9986
15	0.9999	0.9999	0.9999	0.9998	0.9998	0.9998	0.9997	0.9996	0.9996	0.9995
16	1	1	1	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998
17	1	1	1	1	1	1	1	1	1	0.9999
18	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	0.0022	0.0020	0.0018	0.0017	0.0015	0.0014	0.0012	0.0011	0.0010	0.0009
1	0.0159	0.0146	0.0134	0.0123	0.0113	0.0103	0.0095	0.0087	0.0080	0.0073
2	0.0577	0.0536	0.0498	0.0463	0.0430	0.0400	0.0371	0.0344	0.0320	0.0296
3	0.1425	0.1342	0.1264	0.1189	0.1118	0.1052	0.0988	0.0928	0.0871	0.0818
4	0.2719	0.2592	0.2469	0.2351	0.2237	0.2127	0.2022	0.1920	0.1823	0.1730
5	0.4298	0.4141	0.3988	0.3837	0.3690	0.3547	0.3406	0.3270	0.3137	0.3007
6	0.5902	0.5742	0.5582	0.5423	0.5265	0.5108	0.4953	0.4799	0.4647	0.4497
7	0.7301	0.7160	0.7017	0.6873	0.6728	0.6581	0.6433	0.6285	0.6136	0.5987
8	0.8367	0.8259	0.8148	0.8033	0.7916	0.7796	0.7673	0.7548	0.7420	0.7291
9	0.9090	0.9016	0.8939	0.8858	0.8774	0.8686	0.8596	0.8502	0.8405	0.8305
10	0.9531	0.9486	0.9437	0.9386	0.9332	0.9274	0.9214	0.9151	0.9084	0.9015
11	0.9776	0.9750	0.9723	0.9693	0.9661	0.9627	0.9591	0.9552	0.9510	0.9467
12	0.9900	0.9887	0.9873	0.9857	0.9840	0.9821	0.9801	0.9779	0.9755	0.9730
13	0.9958	0.9952	0.9945	0.9937	0.9929	0.9920	0.9909	0.9898	0.9885	0.9872
14	0.9984	0.9981	0.9978	0.9974	0.9970	0.9966	0.9961	0.9956	0.9950	0.9943
15	0.9994	0.9993	0.9992	0.9990	0.9988	0.9986	0.9984	0.9982	0.9979	0.9976
16	0.9998	0.9997	0.9997	0.9996	0.9996	0.9995	0.9994	0.9993	0.9992	0.9990
17	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	0.9998	0.9997	0.9997	0.9996
18	1	1	1	1	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
19	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	0.0008	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004	0.0003
1	0.0067	0.0061	0.0056	0.0051	0.0047	0.0043	0.0039	0.0036	0.0033	0.0030
2	0.0275	0.0255	0.0236	0.0219	0.0203	0.0188	0.0174	0.0161	0.0149	0.0138
3	0.0767	0.0719	0.0674	0.0632	0.0591	0.0554	0.0518	0.0485	0.0453	0.0424
4	0.1641	0.1555	0.1473	0.1395	0.1321	0.1249	0.1181	0.1117	0.1055	0.0996
5	0.2881	0.2759	0.2640	0.2526	0.2414	0.2307	0.2203	0.2103	0.2006	0.1912
6	0.4349	0.4204	0.4060	0.3920	0.3782	0.3646	0.3514	0.3384	0.3257	0.3134
7	0.5838	0.5689	0.5541	0.5393	0.5246	0.5100	0.4956	0.4812	0.4670	0.4530
8	0.7160	0.7027	0.6892	0.6757	0.6620	0.6482	0.6343	0.6204	0.6065	0.5925
9	0.8202	0.8096	0.7988	0.7877	0.7764	0.7649	0.7531	0.7411	0.7290	0.7166
10	0.8942	0.8867	0.8788	0.8707	0.8622	0.8535	0.8445	0.8352	0.8257	0.8159
11	0.9420	0.9371	0.9319	0.9265	0.9208	0.9148	0.9085	0.9020	0.8952	0.8881
12	0.9703	0.9673	0.9642	0.9609	0.9573	0.9536	0.9496	0.9454	0.9409	0.9362
13	0.9857	0.9841	0.9824	0.9805	0.9784	0.9762	0.9739	0.9714	0.9687	0.9658
14	0.9935	0.9927	0.9918	0.9908	0.9897	0.9886	0.9873	0.9859	0.9844	0.9827
15	0.9972	0.9969	0.9964	0.9959	0.9954	0.9948	0.9941	0.9934	0.9926	0.9918
16	0.9989	0.9987	0.9985	0.9983	0.9980	0.9978	0.9974	0.9971	0.9967	0.9963
17	0.9996	0.9995	0.9994	0.9993	0.9992	0.9991	0.9989	0.9988	0.9986	0.9984
18	0.9998	0.9998	0.9998	0.9997	0.9997	0.9996	0.9996	0.9995	0.9994	0.9993
19	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	0.9998	0.9997
20	1	1	1	1	1	1	0.9999	0.9999	0.9999	0.9999
21	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
1	0.0028	0.0025	0.0023	0.0021	0.0019	0.0018	0.0016	0.0015	0.0014	0.0012
2	0.0127	0.0118	0.0109	0.0100	0.0093	0.0086	0.0079	0.0073	0.0068	0.0062
3	0.0396	0.0370	0.0346	0.0323	0.0301	0.0281	0.0262	0.0244	0.0228	0.0212
4	0.0940	0.0887	0.0837	0.0789	0.0744	0.0701	0.0660	0.0621	0.0584	0.0550
5	0.1822	0.1736	0.1653	0.1573	0.1496	0.1422	0.1352	0.1284	0.1219	0.1157
6	0.3013	0.2896	0.2781	0.2670	0.2562	0.2457	0.2355	0.2256	0.2160	0.2068
7	0.4391	0.4254	0.4119	0.3987	0.3856	0.3728	0.3602	0.3478	0.3357	0.3239
8	0.5786	0.5647	0.5507	0.5369	0.5231	0.5094	0.4958	0.4823	0.4689	0.4557
9	0.7041	0.6915	0.6788	0.6659	0.6530	0.6400	0.6269	0.6137	0.6006	0.5874
10	0.8058	0.7955	0.7850	0.7743	0.7634	0.7522	0.7409	0.7294	0.7178	0.7060
11	0.8807	0.8731	0.8652	0.8571	0.8487	0.8400	0.8311	0.8220	0.8126	0.8030
12	0.9313	0.9261	0.9207	0.9150	0.9091	0.9029	0.8965	0.8898	0.8829	0.8758
13	0.9628	0.9595	0.9561	0.9524	0.9486	0.9445	0.9403	0.9358	0.9311	0.9261
14	0.9810	0.9791	0.9771	0.9749	0.9726	0.9701	0.9675	0.9647	0.9617	0.9585
15	0.9908	0.9898	0.9887	0.9875	0.9862	0.9848	0.9832	0.9816	0.9798	0.9780
16	0.9958	0.9953	0.9947	0.9941	0.9934	0.9926	0.9918	0.9909	0.9899	0.9889
17	0.9982	0.9979	0.9977	0.9973	0.9970	0.9966	0.9962	0.9957	0.9952	0.9947
18	0.9992	0.9991	0.9990	0.9989	0.9987	0.9985	0.9983	0.9981	0.9978	0.9976
19	0.9997	0.9997	0.9996	0.9995	0.9995	0.9994	0.9993	0.9992	0.9991	0.9989
20	0.9999	0.9999	0.9998	0.9998	0.9998	0.9998	0.9997	0.9997	0.9996	0.9996
21	1	1	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998
22	1	1	1	1	1	1	1	1	0.9999	0.9999
23	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10.0
0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
1	0.0011	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005
2	0.0058	0.0053	0.0049	0.0045	0.0042	0.0038	0.0035	0.0033	0.0030	0.0028
3	0.0198	0.0184	0.0172	0.0160	0.0149	0.0138	0.0129	0.0120	0.0111	0.0103
4	0.0517	0.0486	0.0456	0.0429	0.0403	0.0378	0.0355	0.0333	0.0312	0.0293
5	0.1098	0.1041	0.0986	0.0935	0.0885	0.0838	0.0793	0.0750	0.0710	0.0671
6	0.1978	0.1892	0.1808	0.1727	0.1649	0.1574	0.1502	0.1433	0.1366	0.1301
7	0.3123	0.3010	0.2900	0.2792	0.2687	0.2584	0.2485	0.2388	0.2294	0.2202
8	0.4426	0.4296	0.4168	0.4042	0.3918	0.3796	0.3676	0.3558	0.3442	0.3328
9	0.5742	0.5611	0.5479	0.5349	0.5218	0.5089	0.4960	0.4832	0.4705	0.4579
10	0.6941	0.6820	0.6699	0.6576	0.6453	0.6329	0.6205	0.6080	0.5955	0.5830
11	0.7932	0.7832	0.7730	0.7626	0.7520	0.7412	0.7303	0.7193	0.7081	0.6968
12	0.8684	0.8607	0.8529	0.8448	0.8364	0.8279	0.8191	0.8101	0.8009	0.7916
13	0.9210	0.9156	0.9100	0.9042	0.8981	0.8919	0.8853	0.8786	0.8716	0.8645
14	0.9552	0.9517	0.9480	0.9441	0.9400	0.9357	0.9312	0.9265	0.9216	0.9165
15	0.9760	0.9738	0.9715	0.9691	0.9665	0.9638	0.9609	0.9579	0.9546	0.9513
16	0.9878	0.9865	0.9852	0.9838	0.9823	0.9806	0.9789	0.9770	0.9751	0.9730
17	0.9941	0.9934	0.9927	0.9919	0.9911	0.9902	0.9892	0.9881	0.9870	0.9857
18	0.9973	0.9969	0.9966	0.9962	0.9957	0.9952	0.9947	0.9941	0.9935	0.9928
19	0.9988	0.9986	0.9985	0.9983	0.9980	0.9978	0.9975	0.9972	0.9969	0.9965
20	0.9995	0.9994	0.9993	0.9992	0.9991	0.9990	0.9989	0.9987	0.9986	0.9984
21	0.9998	0.9998	0.9997	0.9997	0.9996	0.9996	0.9995	0.9995	0.9994	0.9993
22	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	0.9998	0.9997	0.9997
23	1	1	1	1	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
24	1	1	1	1	1	1	1	1	1	1

$P(X \leq x)$ où $X \sim \mathcal{P}(\lambda)$

	λ									
x	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0049	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
4	0.0151	0.0076	0.0037	0.0018	0.0009	0.0004	0.0002	0.0001	0.0000	0.0000
5	0.0375	0.0203	0.0107	0.0055	0.0028	0.0014	0.0007	0.0003	0.0002	0.0001
6	0.0786	0.0458	0.0259	0.0142	0.0076	0.0040	0.0021	0.0010	0.0005	0.0003
7	0.1432	0.0895	0.0540	0.0316	0.0180	0.0100	0.0054	0.0029	0.0015	0.0008
8	0.2320	0.1550	0.0998	0.0621	0.0374	0.0220	0.0126	0.0071	0.0039	0.0021
9	0.3405	0.2424	0.1658	0.1094	0.0699	0.0433	0.0261	0.0154	0.0089	0.0050
10	0.4599	0.3472	0.2517	0.1757	0.1185	0.0774	0.0491	0.0304	0.0183	0.0108
11	0.5793	0.4616	0.3532	0.2600	0.1848	0.1270	0.0847	0.0549	0.0347	0.0214
12	0.6887	0.5760	0.4631	0.3585	0.2676	0.1931	0.1350	0.0917	0.0606	0.0390
13	0.7813	0.6815	0.5730	0.4644	0.3632	0.2745	0.2009	0.1426	0.0984	0.0661
14	0.8540	0.7720	0.6751	0.5704	0.4657	0.3675	0.2808	0.2081	0.1497	0.1049
15	0.9074	0.8444	0.7636	0.6694	0.5681	0.4667	0.3715	0.2867	0.2148	0.1565
16	0.9441	0.8987	0.8355	0.7559	0.6641	0.5660	0.4677	0.3751	0.2920	0.2211
17	0.9678	0.9370	0.8905	0.8272	0.7489	0.6593	0.5640	0.4686	0.3784	0.2970
18	0.9823	0.9626	0.9302	0.8826	0.8195	0.7423	0.6550	0.5622	0.4695	0.3814
19	0.9907	0.9787	0.9573	0.9235	0.8752	0.8122	0.7363	0.6509	0.5606	0.4703
20	0.9953	0.9884	0.9750	0.9521	0.9170	0.8682	0.8055	0.7307	0.6472	0.5591
21	0.9977	0.9939	0.9859	0.9712	0.9469	0.9108	0.8615	0.7991	0.7255	0.6437
22	0.9990	0.9970	0.9924	0.9833	0.9673	0.9418	0.9047	0.8551	0.7931	0.7206
23	0.9995	0.9985	0.9960	0.9907	0.9805	0.9633	0.9367	0.8989	0.8490	0.7875
24	0.9998	0.9993	0.9980	0.9950	0.9888	0.9777	0.9594	0.9317	0.8933	0.8432
25	0.9999	0.9997	0.9990	0.9974	0.9938	0.9869	0.9748	0.9554	0.9269	0.8878
26	1	0.9999	0.9995	0.9987	0.9967	0.9925	0.9848	0.9718	0.9514	0.9221
27	1	0.9999	0.9998	0.9994	0.9983	0.9959	0.9912	0.9827	0.9687	0.9475
28	1	1	0.9999	0.9997	0.9991	0.9978	0.9950	0.9897	0.9805	0.9657
29	1	1	1	0.9999	0.9996	0.9989	0.9973	0.9941	0.9882	0.9782

Bibliographie

[1] Emmanuel Monfrini. *Statistiques appliquées.* Ancien polycopié du cours MAT4103 de Télécom SudParis.

[2] Mathieu Ribatet. *Sondages et Enquêtes.* Polycopié dans le cadre du Master MIND de l'université de Montpellier 2.
http://mribatet.perso.math.cnrs.fr/docs/Sondages/cours.pdf

[3] Rivoirard et Stoltz (2009). *Statistique en action.* Cours et problèmes corrigés pour Master et Agrégation de mathématiques. Vuibert.

[4] Jean-Jacques Ruch (2012). *Statistique : Tests d'hypothèses.* Polycopié pour la préparation à l'Agrégation à Bordeaux 1.
https://www.math.u-bordeaux.fr/∼mchabano/Agreg/ProbaAgreg1213-COURS3-Stat2.pdf

[5] Yves Tillé (2019) *Théorie des Sondages : Échantillonnage et estimation en populations finies.* 2ème édition. Dunod.

[6] Wikipédia. Test statistique.
https://fr.wikipedia.org/wikiTest_statistique

INSTITUT
POLYTECHNIQUE
DE PARIS