Eléments de Statistique

Jean VAILLANT

Septembre 2010

Table des matières

Chapitre 1

Initiation à la théorie de l'échantillonnage

1.1 Notions de base en échantillonnage

L'étude de propriétés caractéristiques d'un ensemble, quand on ne dispose pas encore de données, nécessite d'examiner, d'observer des éléments de cet ensemble. La manière de recueillir ces données fait l'objet d'une théorie mathématique appelée *théorie des sondages* ou encore *théorie de l'échantillonnage*, en anglais *sampling theory*. Cette théorie concerne l'optimisation de la collecte des données selon divers critères et répond à certaines interrogations sur la façon de procéder à cette collecte en rapport avec l'information disponible et l'effort d'échantillonnage consenti.

1.1.1 Population, individu statistiques

La **population statistique** est l'ensemble sur lequel des méthodes et techniques de présentation, de description et d'inférence statistique sont appliquées. Il ne s'agit donc pas forcément d'une population au sens biologique du terme.

Les **individus statistiques** ou **unités statistiques** sont les éléments de la population statistique.

Les exemples sont innombrables :

(1) On désire étudier la préférence pour tel ou tel candidat dans une circonscription. La population statistique est l'ensemble des électeurs

de la circonscription.
(2) On s'intéresse à l'action d'un parasite sur les pontes de la pyrale de la canne à sucre dans une région. La population statistique est l'ensemble des plantes des parcelles cultivées en canne à sucre de la région étudiée.
(3) On s'intéresse à la répartition d'une maladie sur les arbres d'une forêt. La population statistique est l'ensemble des arbres de cette forêt.
(4) On désire évaluer le budget mensuel moyen des étudiants d'une université. La population statistique est l'ensemble des étudiants de cette université.

1.1.2 Sondage, échantillonnage

On appelle **sondage** toute observation partielle d'une population statistique c'est-à-dire l'observation d'une partie de cette population. On cherche généralement à extrapoler les résultats observés à la totalité de la population. Une **unité de sondage** (ou **unité d'échantillonnage**) est un regroupement d'unités statistiques.

Une **méthode de sondages** (ou **d'échantillonnage**) décrit la façon dont la population statistique sera observée partiellement à travers un de ses sous-ensembles appelé **échantillon**.

Plan de sondages, plan d'échantillonnage, procédure d'échantillonnage ont des définitions équivalentes à celles de méthode de sondages.

Il est important de ne pas confondre **sondage** et **sondage d'opinion**. Les sondages d'opinion vise à obtenir des informations sur l'état d'esprit d'une population humaine. Il s'agit donc d'une forme particulière de sondage : les individus statistiques sont des personnes interrogées à travers un questionnaire sur leur opinion. Parmi les exemples ci-dessus, seul [1] est un sondage d'opinion ([4] n'en est pas un, bien que l'on y interroge des personnes). Les sondages d'opinion sont très médiatisés, particulièrement pendant les périodes préélectorales.

La **théorie des sondages** est un ensemble d'outils statistiques permettant l'étude d'une population statistique à partir de l'examen d'un échantillon tiré de celle-ci (Tillé, 2001). On parle aussi, de façon équivalente, de la **théorie de l'échantillonnage**. Cette dernière expression est davantage utilisée en sciences agronomiques ou biologiques.

1.1.3 Population finie, taux de sondage

La **taille de la population statistique** est l'effectif de cette population c'est-à-dire le nombre d'individus statistiques dont elle est constituée. Une **population finie** est une

population dont la taille est finie. Une **population infinie** est une population dont la taille est infinie. Dans la pratique, on peut considérer une population finie comme étant infinie si elle est d'effectif très grand.

La **taille de l'échantillon** est l'effectif de cet échantillon c'est-à-dire le nombre d'individus statistiques observés dans la population statistique.

Le **taux de sondage** (ou **d'échantillonnage**), dans le cas de population finie, est le rapport

$$\frac{\text{taille d'échantillon}}{\text{taille de population}}.$$

Le **facteur correctif de population finie** est

$$1 - \frac{\text{taille d'échantillon}}{\text{taille de population}}.$$

Quand la population statistique est observée complètement, c'est-à-dire que l'échantillon est la population statistique toute entière, on parle d'**échantillonnage exhaustif** ou de **recensement**. Le taux de sondage est alors de 100%.

Pour des raisons de coûts financiers ou techniques, il est, dans bien des cas, impossible de faire un recensement. L'utilisation de sondages est alors incontournable.

1.1.4 Population cible, base de sondage

Dans certaines études, on souligne la notion de **population cible** puis celle de **base de sondage**. On a les définitions suivantes :

- La **population cible** est l'ensemble pour lequel on veut recueillir des informations et sur lequel doivent porter les conclusions de l'étude. Elle peut être distincte de la population statistique, en particulier quand ses éléments ne peuvent être tous répertoriés ou sont soumis à des contraintes liées à l'étude menée. Dans la table 1, les exemples 2, 4 et 5 correspondent à une situation où la population cible est différente de la population statistique.
- La **base de sondage** est déterminée, après avoir définie la population cible. Idéalement, c'est une liste de tous les individus de la population cible : liste électorale, liste des entreprises, liste d'étudiants, liste des arbres d'une parcelle sylvicole, liste des parcelles d'un domaine expérimental, etc... L'échantillon est alors extrait de cette liste à l'aide d'un algorithme de tirages des individus. Par contre, il arrive

	Population statistique	Unité statistique	Population cible	Caractère étudié	Paramètre à estimer
1	Les arbres d'une forêt	Arbre de la forêt	Ensemble des arbres de cette forêt	Présence-absence d'une maladie	Proportion d'arbres malades dans la forêt
2	Les parcelles obtenues par quadrillage d'un champ de tournesol	Parcelle	Ensemble des plantes de tournesol du champ	Nombre de plantes crispées (action du puceron du tournesol)	Nombre moyen de plantes crispées dans le champ
3	Les étudiants d'une université	Etudiant	Ensemble des étudiants de cette université	Budget annuel	Budget annuel moyen par étudiant
4	Les entreprises répertoriées en début d'année dans une région	Entreprise de cette région	Ensemble des entreprises de cette région	Taux d'endettement	Taux d'endettement moyen des entreprises de cette région
5	Les plants de canne à sucre d'une région	Plant de canne à sucre	Ensemble des pontes de pyrale dans la région	Nombre de pontes parasitées par un oophage	Nombre moyen de pontes parasitées dans la région (efficacité de l'oophage)

TABLE 1.1 – Exemples de population statistique, cible et caractère étudié.

fréquemment qu'une telle base de sondage ne soit pas accessible directement. La **base de sondage** est par définition une liste d'individus statistiques identifiés permettant d'avoir accès à la majorité des individus de la population cible. Tous les individus de la population cible ne sont donc pas forcément inclus dans cette base.

En résumé, une base de sondage est donc une liste d'individus de la population statistique à partir de laquelle (liste) on tire l'échantillon avec, pour chaque individu, divers renseignements utiles à la réalisation de l'étude par échantillonnage. Un exemple de telle base est le registre des exploitants de tel département archivé par la Chambre d'Agriculture de ce département.

Dans certains cas, on a plusieurs choix possibles pour la base de sondage. Ce choix dépendra des objectifs de l'étude, des données disponibles sur la base de sondage, de la qualité de la base de sondage et du budget de l'étude, comme le montrent les exemples suivants :

1) *Enquête auprès d'exploitants agricoles d'un département* : afin d'étudier l'utilisation d'intrants en agriculture en 2004 et analyser les risques

pour l'environnement, on a considéré le registre des exploitants de 2003 archivé à la chambre d'Agriculture de ce département. En tenant compte des départs et arrivées enregistrés en 2004 ainsi que du secteur informel, on estime que cette base de sondage permet d'atteindre environ 89% des exploitants de ce département.

2) *Enquête concernant la dengue auprès des ménages d'une commune* : Le comportement vis-à-vis de la dengue et de son vecteur, le moustique Aedes aegypti, veut être étudié afin de mettre en place une campagne efficace d'éradication de ce moustique. L'individu statistique est le logement. Il est assimilé au **ménage** qui est, par définition, l'ensemble des individus qui habitent le même logement. A la suite de chaque recensement, l'INSEE dispose d'une base de sondage comprenant tous les logements recensés. Cette base contient toutes les constructions achevées lors du recensement le plus récent. Elle est complétée par une base de sondage des logements neufs éditée par la Direction Départementale de l'Equipement. Ces bases de sondage représentent des fichiers énormes au niveau de la France. Pour une commune donnée, on peut extraire une base de sondage à partir de ces fichiers. Selon la commune concernée, le pourcentage des ménages couverts par une telle base de sondage varie à cause des éventuelles lenteurs de mise à jour des fichiers et des constructions illégales.

3) *Etude de la pression anthropique sur le crabe de terre en zone littorale* : Pour analyser l'impact anthropique sur la dynamique du crabe de terre, les différents terriers de la zone sont répertoriés. Ils constituent la base de sondage. La population cible est l'ensemble des crabes de terre.

L'existence d'une base adéquate de sélection des individus statistiques est donc un aspect important de la faisabilité de la plupart des plans d'échantillonnage.

1.1.5 Population fixe, superpopulation

Considérons une population notée $\mathcal{P}$ de taille N dont les individus sont répertoriés.

On peut donc affecter un numéro allant de 1 à N à chaque individu et assimiler notre population à l'ensemble des N premiers entiers naturels non nuls :

$$\mathcal{P} = \{1, \cdots, N\}.$$

Sur cette population statistique, on étudie un caractère $\mathcal{Y}$ prenant la valeur y_i sur l'individu i. **Avant l'exécution de l'échantillonnage, on a N valeurs inconnues $y_1, y_2, \cdots, y_N$.** Le fait d'échantillonner dans la population **permet d'accéder à certaines de ces valeurs**.

Le but de l'échantillonnage est souvent d'estimer une quantité $h(y_1, y_2, \cdots, y_N)$ dépendant donc par conséquence des $y_1, y_2, \cdots, y_N$. Cette quantité $h(y_1, y_2, \cdots, y_N)$ est appelé **fonction d'intérêt**. L'exemple le plus courant de fonction d'intérêt est **la moyenne de la population** :

$$\bar{y} = \frac{1}{N}(y_1 + y_2 + \cdots + y_N).$$

Remarque : la proportion est une forme particulière de moyenne : il s'agit de la moyenne d'un caractère ne pouvant prendre que les valeurs 0 ou 1 (**caractère binaire**).

On peut s'intéresser également à la fonction d'intérêt **variance de la population** :

$$V = \frac{1}{N}((y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_N - \bar{y})^2).$$

Il est important de noter que, dans certaines situations, des suppositions de nature probabiliste sont faites sur les valeurs inconnues $y_1, y_2, \cdots, y_N$. Par exemple, on peut supposer que y_i est la réalisation d'une variable aléatoire suivant la loi gaussienne. On parle alors de **modèle de superpopulation**, par opposition au cas où aucune supposition distributionnelle n'est faite sur les $y_1, y_2, \cdots, y_N$ auquel cas on parle de **population fixe**. On parle donc de population fixe lorsque les y_i sont considérées fixes, ou en d'autres termes, quand on ne fait aucune supposition probabiliste sur les valeurs possibles du caractère $\mathcal{Y}$ sur les différents individus de la population.

1.1.6 Plan stochastique, plan empirique

Il existe deux grandes catégories de plans d'échantillonnage :

- **les plans probabilistes**, dits aussi **plans stochastiques**. Ces plans se caractérisent par le fait que les individus statistiques devant faire partie de l'échantillon sont sélectionnés par tirages probabilistes. Chaque individu de la population statistique a une probabilité connue d'être inclus dans l'échantillon (cette probabilité est appelée **probabilité d'inclusion d'ordre un** de l'individu pour le plan d'échantillonnage considéré). Avec de tels plans, il est possible d'utiliser la théorie des probabilités : les observations sur l'échantillon sont des variables aléatoires. On peut utiliser des outils d'inférence statistique pour estimer des paramètres de la population et également évaluer les précisions d'estimation.
- **les plans non probabilistes**, dits aussi **plans empiriques** ou **plans par choix raisonné**. L'échantillon est construit par des procédés comportant une part d'arbitraire et ne permettant pas l'évaluation de la précision d'estimation.

Les plans probabilistes classiques sont les suivants : plan aléatoire simple, plan aléatoire systématique, plan aléatoire stratifié, plan aléatoire en groupes.

Les plans non probabilistes sont utilisés dans les études qualitatives où il n'est pas envisagé une extrapolation à la population statistique dans son entier. Quelques exemples :

- Plan par commodité : on choisit des individus statistiques qui sont d'accès facile
- Enquête boule de neige : on choisit quelques individus (au sein d'une population humaine) qui sont pertinents pour l'étude, et ensuite on leur demande de proposer d'autres individus pour l'enquête.
- Plan par quotas : on construit un échantillon qui respecte les proportions connues pour certaines catégories de la population.

1.1.7 Représentativité d'un échantillon

La définition d'**échantillon représentatif** diffère selon que le plan d'échantillonnage est probabiliste ou non probabiliste :

- un plan probabiliste fournit un échantillon représentatif dès lors que chaque individu de la population a une *probabilité connue et non nulle d'être inclus dans l'échantillon.*
- un plan non probabiliste fournit un échantillon représentatif si la *structure de l'échantillon pour certaines variables clés est similaire à celle de la population cible.* Par exemple, on peut vouloir construire un échantillon pour lequel les proportions de catégories d'individus soient similaires dans l'échantillon à celles de la population cible (c'est le principe de la méthode dite des quotas).

En population fixe, un échantillon n'est représentatif que de la population au sein de laquelle il a été sélectionné.

1.1.8 Précision statistique, distribution d'échantillonnage

Une **statistique** est une valeur calculée à partir d'observations effectuées sur les individus de l'échantillon. Comme l'objectif de l'échantillonnage est généralement d'**inférer** sur la population statistique (c'est-à-dire tirer des conclusions concernant cette population), le calcul de cette statistique correspond souvent à **l'estimation d'un paramètre** (c'est-à-dire l'évaluation numérique de ce paramètre), paramètre de la population sur laquelle est prélevé l'échantillon. Les deux exemples les plus courants sont indiqués dans la table 2.

L'ensemble des valeurs possibles d'une statistique, affectées de leur probabilité de réalisation s'appelle la **distribution d'échantillonnage** de cette statistique. Les figures 1 et 2 nous montrent deux exemples de distributions d'échantillonnage.

La distribution d'échantillonnage d'une statistique peut correspondre à une loi de probabilité usuelle. Ainsi, le nombre d'individus malades observé dans l'échantillon suit :

- une loi binomiale si les tirages sont effectués avec remise,
- une loi hypergéométrique si les tirages sont effectués sans remise,
- une loi approximativement gaussienne si les tirages sont suffisamment nombreux.

La **précision statistique** d'une méthode d'estimation d'un paramètre de la population est définie comme *une mesure de l'écart entre l'estimation obtenue à partir de l'échantillon et la vraie valeur du paramètre*. Cet écart est attribuable à deux types d'erreur :

- **l'erreur d'échantillonnage** : c'est l'erreur liée à l'aléa de tirage de l'échantillon car, à partir d'un échantillon, quand on calcule une statistique, on obtient une valeur parmi toutes les valeurs possibles de la distribution d'échantillonnage de cette statistique. L'erreur d'échantillonnage diminue généralement avec l'accroissement de la taille d'échantillon.
- **l'erreur d'observation** : elle est la conséquence d'erreur de mesure, de notation lors de la cueillette de l'information, mais aussi, dans le cas d'enquêtes auprès de personnes, des réponses erronées, des refus de réponse. Ce type d'erreur peut être minimisé par une formation approfondie des observateurs ou des enquêteurs et par le contrôle de la qualité du travail effectué aux différentes étapes du plan d'échantillonnage.

Un **estimateur d'une fonction d'intérêt** $\theta(y_1, y_2, \cdots, y_N)$ (ou plus simplement d'un paramètre θ) est une statistique T qui fournit une évaluation pertinente de θ.

L'**espérance mathématique** $E(T)$ d'une statistique T est une moyenne théorique qui est la somme des valeurs possibles de T pondérées par leurs probabilités de réalisation. On l'appelle aussi **valeur espérée** de T.

Statistique	**Paramètre**
Moyenne de l'échantillon	Moyenne de la population
Proportion dans l'échantillon	Proportion dans la population

TABLE 1.2 – Exemples de statistique associée à un paramètre

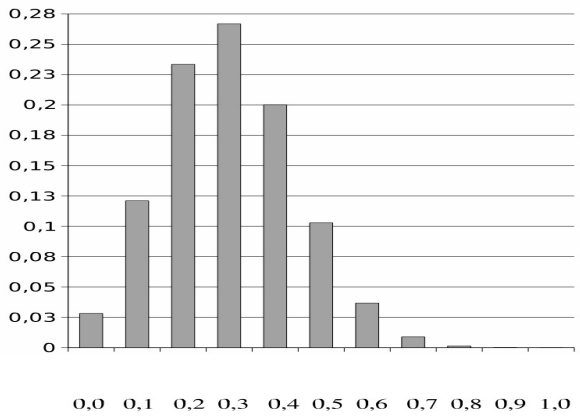

FIGURE 1.1 – Exemple de distribution d'échantillonnage de la proportion observée pour une taille d'échantillon $n = 10$ obtenu par **tirages avec remise** dans une population de taille 50. La proportion dans la population (inconnue et à estimer dans la pratique) est $p = 0,3$.

Le **biais** $B(T)$ de la statistique T pour le paramètre θ est l'écart entre la valeur espérée de T et θ :

$$B_\theta(T) = E(T) - \theta.$$

La **variance d'échantillonnage** $V(T)$ de la statistique T est par définition

$$V(T) = E((T - E(T))^2).$$

L'**écart quadratique moyen** $EQM(T)$ d'estimation de θ par T est défini de la façon suivante

$$EQM_\theta(T) = E((T - \theta)^2).$$

Le lien entre écart quadratique, biais et variance d'échantillonnage est :

$$EQM_\theta(T) = V(T) + (B_\theta(T))^2.$$

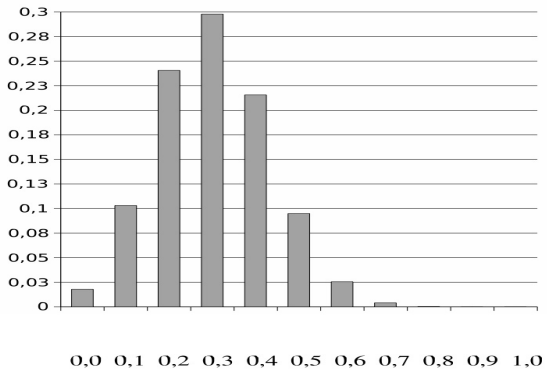

FIGURE 1.2 – Exemple de distribution d'échantillonnage de la proportion observée pour une taille d'échantillon $n = 10$ obtenu par **tirages sans remise** dans une population de taille 50. La proportion dans la population (inconnue et à estimer dans la pratique) est $p = 0,3$.

L'estimateur T est dit **sans biais** pour θ si $E(T) = \theta$, ce qui est équivalent à $B_\theta(T) = 0$. La figure 1.3 illustre les notions de distribution, de biais et de variabilié d'échantillonnage. Elle fait le parallèle suivant entre qualités d'un estimateur et d'un tireur sur cible. Un estimateur fournit, pour un échantillon donné, une évaluation numérique du paramètre considéré ; un tireur obtient, pour un tir effectué, un impact sur la cible alors qu'il cherche à atteindre le centre de la cible. Les estimations varient d'un échantillon à l'autre ; les points d'impact varient d'un tir à l'autre. On espère qu'avec un bon estimateur on a des valeurs estimées qui ne sont pas trop éloignées du paramètre ; et des points d'impact pas trop éloignés du centre de la cible pour un bon tireur.

Généralement, on recherche un estimateur qui soit de moindre écart quadratique moyen. La situation la plus intéressante, dans la pratique, est de disposer d'un estimateur sans biais et de moindre variance. La **convergence** est une autre propriété très recherchée :

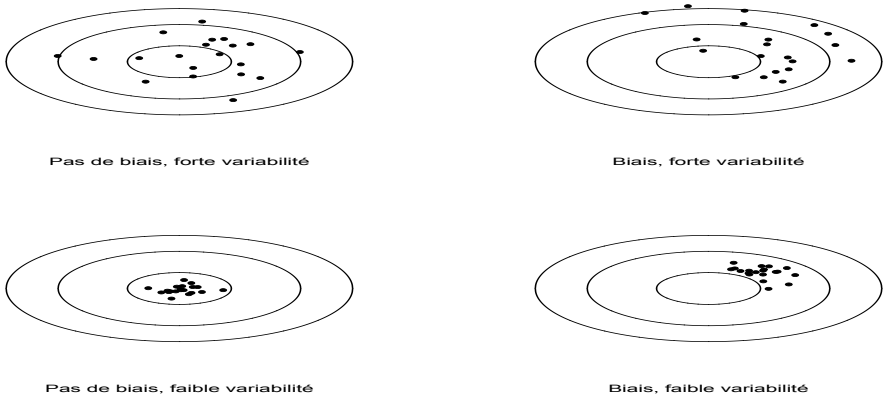

FIGURE 1.3 – Illustration du biais et de la variabilité d'un estimateur

elle signifie que plus on a des données, plus notre estimation se rapproche de la vraie valeur du paramètre inconnu.

Les trois méthodes d'estimation les plus répandues sont celles du maximum de vraisemblance, des moments et des moindres carrés.

Estimation par maximum de vraisemblance : La vraisemblance est une fonction du paramètre θ conditionnelle aux observations. Elle traduit la probabilité d'observer l'échantillon obtenu pour la valeur θ du paramètre.

Par exemple, on choisit n individus par tirages indépendants dans une population dont une proportion p est malade. Notons X le nombre d'individus malades dans l'échantillon. La vraisemblance est par conséquent la fonction :

$$p \longmapsto C_n^X p^X (1-p)^{n-X}$$

car X suit une loi binomiale de paramètres n et p en tant que nombre de succès au bout de n épreuves avec pour chaque épreuve une probabilité de succès égale à p.

La valeur qui maximise cette fonction n'est autre que X/n c'est-à-dire la proportion de malades observée dans l'échantillon.

Estimation des moments (ou M-estimation) : Soit k un entier supérieur à 1. Le moment d'ordre k d'une variable aléatoire X, noté $m_k(X)$, est l'espérance de X^k : $m_k(X) = E(X^k)$. La méthode des moments est utilisée quand on sait expliciter les moments d'une statistique en fonction des paramètres inconnus que l'on veut estimer. Pour un nombre r de paramètres à estimer, on utilise les moments d'ordre 1 à r pour construire un système

de r équations à r inconnues. Dans ce système, on introduit les moments dit empiriques puis on calcule les r solutions qui sont appelées estimateurs des moments.

Prenons l'exemple de comptages $x_1, x_2, \cdots x_n$ d'individus d'une espèce dans n unités expérimentales de même taille. Pour des individus à comportement indépendant mais avec des affinités communes, on admet que la loi de probabilité du nombre d'individus X dans une unité expérimentale suit une loi binomiale négative d'espérance μ et de paramètre d'agrégation γ. Pour estimer μ et γ par la méthode des moments, on utilise les expressions des moments d'ordre 1 et 2 de X :

$$E(X) = \mu \quad \text{et} \quad E(X^2) = \mu + (1 + \frac{1}{\gamma})\mu^2.$$

En remplaçant dans ces équations, les termes de gauche par les moments empiriques $\widehat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\widehat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, on obtient :

$$\widehat{m}_1 = \mu \quad \text{et} \quad \widehat{m}_2 = \mu + (1 + \frac{1}{\gamma})\mu^2.$$

La résolution en μ et γ de ce sytème d'équations nous donne les solutions :

$$\widehat{\mu} = \widehat{m}_1 \quad \text{et} \quad \widehat{\gamma} = \frac{\widehat{m}_1^2}{\widehat{m}_2 - \widehat{m}_1 - \widehat{m}_1^2}$$

qui sont donc les M-estimateurs de μ et de γ.

Estimation des moindres carrés : Le principe est de minimiser la somme des carrés d'écart entre valeurs observées et valeurs espérées sous un modèle.

Une application classique de cette méthode est l'estimation des paramètres du modèle de régression linéaire simple. Ce modèle associe une variable quantitative (dite *réponse*) à une variable explicative de la facçon suivante : $y_i = ax_i + b + \epsilon_i$, où y_i et x_i sont les valeurs prises par la réponse et la variable explicative sur l'individu statistique i. La pente a et l'interception b sont appelées paramètre de régression. ϵ_i est l'erreur expérimentale pour l'individu i.

La méthode des moindres carrés vise ici à minimiser la quantité $\sum_{i=1}^{n}(y_i - ax_i - b)^2$ par

rapport à a et b. Les valeurs qui réalisent ce minimum sont :

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{et} \quad \widehat{b} = \bar{y} - \widehat{a}\bar{x}$$

où $\bar{x}$ est la moyenne des x_i et $\bar{y}$ celle des y_i.

1.2 Quelques plans d'échantillonnage classiques

Nous considérons le cas d'une population statistique de taille N.

1.2.1 Plan aleatoire simple

Le plan aléatoire simple (PAS) de taille n consiste à effectuer n tirages équiprobables dans la population statistique. Les tirages peuvent être avec ou sans remise.

Pour un plan aléatoire simple sans remise (PASSR), on a C_N^n échantillons possibles, ayant tous la même probabilité de réalisation.

Pour un plan aléatoire simple avec remise (PASAR), on a N^n échantillons possibles, ayant tous la même probabilité de réalisation.

Le plan aléatoire simple est utilisé en phase exploratoire quand on désire estimer un paramètre de la population et qu'il n'y a pas de structure spatiale à étudier.

La moyenne d'échantillon pour un PAS est sans biais pour la moyenne de la population.

1.2.2 Plan aléatoire stratifié

La population est divisée en H strates de taille $N_1, \cdots, N_H$. La moyenne d'échantillon dans la strate h est notée $\overline{y}_h$. La procédure d'échantillonnage consiste à exécuter un PASSR de taille n_h dans la strate h, indépendamment des autres strates.

Le nombre d'échantillons possibles est $\prod_{h=1}^{H} C_{N_h}^{n_h}$.

La moyenne d'échantillon global n'est pas forcément sans biais pour $\overline{Y}$ pour ce type d'échantillonnage. On utilise donc la moyenne dite stratifiée qui, elle, est sans biais :

$$\overline{y}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_h$$

et de variance $Var(\overline{y}_{st}) = \sum_{h=1}^{H} \Big(\frac{N_h}{N}\Big)^2 (1 - \frac{n_h}{N_h}) \frac{V_h}{n_h} \frac{N_h}{N_h - 1}$ où V_h est la variance de la strate h.

Le plan aléatoire stratifié est intéressant quand la variabilité intra-strate est faible.

1.2.3 Plan aléatoire en grappes

La population est divisée en G grappes, pas forcément de même taille. L'échantillonnage consiste à choisir g grappes selon un plan aléatoire simple sans remise.

Le nombre d'échantillons possibles est C_G^g.

Un estimateur sans biais de la moyenne de la population est donnée par

$\bar{y}_{grappes} = \dfrac{G}{g \times N} \times$ [somme des valeurs observées sur les grappes échantillonnées].

Le plan aléatoire en grappes est intéressant quand la variabilité inter-grappe est faible.

1.2.4 Plan aléatoire systématique

Dans le cas d'une population ordonnée, de taille $N = kn$, le plan consiste à choisir un individu dans $\{1, \cdots, k\}$, soit i, et à constituer l'échantillon $\{i, i+k, \cdots, i+(n-1)k\}$.

On a seulement N/n échantillons possibles. La moyenne d'échantillon est sans biais pour la moyenne de la population.

Ce plan est utilisé quand une exploration spatiale a un intérêt. Il est décommandé en cas de périodicité supposée de la variable étudiée si l'on veut estimer la moyenne de la population.

1.3 Recherche d'une procédure d'échantillonnage

La figure 1.4 schématise le processus décisionnel permettant de mettre au point un plan d'échantillonnage. Il s'agir de trouver un juste milieu entre l'effort d'échantillonnage qui sera consenti et la fiabilité du plan en termes de précision ou de minimisation des risques d'erreur.

Le processus décisionnel décrit en figure 3 conduit souvent à la succession d'étapes méthodologiques que voilà :

1. *Etude bibliographique.* Il s'agit de mettre à profit des études antérieures pour construire un plan de sondage performant.
2. *Définition claire des objectifs de l'échantillonnage (ou sondage).* Cette étape doit déboucher sur la définition des variables à prendre en compte et la confection d'une feuille de saisie (ou d'un questionnaire quand il s'agit de sondages d'opinion).
3. *Définition de la population à étudier.* Elle doit être définie sans ambiguité. On définit d'abord la population cible puis on détermine la liste des unités statistiques sélectionnables, autrement dit la base de sondage.
4. *Construction du plan de sondage.* Il s'agit de déterminer la façon dont les individus doivent être sélectionnés, d'organiser l'observation en fonction des contraintes naturelles et techniques. Si les individus sont sélectionnés selon une procédure aléatoire, on parle de plan probabiliste. Sinon, on parle de plan empirique.
5. *Collecte des informations.* L'exécution du plan doit respecter les règles établies à l'étape précédente. La collecte des informations peut se faire à partir d'une base de données informatique, mais aussi à l'aide d'observations sur le terrain, en plein champ. Il peut s'agir également d'enquêtes faites par un enquêteur par entretien, par courrier, par téléphone, par examen. Quelque soit la procédure, il est nécessaire que la qualité, la fiabilité des données soient garanties.
6. *Encodage et archivage des données.* Il s'agit de choisir les logiciels ou programmes informatiques les plus appropriés.
7. *Traitement statistique des données.* Les méthodes doivent tenir compte des caractéristiques du plan d'échantillonnage.

Une partie de la théorie des sondages consiste en l'étude des propriétés de la distribution d'échantillonnage de la moyenne d'échantillon pour différents plans d'échantillonnage dans le cadre d'une population finie fixe. En d'autres termes, on aimerait connaître les propriétés de la série statistique que l'on obtiendrait si, pour la population statistique considérée, l'on pouvait :

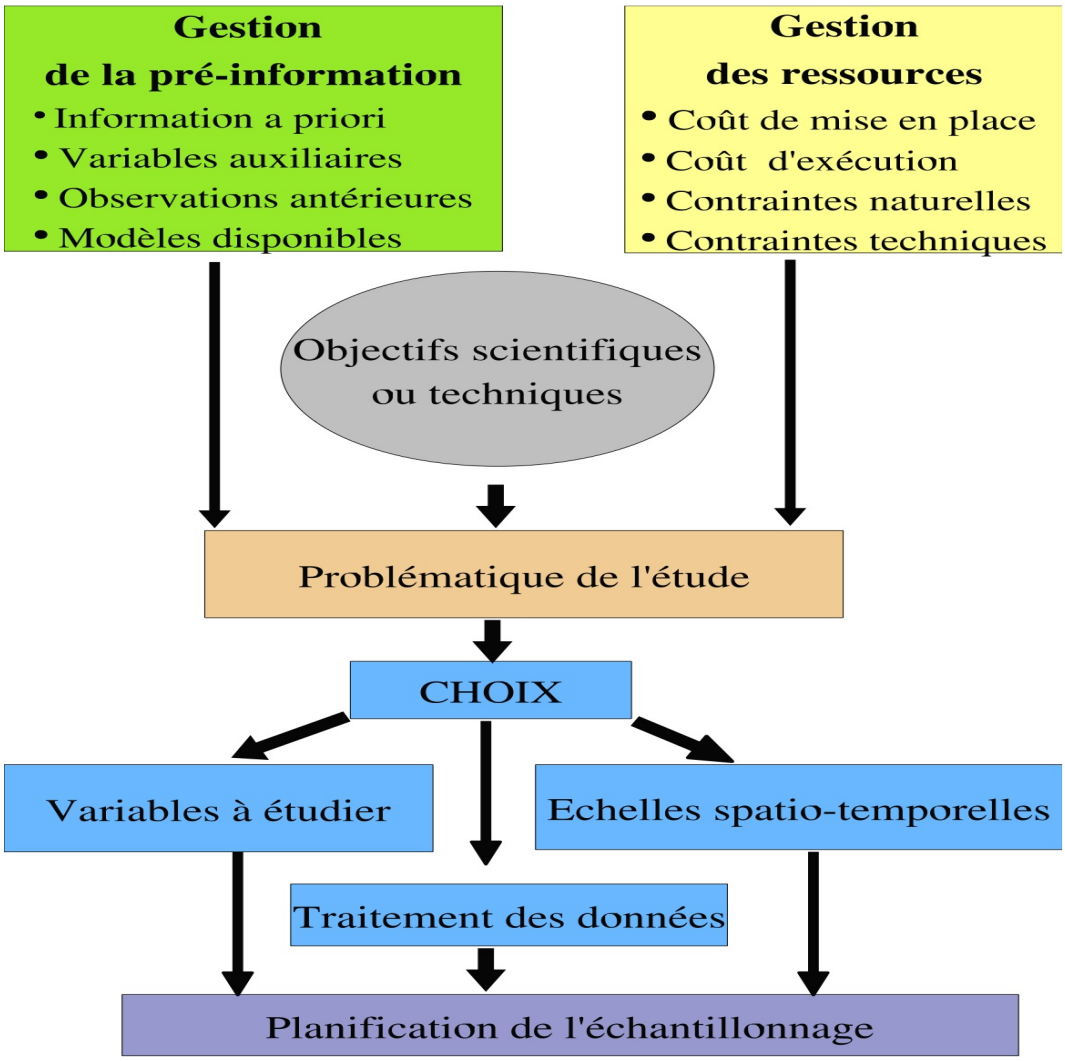

FIGURE 1.4 – Schéma du processus décisionnel pour le choix d'un plan d'échantillonnage

1. réaliser tous les échantillons possibles avec ce plan d'échantillonnage,
2. calculer la moyenne de chacun de ces échantillons,
3. constituer une série statistique avec ces moyennes d'échantillon.

Illustration numérique : On a une population $\mathcal{P} = \{1, 2, 3, 4\}$ sur laquelle un caractère $\mathcal{Y}$ prend les valeurs $y_1 = 17, y_2 = 8, y_3 = 8$ et $y_4 = 23$. On échantillonne en effectuant deux tirages sans remise dans $\mathcal{P}$. Les trois étapes décrites ci-dessus deviennent :

1. Les échantillons possibles sont $\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}$ et $\{3, 4\}$
2. Les moyennes de ces échantillons sont respectivement 12,5 ; 12,5 ; 20 ; 8 ; 15,5 et 15,5.
3. La série statistique est donc 12,5 12,5 20 8 15,5 15,5 La distribution d'échantillonnage de la moyenne d'échantillon est donc P(8)=1/6, P(12,5)=1/3, P(15,5)=1/3, P(20)=1/6. Son espérance mathématique est 14 et sa variance 13,5.

Bien sûr, dans la pratique, la taille de population N est bien plus grande et il est utile de rappeler que l'on ne connaîtra les valeurs y_i que pour les individus i inclus dans l'échantillon. On verra qu'il est possible, pour beaucoup de plans d'échantillonnage, d'exprimer l'espérance m et la variance V de la distribution d'échantillonnage de la moyenne d'échantillon puis d'estimer ces paramètres m et V à l'aide des données de l'échantillon observé. On tâchera de répondre plus particulièrement aux questions essentielles suivantes :

Comment échantillonner (quel plan d'échantillonnage appliquer) ?

Quelle taille d'échantillon adopter (quel taux de sondage appliquer) ?

Comment estimer une fonction d'intérêt avec une bonne précision ?

Ouvrages conseillés :

GRAIS Bernard. (1992) : Méthodes statistiques. Dunod, 3ième édition.

GROSBRAS Jean-Marie. (1987) : Méthodes statistiques des sondages, Economica.

MORIN Hervé. (1993) Théorie de l'échantillonnage, Les presses de l'Université Laval.

TILLE Yves. (2001) Théorie des sondages, Dunod.

Chapitre 2

Eléments pour le choix de la taille d'échantillon

Considérons une population statistique de taille N. Ses individus sont répertoriés et un numéro compris entre 1 à N est affecté à chacun d'eux. On s'intéresse à la taille d'échantillon nécessaire pour estimer la proportion p d'individus vérifiant une certaine propriété. Par exemple, la population statistique est l'ensemble des plantes d'une parcelle, et p est la proportion de plantes infectées.

2.1 Quelques formules pour l'échantillonnage aléatoire simple

On considère un plan d'échantillonnage de taille n. On note $\widehat{p}$ la proportion d'individus vérifiant la propriété dans l'échantillon de taille n.

Dans ce qui suit, $u_{1-\frac{\alpha}{2}}$ est le fractile d'ordre $1-\frac{\alpha}{2}$ de la loi Normale centrée réduite ($u_{0,975} = 1,96$).

2.1.1 Cas avec remise

$\widehat{p}$ est sans biais pour p et de variance $Var(\widehat{p}) = \dfrac{p(1-p)}{n}$.

$Var(\widehat{p})$ est estimée sans biais par $\dfrac{\widehat{p}(1-\widehat{p})}{n-1}$.

2.1.2 Cas sans remise

$\widehat{p}$ est sans biais pour p et de variance $Var(\widehat{p}) = \dfrac{p(1-p)}{n}\dfrac{N-n}{N-1}$ ou encore

$$Var(\widehat{p}) = (1-\frac{n}{N})\frac{p(1-p)}{n}\frac{N}{N-1}.$$

$Var(\widehat{p})$ est estimée sans biais par $(1-\dfrac{n}{N})\dfrac{\widehat{p}(1-\widehat{p})}{n-1}$.

2.2 Choix d'une taille d'échantillonnage sans remise

Le fait que l'erreur absolue soit supérieure à d avec un risque d'erreur au plus égal à α s'exprime par

$$Proba(|\widehat{p}-p| \geq d) \leq \alpha.$$

Si n est suffisamment grand (dans la pratique $n \geq 50$), on peut utiliser l'approximation de la loi hypergéométrique par la loi normale de même espérance et de même variance et ceci nous conduit à

$$d \geq u_{1-\frac{\alpha}{2}}\sqrt{Var(\widehat{p})} \quad \text{puis} \quad d \geq u_{1-\frac{\alpha}{2}}\sqrt{(1-\frac{n}{N})\frac{p(1-p)}{n}\frac{N}{N-1}}.$$

On obtient

$$n \geq \frac{Np(1-p)u^2_{1-\frac{\alpha}{2}}}{p(1-p)u^2_{1-\frac{\alpha}{2}} + (N-1)d^2} \tag{2.1}$$

Remarque : Dans l'expression (2.1), la valeur inconnue p intervient alors qu'elle est bien-sûr inconnue dans la pratique, puisque le but de l'échantillonnage est de l'estimer. Par conséquent, on se sert des deux propriétés suivantes : (1) le membre de droite de l'égalité (2.1) est une fonction croissante de $p(1-p)$; (2) $p(1-p) \leq 1/4$

Ainsi, (2.1) devient

$$\boxed{n \geq \frac{N u^2_{1-\frac{\alpha}{2}}}{u^2_{1-\frac{\alpha}{2}} + 4(N-1)d^2}} \tag{2.2}$$

Exemple numérique : On dispose d'une parcelle de 10000 plantes. On désire évaluer numériquement la proportion P de plantes infectées dans la parcelle avec une erreur absolue inférieure à 4% et un risque d'erreur $\alpha \leq 0,05$.

On applique (2.2) avec $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$;$N = 10000$; $d = 0,04$.

On obtient $n = 567$.

Pour d=6%, on aurait $n = 259$ et pour d=10%, on aurait $n = 96$.

2.3 Choix d'une taille d'échantillonnage avec remise

A partir de l'expression de $\mathrm{Var}(\widehat{p})$, on déduit

$$d \geq u_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

Par conséquent,

$$n \geq \frac{p(1-p)u^2_{1-\frac{\alpha}{2}}}{d^2} \tag{2.3}$$

Le membre de droite de l'égalité (2.3) est une fonction croissante de $p(1-p)$. Comme $p(1-p) \leq 1/4$, il en résulte :

$$\boxed{n \geq \frac{u^2_{1-\frac{\alpha}{2}}}{4d^2}} \tag{2.4}$$

On applique (2.4) avec $u_{1-\frac{\alpha}{2}} = u_{0,975} = 1,96$ et $d = 0,04$. On obtient $n \geq 601$.

Pour d=6%, on aurait $n = 267$ et pour d=10%, on aurait $n = 97$.

2.4 Utilisation de l'information a priori sur p

Dans (2.1) et (2.3) le membre de droite de l'égalité est une fonction croissante de $p(1-p)$. Si on sait que $p \leq p_0 \leq 1/2$, p_0 étant une valeur seuil que ne peut atteindre p, alors on a :

Cas sans remise :

$$\boxed{n \geq \frac{N u^2_{1-\frac{\alpha}{2}}}{u^2_{1-\frac{\alpha}{2}} + \frac{(N-1)d^2}{p_0(1-p_0)}}} \tag{2.5}$$

Cas avec remise :

$$\boxed{n \geq \frac{p_0(1-p_0)u^2_{1-\frac{\alpha}{2}}}{d^2}} \tag{2.6}$$

(2.5) et (2.6) permettent de diminuer de façon substantielle la taille d'échantillonnage n.

Reprenons l'exemple numérique présenté plus haut. Si on sait que la proportion p ne peut atteindre 20%, on posera $p_0 = 0,20$ dans (2.5) et (2.6) :

Cas sans remise :

$n \geq 370$ pour $d = 0,04$, $\alpha = 0,05$ et $p_0 = 0,20$

$n \geq 168$ pour $d = 0,06$, $\alpha = 0,05$ et $p_0 = 0,20$

$n \geq 62$ pour $d = 0,10$, $\alpha = 0,05$ et $p_0 = 0,20$.

Cas avec remise :

$n \geq 385$ pour $d = 0,04$, $\alpha = 0,05$ et $p_0 = 0,20$

$n \geq 171$ pour $d = 0,06$, $\alpha = 0,05$ et $p_0 = 0,20$

$n \geq 62$ pour $d = 0,10$, $\alpha = 0,05$ et $p_0 = 0,20$.

2.5 Conclusions

La taille n d'échantillon, proposée dans les situations précédentes, permet d'affirmer, avec un risque égal à α de se tromper, que l'écart entre la vraie valeur de p et son évaluation $\widehat{p}$ est inférieure en valeur absolue à un nombre d fixé à l'avance. La taille n proposée est donc fonction de α et de d.

Si, en outre, on sait que la proportion p que l'on cherche à évaluer est inférieure à une valeur seuil p_0 vérifiant $p_0 \leq 1/2$, on peut diminuer substantiellement la taille d'échantillon. La taille n est alors fonction de α, de d et aussi de p_0.

Chapitre 3

Eléments de Statistique Descriptive

3.1 Vocabulaire général de la Statistique Descriptive

La Statistique Descriptive a pour souci la description des données et la mise en évidence de certains aspects de l'information contenue dans ces données. Elle utilise pour cela des méthodes graphiques et des résumés numériques donnant des indications de différentes natures sur les données.

Voici par ordre alphabétique une liste de termes fréquemment utilisés en Statistique descriptive.

Amplitude d'une classe (ou d'un intervalle) : C'est la longueur de l'intervalle. L'amplitude de la classe $]a_{i-1}; a_i]$ est $a_i - a_{i-1}$. Exemple : la classe]16 ;43] est d'amplitude $43 - 16 = 27$ (unités de mesure).

Caractère qualitatif : Un caractère statistique est qualitatif si ses valeurs, ou modalités, s'expriment de façon littérale ou par un codage sur lequel les opérations arithmétiques telles que moyenne, somme,$\cdots$, n'ont pas de sens. Exemples : *Sexe de la personne interrogée, Situation familiale, Numéro de son département de naissance ; Etat du temps constaté à une station expérimentale chaque jour ; Variété de la plante observée, Etat sanitaire, numéro de Site.*

Caractère quantitatif : Un caractère statistique est quantitatif si ses valeurs sont des nombres sur lesquels des opérations arithmétiques telles que somme, moyenne,$\cdots$, ont un sens. Exemples : *Taille, Poids, Salaire, Rendement, Note à un examen, PNB/habitant,*

Espérance de vie, Nombre d'habitants, Taux d'infestation.

Caractère statistique (ou variable statistique) : C'est ce qui est observé ou mesuré sur les individus d'une population statistique. Il peut s'agir d'une variable qualitative ou quantitative.

Classe modale : C'est la classe correspondant à une *fréquence par amplitude* maximale. Elle correspond au maximum de l'histogramme (plus grand effectif par unité d'amplitude). Dans le cas d'une classe modale unique, on parle de distribution continue unimodale.

Classe statistique : Intervalle de valeurs d'une variable statistique. L'ensemble des classes forment une partition de l'ensemble des valeurs possibles de la variable. Par exemple, si tous les salaires des employés d'une entreprise se situent entre 1000 et moins de 20000 EUR, on peut construire (par exemple) les classes :

$$]1000; 3000],]3000; 5000],]5000; 7000],]7000; 20000]$$

Les classes statistiques sont exclusives c'est-à-dire une valeur observée appartient à une classe et une seule.

Remarque : on peut utiliser une distribution en classes statistiques pour une variable discrète pouvant prendre beaucoup de valeurs distinctes. Exemple : *nombre d'insectes par unité d'échantillonnage* dans le cas de pullulation.

Coefficient de corrélation (linéaire) : Le coefficient de corrélation entre deux variables quantitatives X et Y est le nombre r vérifiant : $r = \dfrac{s_{xy}}{s_x s_y}$

où s_{xy} est la covariance entre X et Y, et s_x, s_y les écarts-types de X et Y.

Ce coefficient est toujours compris entre -1 et + 1.

S'il est très proche de + 1 ou - 1 , X et Y sont presque parfaitement corrélées linéairement, c'est-à-dire qu'elles sont liées entre elles par une relation presque affine ; le nuage de points est presque aligné le long d'une droite (croissante si $r = +1$, décroissante si $r = -1$). S'il n'y a aucun lien linéaire entre X et Y, ce coefficient est nul, ou presque nul.

Coefficient de Spearman (ou coefficient de corrélation des rangs) : Il est utilisé dans le cas de deux variables ordinales X et Y (qui peuvent être deux variables quantitatives transformées en rangs). Il s'agit du coefficient de corrélation entre le rang des individus pour X et le rang des individus pour Y.

Coefficient de variation : C'est le rapport écart-type sur la moyenne. Il est calculé pour des variables statistiques positives par exemple taille, durée, poids. C'est un nombre sans dimension (c'est-à-dire qu'il est indépendant du choix des unités de mesure). Il permet de comparer la dispersion autour de la moyenne de variables statistiques ayant des échelles ou des unités de mesure différentes.

Courbe cumulative : On l'utilise quand la variable quantitative est continue. Il s'agit d'une fonction continue, affine par morceaux. Pour la tracer, on relie les points $(x_i, F(x_i))$, pour les points distincts x_i de la série statistique.

Diagramme circulaire (ou à secteurs angulaires ou camembert) : Il s'agit d'un disque divisé en sections angulaires. Chaque section correspond à une modalité de la variable qualitative et a un angle proportionnel à la fréquence de cette modalité.

Diagramme cumulatif : C'est le tracé de la fonction qui à tout x associe $F(x) =$ proportion d'observations $\leq x$. Il s'obtient au moyen des effectifs cumulés croissants. On a une fonction dite en escalier. On l'utilise dans le cas d'une variable quantitative discrète.

Diagramme figuratif : Chaque modalité de la variable qualitative est représentée par une image (ordinateur, maison, plante, avion,...) rappelant la variable (ou la population) statistique étudiée, et de taille proportionnelle à la fréquence de cette modalité.

Dispersion : Un indicateur statistique est dit de dispersion s'il s'agit d'un nombre clé caractérisant la variabilité des observations dans la série statistique. Ainsi l'étendue donne l'écart entre la plus petite et la plus grande valeur dans la série statistique ; l'écart interquartile donne la plage de variation des observations situées dans le second et troisième quarts de la série statistique réordonnée.

Distribution statistique : Ensemble des modalités, valeurs, ou classes d'une variable, avec les effectifs observés correspondants.

Ecart interquartile : C'est la différence I entre le 1er et le 3ème quartile : $I = Q_3 - Q_1$.

Ecart-type : pour une distribution d'effectifs $(x_1, n_1), \cdots, (x_k, n_k)$, où x_i a pour effectif associé n_i,, l'écart-type noté s_x est donné par la formule :

$$s_x = \sqrt{\frac{1}{n}(n_1(x_1 - \bar{x})^2 + \cdots + n_k(x_k - \bar{x})^2)}$$

où $\bar{x}$ est la moyenne de la série.

Etendue : C'est l'écart entre la plus petite et la plus grande valeur dans la série statistique.

Fractiles (ou quantiles) : On appelle fractiles des valeurs divisant une série en plusieurs parties. Pour une valeur α comprise entre 0 et 1, le fractile d'ordre α noté q_α est, par définition, tel que la proportion de valeurs inférieures à q_α vaut α. On a donc $F(q_\alpha) = \alpha$. Les fractiles divisant la série en k parties d'effectifs égaux ont parfois une dénomination commune : Les 3 quartiles divisent la série en 4 parties d'effectifs égaux, les 9 déciles en 10, les 99 centiles en 100. Les 3 quartiles sont notés Q_1, Q_2, Q_3 (Q_2 étant la médiane).

Fréquence (ou fréquence relative) : C'est la proportion (ou le pourcentage) d'individus pour lesquels une variable statistique a pris une valeur donnée. Si, sur 150 familles, 50 ont 2 enfants, on dira que la fréquence f_i correspondant à la valeur $x_i = 2$ de la variable

nombre d'enfants, est : 0.33 ou 1/3 ou 33.33%.

Fréquence cumulée : Résultat de l'addition, de proche en proche, des fréquences d'une distribution observée, soit en commençant par le 1er :

$F_1 = f_1$, $F_2 = f_1 + f_2, \cdots,$ $F_i = f_1 + f_2 + \cdots + f_i$ (fréquences cumulées croissantes),

soit en commençant par le dernier :

$F_K^* = f_K, F_{K-1}^* = f_K + f_{K-1}, \cdots, F_i^* = f_K + f_{K-1} + \cdots + f_i$ (fréquences cumulées décroissantes).

Histogramme : Graphique permettant de représenter une distribution continue regroupée en classes : rectangles juxtaposés dont les bases sont les classes, et les surfaces sont proportionnelles aux effectifs (ou fréquences) associés.

Indépendance : Deux variables statistiques X et Y sont dites indépendantes si la distribution de Y conditionnelle à $X = x$, pour tout x, est constante (c'est-à-dire ne dépend pas de x). Cela signifie que les profils des lignes du tableau de contingence sont identiques, ou de façon équivalente que les profils des colonnes du tableau de contingence sont identiques, et donc que la distribution de fréquences conditionnelle est égale à la distribution de fréquences marginale.

Indicateur statistique (ou résumé numérique) : C'est un nombre permettant de résumer numériquement les traits principaux d'une distribution statistique. On parle aussi de résumé numérique. On distingue principalement deux types d'indicateurs :
- les indicateurs de position (ou de tendance centrale) qui donne une idée de l'ordre de grandeur de la série ;
- les indicateurs de dispersion qui donnent une idée de la variabilité dans la série.

Inégalité de (Bienaymé-) Tchébichev : Pour toute série statistique $x_1, \cdots, x_n$ de moyenne $\bar{x}$ et d'écart-type s_x , la proportion de valeurs dans l'intervalle $[\bar{x} - k \times s_x ; \bar{x} + k \times s_x]$ est supérieure à $1 - \dfrac{1}{k^2}$, pour tout nombre $k \geq 1$. Par exemple, au moins 75% des valeurs appartiennent à : $[\bar{x} - 2s_x ; \bar{x} + 2s_x]$, c'est-à-dire s'écartent de moins de 2 écart-types de la moyenne.

Intervalle interquartile : C'est l'intervalle dont les bornes sont le 1er et le 3ème quartile : $[Q_1, Q_3]$. Il contient 50% des observations ; rappelons que 25% des valeurs de la série statistique sont inférieures à Q_1 et 25% sont supérieures à Q_3.

Intervalle médian : C'est l'intervalle dont toutes les valeurs vérifient la propriété de la médiane pour la série statistique étudiée.

Médiane : C'est le fractile d'ordre 0.5. La médiane est notée M_e et vérifie $F(M_e) = 0.5$. Il y a autant de valeurs inférieures à M_e que de valeurs supérieures à M_e dans la série

statistique.

Mode : C'est la valeur la plus fréquente dans la série statistique. Le mode n'est pas forcément unique. Quand il existe plusieurs modes, la distribution statistique est dite multimodale.

Moyenne : Pour une distribution d'effectifs $(x_1, n_1), \cdots, (x_k, n_k)$, où x_i a pour effectif associé n_i, la moyenne notée $\bar{x}$ est la somme des valeurs divisée par le nombre de valeurs. Elle est donné par la formule : $\bar{x} = \frac{1}{n}(n_1 x_1 + \cdots + n_k x_k)$.

Nuage de points : Ensemble de points isolés représentés dans un graphique cartésien. Une séries à deux caractères quantitatifs $(x_1, y_1); (x_2, y_2); \cdots; (x_n, y_n)$ peut être représentée par les n points $M_1, M_2, \cdots, M_n$ de coordonnées $(x_1, y_1); (x_2, y_2); \cdots; (x_n, y_n)$.

Population statistique : Une population statistique est l'ensemble sur lequel on effectue des observations. Exemples : ensemble de personnes interrogées pour une enquête ; ensemble de parcelles cultivées sur lesquelles on mesure un rendement ; ensemble de pays pour lesquels on dispose de données géographiques ou économiques, ...

Position : Un indicateur statistique est dit de position (ou de tendance centrale) s'il s'agit d'un nombre clé permettant de préciser où se répartit une certaine fraction des observations. Ainsi les quartiles permettent de situer le quart inférieur, la moitié, le quart supérieur des observations.

Profil : C'est une distribution conditionnelle de fréquences (et non d'effectifs). Dans un tableau de contingence à I lignes et J colonnes, le profil de la ligne i est obtenu en divisant les effectifs $n_{i1}, n_{i2}, \cdots, n_{iJ}$ de cette ligne par la somme $n_{i.}$ de ces effectifs. On obtient : $\frac{n_{i1}}{n_{i.}}, \frac{n_{i2}}{n_{i.}}, \cdots, \frac{n_{iJ}}{n_{i.}}$. De même, le profil de la colonne j est : $\frac{n_{1j}}{n_{.j}}, \frac{n_{2j}}{n_{.j}}, \cdots, \frac{n_{Ij}}{n_{.j}}$. où $n_{.j}$ est la somme des effectifs de cette colonne.

Quartiles : Ce sont les 3 fractiles d'ordre 0,25, 0.5 et 0,75. Ils sont notés dans l'ordre croissant Q_1, Q_2, Q_3. Ils divisent la distribution statistique en quatre parties d'égale fréquence. Q_1 est le premier quartile, Q_3 le troisième. Q_2 est la médiane. (voir fractiles).

Résumé numérique : Voir indicateur statistique.

Série statistique (ou distribution observée) : Séquence des modalités, ou valeurs d'une variable statistique. L'ordre correspond souvent à l'ordre chronologique de recueil des observations.

Statistique Descriptive : Ensemble des méthodes et techniques permettant de présenter, de décrire, de résumer des données nombreuses et variées.

Statistique Descriptive univariée : La Statistique Descriptive univariée consiste en la description de chacun des caractères statistiques, un par un, et non des liens éventuels

existant entre eux.

Statistique Descriptive bivariée : La Statistique Descriptive bivariée consiste en la description de deux variables mesurées simultanément sur les mêmes individus. Elle permet de mettre en évidence le type de lien existant éventuellement entre ces variables.

Tableau de contingence : C'est le tableau d'effectifs obtenu par tri croisé d'une série bivariée (ou multivariée).

Tri à plat d'une série statistique brute : C'est l'inventaire des modalités ou valeurs rencontrées dans la série, avec les effectifs correspondants.

Tri croisé d'une série bivariée : C'est l'inventaire des modalités ou valeurs rencontrées conjointement dans une série comportant deux variables mesurées pour chaque individu statistique, avec les effectifs correspondants.

Variable statistique (ou caractère statistique) : C'est ce qui est observé ou mesuré sur les individus d'une population statistique. Il peut s'agir d'une variable qualitative ou quantitative.

Variance : Pour une distribution d'effectifs $(x_1, n_1), \cdots, (x_k, n_k)$, où x_i a pour effectif associé n_i, la variance notée s_x^2 est donnée par la formule :

$s_x^2 = \frac{1}{n}(n_1(x_1 - \bar{x})^2 + \cdots + n_k(x_k - \bar{x})^2)$. La variance est le carré de l'écart-type.

3.2 Quelques formules de résumés numériques

Les indicateurs statistiques (ou résumés numériques) sont des valeurs fournissant des indications sur le(s) caractère(s) étudié(s). Comme précisé auparavant, on distingue principalement deux types d'indicateurs. D'une part, les indicateurs de position (ou de tendance centrale) qui donne une idée de l'ordre de grandeur de la série, par exemple : moyenne, mode, médiane, quantiles. D'autre part, les indicateurs de dispersion qui donnent une idée de la variabilité dans la série, par exemple : écart-type, étendue, écart interquartile.

Soit une série statistique de longueur n correspondant aux observations d'un caractère quantitatif sur un échantillon de taille n. Soit k le nombre de valeurs distinctes dans cette série. On note n_i l'effectif correspondant à la valeur (ou modalité) x_i, et $f_i = \frac{n_i}{n}$ sa fréquence relative. Le tableau (3.1) donne l'expression mathématique pour quelques indicateurs statistiques de **tendance centrale** et de **dispersion**. Le tableau (3.2) fournit quelques indicateurs de **forme** et de **concentration**.

Moyenne de la série	$\bar{x} = \dfrac{n_1 x_1 + ... + n_k x_k}{n} = f_1 x_1 + ... + f_k x_k = \sum_{i=1}^{k} f_i x_i$
Variance de la série	$s^2 = \dfrac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2$
Propriété fondamentale :	$s^2 = \left(\dfrac{1}{n} \sum_{i=1}^{k} n_i x_i^2 \right) - \bar{x}^2$
Ecart-type de la série	$s = \sqrt{\dfrac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2}$
Variance corrigée de la série	$s'^2 = \dfrac{1}{n-1} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = \dfrac{n}{n-1} s^2$
Coefficient de variation de la série (n'a de sens que si la variable est positive)	$C_v = \dfrac{s}{\bar{x}}$

TABLE 3.1 – Quelques indicateurs de tendance centrale et de dispersion

Coefficient d'asymétrie de Yule	$c_Y = \dfrac{(Q_3 - M_e) - (M_e - Q_1)}{Q_3 - Q_1}$
Premier coefficient d'asymétrie de Pearson	$c_P = \dfrac{\bar{x} - M_o}{s}$ où M_o est le mode de la série
Second coefficient d'asymétrie de Pearson	$\beta_1 = \Big(\dfrac{1}{s^3}\sum_{i=1}^{k} f_i(x_i - \bar{x})^3\Big)^2$
Coefficient d'asymétrie de Fisher	$\gamma_1 = \dfrac{1}{s^3}\sum_{i=1}^{k} f_i(x_i - \bar{x})^3$
Coefficient d'aplatissement de Pearson	$\beta_2 = \dfrac{1}{s^4}\sum_{i=1}^{k} f_i(x_i - \bar{x})^4$
Coefficient d'aplatissement de Fisher	$\gamma_2 = \beta_2 - 3$
Indice de Gini	$I_G = \dfrac{1}{2\bar{x}n^2}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i n_j \lvert x_i - x_j\rvert$

TABLE 3.2 – Quelques indicateurs de forme et de concentration

Obtention de la médiane M_e d'une distribution en classes statistiques :

Soit N_i l'effectif cumulé en a_i. On rappelle que N_i est le nombre d'observations inférieures ou égales à a_i dans la distribution et que M_e est telle que l'effectif cumulé en $M_e = n/2$. Soit la classe $]a_{i-1}, a_i]$ telle que $N_{i-1} \leq n/2$ et $N_i \geq n/2$. On a :

$$M_e = a_{i-1} + (n/2 - N_{i-1})\frac{a_i - a_{i-1}}{n_i} \quad .$$

Plus généralement, on obtient, avec ce même procédé d'interpolation linéaire, le quantile d'ordre α noté q_α qui est la valeur telle que l'effectif cumulé en q_α est $n \times \alpha$. On a

$$q_\alpha = a_{i-1} + (n \times \alpha - N_{i-1})\frac{a_i - a_{i-1}}{n_i}$$

où la classe $]a_{i-1}, a_i]$ vérifie cette fois $N_{i-1} \leq n \times \alpha$ et $N_i \geq n \times \alpha$.

Distribution à deux caractères quantitatifs : Considèrons une série statistique double de longueur n ayant I valeurs distinctes pour le premier caractère et J valeurs distinctes pour le second caractère. On note n_{ij} l'effectif correspondant au couple de valeurs (x_i, y_j), et $f_{ij} = \frac{n_{ij}}{n}$ sa fréquence. Quelques indicateurs statistiques de dépendance sont proposés au tableau (3.3).

Distribution à deux caractères qualitatifs : Dans une telle situation, on dispose souvent du tableau de tri croisé constitué des effectifs n_{ij}, à I lignes et J colonnes. n_{ij} est le nombre d'individus statistiques de l'échantillon ayant la $i$ème modalité du premier caractère et la $j$ème modalité du second caractère. On peut alors construire un tableau de même dimension que celui du tri croisé ayant les particularités suivantes :

(a) les totaux des lignes ou des colonnes sont les mêmes que le tableau du tri croisé

(b) Les profils lignes sont égaux ; les profils colonnes sont égaux.

De cette dernière propriété, vient le nom de *tableau d'indépendance parfaite* pour ce tableau : la valeur à la ligne i et à la colonne j est $\nu_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ où $n_{i.} = n_{i1} + ... + n_{iJ}$ est l'effectif total pour la ligne i et $n_{.j} = n_{1j} + ... + n_{Ij}$ celui dela colonne j.

ν_{11}	.	.	ν_{1j}	.	.	ν_{1J}	
.			.			.	
ν_{i1}	.	.	ν_{ij}	.	.	ν_{iJ}	$n_{i.}$
.						.	
ν_{I1}	.	.	ν_{Ij}	.	.	ν_{IJ}	
			$n_{.j}$				

Covariance entre les deux caractères (Formules simple et développée)	$s_{xy} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(x_i - \bar{x})(y_j - \bar{y})$ $s_{xy} = \Big(\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}x_i y_j\Big) - \bar{x}\bar{y}$
Coefficient de corrélation linéaire entre les deux caractères	$\rho_{xy} = \frac{s_{xy}}{s_x s_y}$ où s_x et s_y sont les écart-types respectifs des deux caractères
Coefficients de la droite de régression d'équation $y = \hat{a}x + \hat{b}$	$\hat{a} = \frac{s_{xy}}{s_x^2}$ et $\hat{b} = \bar{y} - \hat{a}\bar{x}$

TABLE 3.3 – Quelques indicateurs de corrélation

La similarité éventuelle entre le tableau du tri croisé et le tableau d'indépendance parfaite permet de se faire une idée du degré de dépendance entre les deux variables statistiques. Pour cela, on calcule le *critère du khi-2 d'indépendance* : $\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \nu_{ij})^2}{\nu_{ij}}$.

Plus ce critère est proche de zéro, plus les ν_{ij} sont proches des n_{ij}. Est ce que cela veut dire qu'on peut accepter l'hypothèse d'indépendance entre les deux variables qualitatives étudiées ? On verra au chapitre 4 comment tester statistiquement cette hypothèse d'indépendance.

3.3 Quelques graphiques

Les représentations graphiques de données doivent être adaptées à la nature des variables statistiques étudiées et leur nombre. Pour une variable qualitative, les graphiques appropriés sont le diagramme circulaire et le diagramme en tuyau d'orgues (figure 3.1). Pour une variable quantitative discrète, le diagramme en bâtons représente parfaitement la distribution de fréquences ou d'effectifs et respecte le fait que les valeurs possibles de la variable sont isolées sur l'axe des réels. Il est de plus souvent dans la pratique remplacé

par le diagramme en tuyaux d'orgue, disons plus esthétique (figure 3.2). L'histogramme est le diagramme le plus utilisé pour une variable quantitative continue, mais la courbe cumulative est également utilisée car elle permet d'évaluer graphiquement certains fractiles, par exemples les 3 quartiles (figure 3.3).

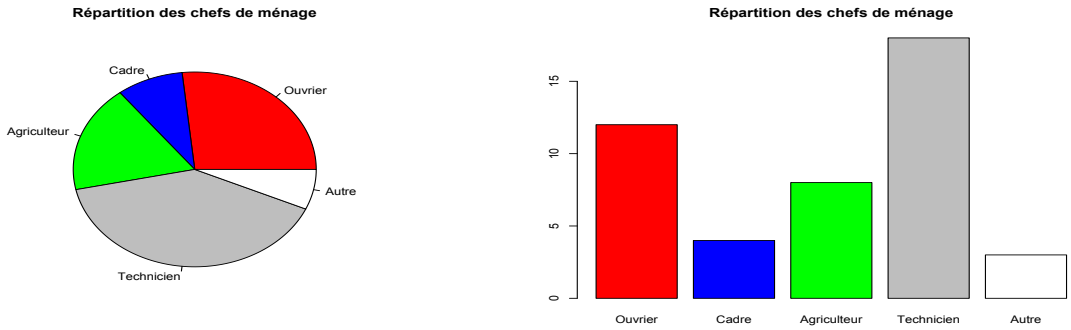

FIGURE 3.1 – Exemples de représentation d'une variable qualitative.

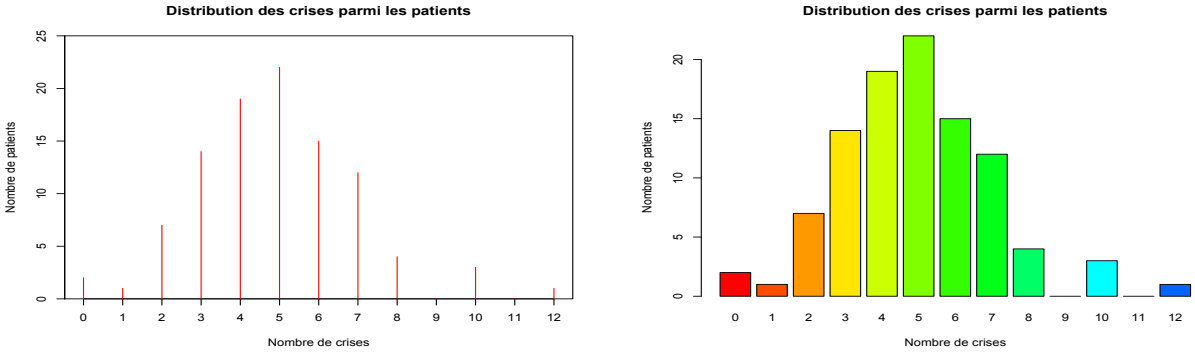

FIGURE 3.2 – Exemples de représentation d'une variable quantitative discrète.

Si l'on considère une distribution statistique à deux variables, plusieurs possibilités de graphiques se présentent également selon la nature des données. Tâchons de passer en revue succinctement les plus fréquentes. Dans le cas de deux variables qualitatives, on peut faire une représentation de l'une des deux variables pour chaque modalité de l'autre. Cette dernière est choisie comme celle qui donne le plus de sens aux distributions conditionnelles, ou simplement celle qui a le moins de modalités. La figure 3.4 illustre le cas où le diagramme choisi est celui à secteurs angulaires. Une autre situation est celle où une variable est quantitative et l'autre qualitative : on peut représenter alors la variable

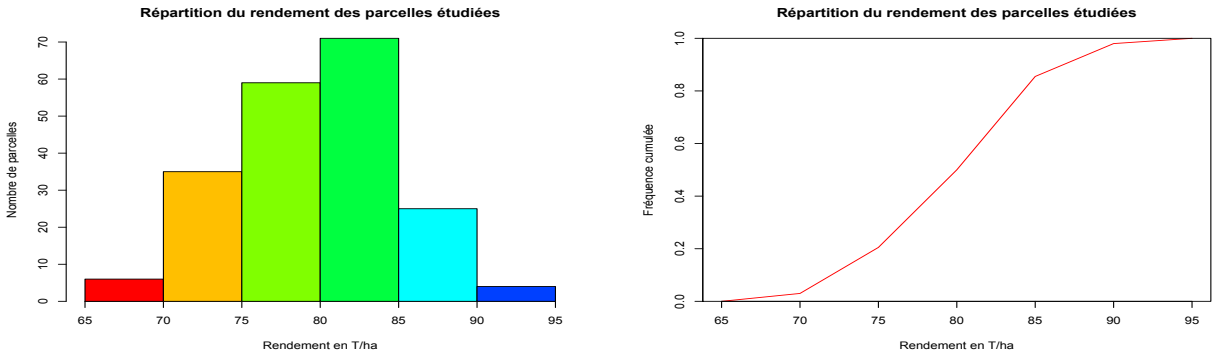

FIGURE 3.3 – Exemples de représentation d'une variable quantitative continue.

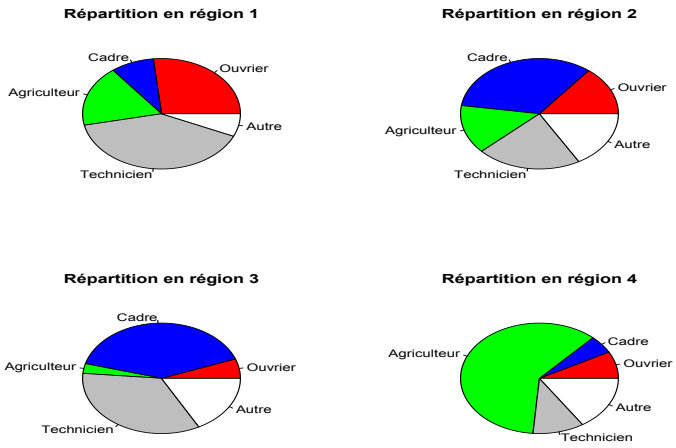

FIGURE 3.4 – Exemple de représentation de 2 variables qualitatives.

quantitative pour chaque modalité de la variable qualitative comme le montre la figure 3.5. On voit sur cette dernière que la représentation peut être soit sur des graphiques séparés, soit simultanément sur un même graphique.

Certains diagrammes se prêtent plus à l'une ou l'autre des méthodes. Par exemple, la boîte à moustaches ou box-plot est souvent utilisée pour représenter plusieurs échantillons d'une même variable statistique, chaque échantillon étant associé à une modalité d'une

variable qualitative (figure 3.5).

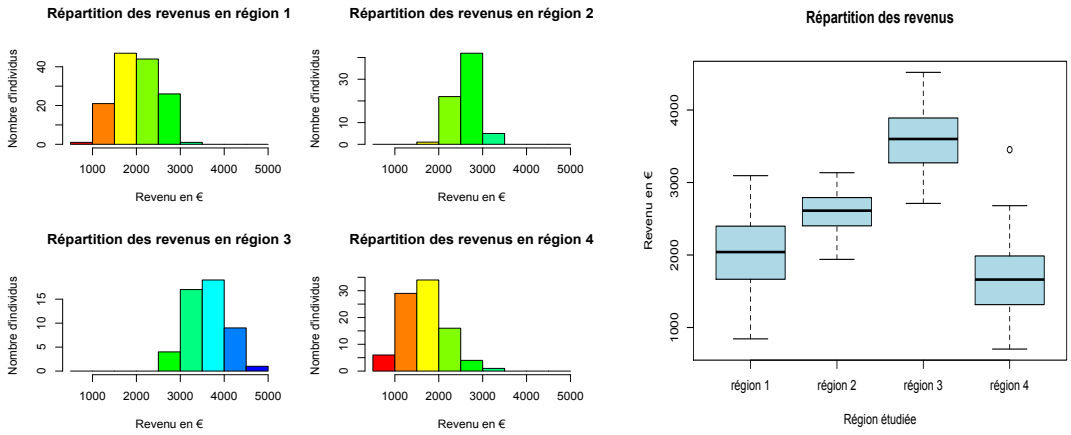

FIGURE 3.5 – Exemples de représentation de 2 variables de nature différente.

Les diagrammes box-plot sont utilisés également quand on dispose de deux variables quantitatives dont une continue et l'autre discrète avec peu de valeurs observées pour cette dernière (figure 3.6). Si le nombre de valeurs observées est grand pour les deux variables étudiées, le nuage de points est préférable. Il permet de visualiser les relations éventuelles entre variables. Dans le cas de deux caractères continus, ce graphique est incontournable (figure 3.7).

Dans ce paragraphe, nous nous sommes intéressés à des distributions statistiques à une ou deux variables. Quand le nombre de variables est plus élevé, d'autres outils graphiques sont possibles. Ils reviennent parfois à représenter les variables 2 à 2. L'utilisation de la couleur permet de visualiser des distributions tridimensionnelles (c'est-à-dire à trois caractères) mais en dimension supérieure à trois, les représentations appropriées ne sont pas faciles. Plus le nombre de variables est important, plus il est difficile de visualiser nos données. C'est la raison pour laquelle les techniques d'exploration des données multidimensionnelles mettent l'accent sur la réduction de dimension : le principe est de représenter les données dans un espace de moindre dimension par des méthodes dites de projection, tout en extrayant un maximum d'information. La qualité de restitution de l'information est mesurée par un critère appelé *taux de représentativité* qui permet de connaitre le degré de fiabilité de la représentation effectuée.

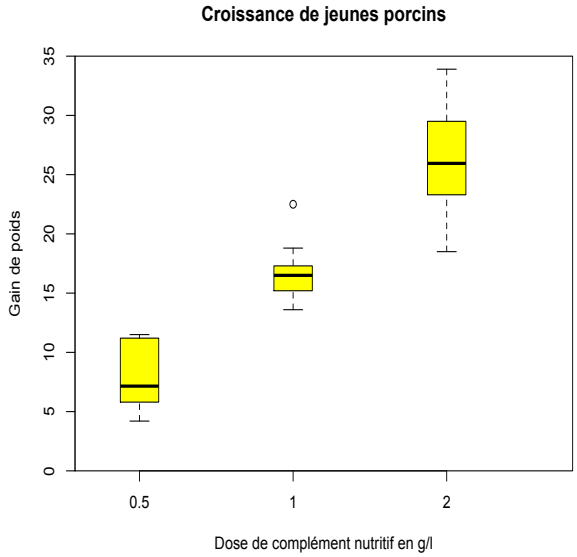

FIGURE 3.6 – Exemple de boîte à moustaches pour 2 variables quantitatives.

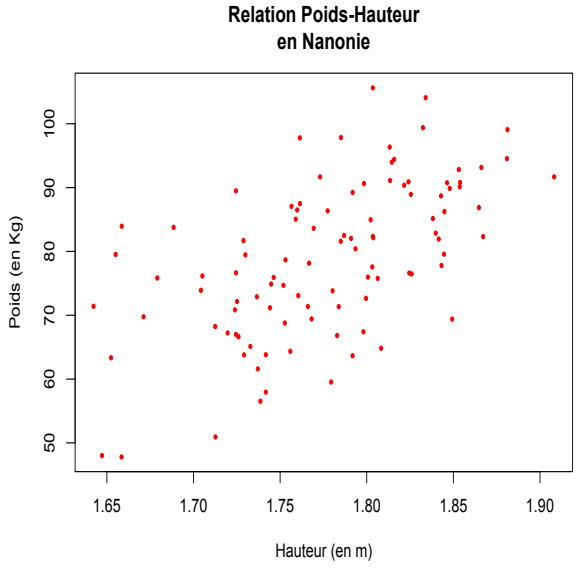

FIGURE 3.7 – Exemple de nuage de points.

3.4 Cas des séries chronologiques

Considérons N relevés sucessifs d'une variable quantitative Y. S'il s'agit de relevés équidistants dans le temps, les valeurs observées peuvent être indicées par des numéros t allant de 1 à N, qui correspondent aux rangs d'observation de ces N relevés. On peut donc écrire la série des valeurs de la façon suivante : $y_1, y_2, \cdots, y_t, \cdots, y_N$.

La première chose à faire est la visualisation de la série. Pour cela, on trace une courbe représentative de cette série, en reliant par un segment de droite les points (t, y_t) et $(t+1, y_{t+1})$ pour t allant de 1 à $N-1$.

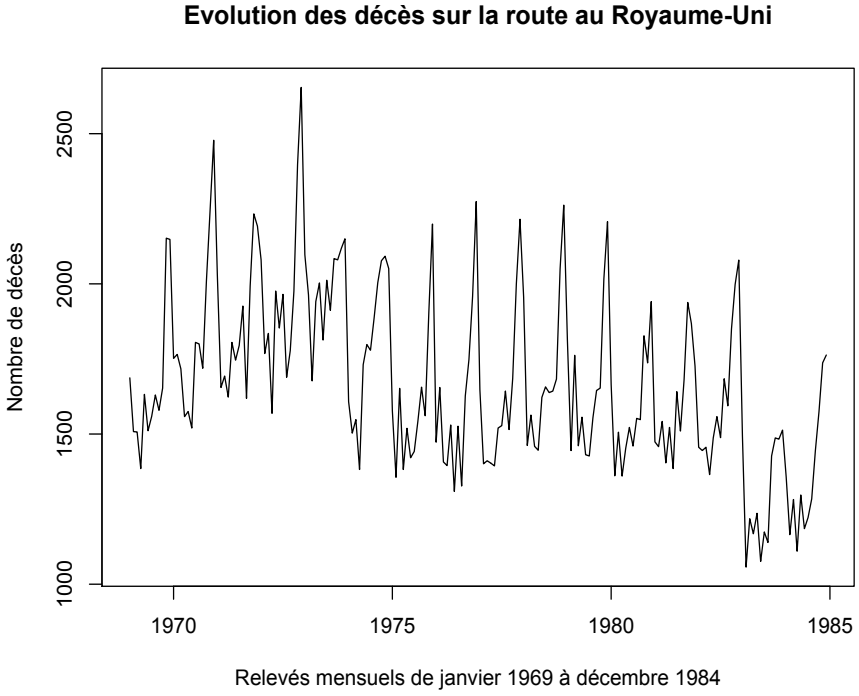

FIGURE 3.8 – Exemple de série chronologique

L'allure de la courbe permet généralement de distinguer :

- un mouvement de longue durée (appelée tendance ou trend) notée (f_t). Le trend schématise la tendance générale du phénomène.
- une composante saisonnière notée (S_t). Elle traduit les variations dites saisonnières qui résultent d'évènements se répétant à l'identique de

période en période. On pose $S_t = c_i$ pour $t = i + kp$ où p est la longueur de la période. c_i est appelé coefficient saisonnier (de la saison i), i variant de $1, \cdots, p$.
- la composante résiduelle (e_t) correspondant à des mouvements perturbateurs irréguliers et imprévisibles.

Traditionnellement, on considère (après transformation éventuelle des données brutes) deux types de schéma :

- le schéma additif : $y_t = f_t + S_t + e_t$
- le schéma multiplicatif : $y_t = f_t \times S_t + e_t$

L'estimation de la tendance à la date t notée $\widehat{f}_t$ est obtenu par lissage de la série chronologique (généralement par le procédé des moyennes mobiles que l'on peut compléter par une courbe de régression). Si on considère qu'il y a un effet saisonnier, on estime les coefficients saisonniers, et l'estimation du $i$ème coefficient saisonnier est notée $\widehat{c}_i$.

	schéma additif	schéma multiplicatif
série ajustée	$\widehat{y}_t = \widehat{f}_t + \widehat{c}_i$	$\widehat{y}_t = \widehat{f}_t \times \widehat{c}_i$
série corrigée des variations saisonnières	$y_t^* = y_t - \widehat{c}_i$	$y_t^* = y_t / \widehat{c}_i$
prévision à l'horizon h	$\widehat{y}_{t+h} = \widehat{f}_{t+h} + \widehat{c}_i$	$\widehat{y}_{t+h} = \widehat{f}_{t+h} \times \widehat{c}_i$

Moyennes mobiles $(M_{t,p})$ d'ordre p : elles ne sont calculables que pour les numéros de relevés t tels que $k < t < N - k$ où k est l'entier tel que $p = 2k + 1$ ou bien $p = 2k$).

Ordre	$p = 2k + 1$	$p = 2k$
Moyenne mobile	$M_{t,p} = \dfrac{1}{2k+1} \displaystyle\sum_{i=-k}^{i=k} y_{t+i}$	$M_{t,p} = \dfrac{1}{2k} \Big[(\displaystyle\sum_{i=-k+1}^{i=k-1} y_{t+i}) + (y_{t-k} + y_{t+k})/2 \Big]$

Chapitre 4

Eléments de Statistique Inférentielle

A l'inverse de la Statistique Descriptive, la Statistique Inférentielle s'appuie sur la théorie des probabilités pour introduire les notions de risque et de précision, en mettant en avant les distributions d'échantillonnage. Ainsi, si l'on s'intéresse à un paramètre θ inconnu d'une population statistique, on peut se donner comme objectif de l'évaluer au mieux à partir d'un échantillon, ou encore de tester des hypothèses le concernant. Par exemple, θ peut être la prévalence d'une maladie dans une population (proportion d'individus atteints par cette maladie). On distingue deux approches pour l'estimation d'un paramètre scalaire θ : l'estimation ponctuelle et l'estimation par intervalle. La première approche consiste à fournir une valeur qui semble la meilleure évaluation de θ vis-à-vis d'un certain critère d'optimalité. On parle d'*estimation ponctuelle*. Ainsi, la proportion de malades dans un échantillon aléatoire simple avec remise est la meilleure estimation de la prévalence par rapport au critère d'écart quadratique moyen. La seconde approche est de se fournir un ensemble de valeurs fortement probables pour θ. Plus précisément, on construit un intervalle tel que la probabilité qu'il contienne θ soit au moins égale à $1-\alpha$, avec $\alpha \in [0,1]$. On parle d'*estimation par intervalle*. Le nombre $1-\alpha$ est appelé niveau de sécurité de l'intervalle ainsi créé. Cet intervalle est appelé *intervalle de confiance de (niveau de) sécurité* $1-\alpha$ *pour* θ. On peut remarquer que la probabilité que θ ne soit pas dans l'intervalle est au plus α. On peut donc évoquer une dualité entre intervalle de confiance et test d'hypothèse puisque α peut être vu comme un majorant du risque de se tromper en déclarant que θ appartient à cet intervalle. Aux paragraphes suivants, les intervalles de confiance et les tests d'hypothèses concernant une proportion, une espérance, une variance dans des situations classiques sont présentés.

4.1 Intervalles de confiance

Un intervalle de confiance I_α de sécurité $1-\alpha$ pour un paramètre inconnu θ est donc un intervalle tel que la probabilité de l'événement $\ll\theta \in I_\alpha\gg$ vaille au moins $1-\alpha$. Dans ce qui suit, u_p est le fractile d'ordre p de la loi $\mathcal{N}(0,1)$, $t_{\nu,p}$ est celui de la loi de Student à ν degrés de liberté, et $\chi^2_{\nu,p}$ celui de la loi de Pearson à ν degrés de liberté. On considère n observations indépendantes de même loi.

Pour une proportion et un nombre d'observations $n \geq 30$:

$$[\ \hat{p}\ \pm\ u_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ]$$

où $\hat{p}$ est la proportion observée dans l'échantillon.

Pour une espérance :

Taille d'échantillon	Loi des observations	Ecart-type	Intervalle de confiance
$n \geq 30$	quelconque	connu	$[\bar{x}\ \pm\ u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$
$n \geq 30$	quelconque	inconnu	$[\bar{x}\ \pm\ u_{1-\frac{\alpha}{2}}\frac{\widehat{\sigma}}{\sqrt{n}}]$
$n < 30$	gaussienne	connu	$[\bar{x}\ \pm\ u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$
$n < 30$	gaussienne	inconnu	$[\bar{x}\ \pm\ t_{n-1,1-\frac{\alpha}{2}}\frac{\widehat{\sigma}}{\sqrt{n}}]$

où $\bar{x}$ et $\hat{\sigma}$ sont respectivement la moyenne et l'écart-type corrigé de la série observée.

Pour une variance (observations gaussiennes) :

$$[\ (n-1)\frac{\hat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}}; (n-1)\frac{\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}}\ ]$$

où $\hat{\sigma}^2$ est la variance corrigée de la série observée.

		Etat de la nature	
		H_0	H_1
Décision	Rejet de H_0	α	$1-\beta$
	Non rejet de H_0	$1-\alpha$	β

TABLE 4.1 – Risques décisionnels conditionnels à l'état de la nature

4.2 Tests de signification

Un test de signification est une procédure permettant de choisir parmi deux hypothèses celles la plus probable au vu des observations effectuées à partir d'un échantillon ou un dispositif expérimental. Ces deux hypothèses sont disjointes c'est-à-dire s'excluent mutuellement. Les hypothèses auxquelles on s'intéresse portent généralement sur un ou plusieurs paramètres de la population statistique étudiée : ainsi, si l'on s'intéresse à un caractère particulier, on pourra par exemple tester l'égalité de l'espérance du caractère avec une valeur de référence. Par exemple, on peut désirer tester l'égalité d'une contenance attendue de bouteilles commercialisées, avec une valeur de référence en l'occurrence la contenance indiquée sur l'étiquette commerciale. Un inspecteur de la direction de la consommation peut choisir un certain nombre de bouteilles dans la production de l'usine concernée. Sachant qu'il y a un aléa d'échantillonnage et une variabilité dans le système de remplissage des bouteilles, comment tranchera t-il entre l'hypothèse ≪la contenance attendue est égale à la contenance annoncée≫ et l'alternative contraire ?

Erreurs décisionnelles et risques : Le principe de base d'un test de signification est de considérer une hypothèse privilégiée H_0 et une alternative H_1, puis de bâtir une règle permettant de décider de rejeter ou pas H_0. Le tableau 4.1 résume les 4 situations possibles. L'erreur de première espèce est de rejeter l'hypothèse privilégiée H_0 alors qu'elle est vraie. L'erreur de seconde espèce est de ne pas rejeter H_0 alors qu'elle est fausse. α est la probabilité de rejeter à tort l'hypothèse H_0 ; α est aussi appelé risque de première espèce, ou niveau du test. β est la probabilité de ne pas rejeter H_0 alors que l'hypothèse alternative H_1 est vraie ; β est appelé risque de seconde espèce. La valeur $1-\beta$ est la puissance du test, et traduit la faculté de rejeter H_0 quand l'alternative H_1 est vraie.

Dans la pratique, α est fixé par l'expérimentateur (les valeurs les plus courantes sont 0,05 ou 0,01. On dit qu'on contrôle le risque de première espèce. Par contre, β peut être difficile à calculer. Heureusement, ce calcul n'est pas nécessaire sauf si l'on veut comparer plusieurs procédure de tests.

Dans la littérature, H_0 est aussi appelée *hypothèse nulle* ou encore *hypothèse principale*. Elle joue un rôle prédominant par rapport à l'hypothèse H_1 qui est souvent l'hypothèse alternative contraire. On cherche à contrôler le risque α de rejeter à tort H_0 en lui imposant une valeur relativement faible (au plus 0,05). Le fait d'imposer une valeur faible à α conduit à n'abandonner l'hypothèse H_0 que dans des cas qui ≪semblent sortir nettement de l'ordinaire≫ si H_0 était vraie.

Probabilité critique (ou p-value ou niveau de signification observé) : Notons bien que

plus α est choisi petit, plus la règle de décision est stricte (ou conservative) dans la mesure où elle aboutit à rejeter H_0 que dans des cas rarissimes et donc à conserver cette hypothèse quelque fois à tort. Une vision moderne, liée à l'explosion de la puisssance des ordinateurs et de processus numériques d'approximation rapides et précis, est d'afficher la p-value ou probabilité critique p_c. Par définition, **la p-value est la plus petite des valeurs de risque de première espèce pour lesquelles la décision serait de rejeter** H_0. La valeur p_c est calculée à partir des observations et de leurs propriétés distributionnelles sous H_0. Comme p_c est le plus petit niveau de signification auquel on rejette l'hypothèse H_0, il est aussi appelé *niveau de signification observé*.

Critère de test, Région critique : Tout test d'une hypothèse H_0 est basé sur un critère C qui est calculé à partir des observations effectuées. C est appelé critère de test (ou statistique de test). C est une quantité dépendant des données observées ou recueillies lors de l'expérimentation ou l'enquête. C'est donc une variable aléatoire dont la valeur observée nous permettra de déterminer quelle hypothèse est la plus plausible, en se référant à la distribution de probabilité de cette variable aléatoire sous H_0. La prise de décision se fera selon une règle dont la forme est généralement :

Rejet de H_0 si $C \in R_c(\alpha)$

Non rejet de H_0 si $C \notin R_c(\alpha)$.

où $R_c(\alpha)$ est donc l'ensemble des valeurs pour lesquelles la statistique de test conduit au rejet de l'hypothèse H_0 au niveau de signification α, et est donc appelé région critique (ou zone de rejet) du test au niveau α.

Le complémentaire de $R_c(\alpha)$ est l'ensemble des valeurs pour lesquelles la statistique de test conduit au non rejet de l'hypothèse H_0. On l'appelle région (ou zone) d'acceptation du test au niveau α.

La région critique ou zone de rejet correspond donc aux valeurs de C qui seraient trop extraordinaires sous l'hypothèse H_0 pour être considérées comme le fruit du hasard d'échantillonnage.

Par définition, la région critique de niveau α d'un test est donc l'ensemble des valeurs pour lesquelles H_0 est rejetée au niveau α.

Unilarité, bilatéralité : Le test est unilatéral si la région critique du test est un intervalle situé d'un seul côté de la distribution de probabilité de C sous H_0. Il est bilatéral si la région critique est la réunion de deux intervalles disjoints situés des deux côtés opposés de la distribution de probabilité de C sous H_0.

Certains tests comme l'analyse de la variance ou le test du khi-deux sont pratiquement toujours unilatéraux.

Problème de test : Avant d'utiliser tout test statistique, il s'agit de bien définir le *problème posé*. Ceci consiste à formuler de façon précise les hypothèses H_0 et H_1 en faisant apparaître le(s)

paramètre(s) concerné(s), et éventuellement les valeurs de référence. En effet, selon les hypothèse formulées, on applique soit un test bilatéral, soit un test unilatéral.

Si l'on revient à l'exemple de l'inspecteur de la direction de la consommation qui s'intéresse à la contenance attendue m des bouteilles produites dans un atelier industriel, l'hypothèse privilégiée peut être que m vaut bien la contenance nominale m_0 indiquée sur l'étiquette contre l'hypothèse de fraude selon laquelle m est strictement inférieure à m_0. On a donc le problème de test :

$H_0 : m = m_0$ contre $H_1 : m < m_0$.

Robustesse d'un test : La majorité des tests statistiques repose sur le respect d'un certain nombre de conditions. On parle parfois de *suppositions du test* pour les distinguer des hypothèses que l'on veut tester. Selon le degré de respect de ces conditions d'application, la validité des résultats se trouve plus ou moins affectée et elle l'est d'autant plus que le test est moins robuste. Ainsi, la robustesse d'un test équivaut à sa tolérance vis-à-vis du respect des conditions (ou suppositions) du test. Un test est dit robuste si le non respect des conditions d'application n'affecte que très peu la validité des résultats.

Puissance d'un test : C'est la faculté de rejeter l'hypothèse priviliégiée quand elle est fausse, et l'aptitude à détecter les écarts entre l'hypothèse privilégiée et les hypothèses alternatives. Les tests peu puissants augmentent la probabilité de commettre une erreur de deuxième espèce. Or, cette erreur peut s'avérer particulièrement grave. Une étude visant à tester l'hypothèse selon laquelle un seuil d'alerte épidémiologique est atteint dans une région, et qui classerait à tort cette région dans la catégorie de celles n'ayant pas dépassé ce seuil, peut engendrer une catastrophe sanitaire.

Choix d'un test : Plusieurs tests de conception très différente peuvent être disponibles pour vérifier une hypothèse privilégiée. Dans un tel cas, le statisticien choisira le test **le plus puissant et le plus robuste** (si tenté qu'on le connaisse). .

Construction d'un test : Pour construire un test de niveau α, on peut procéder selon les étapes suivantes :

1. Définir le problème posé en formulant de façon précise les hypothèses H_0 et H_1.
2. Déterminer un critère pertinent par rapport au problème posé (en fait une statistique de test C) de loi de probabilité connue sous l'hypothèse H_0.
3. Détermination de la région critique $R_c(\alpha)$ en fonction de la loi de probabilité de C et du risque de première espèce α.

Après recueil des données, appliquer le test statistique de niveau α consistera donc à :

1. Calculer la valeur C_{obs} de la statistique de test à partir des observations

2. Vérifier si C_{obs} appartient ou non à $R_c(\alpha)$
3. Conclure au rejet ou non de l'hypothèse H_0 au niveau α.

Courbe caractéristique d'efficacité : Considérons des hypothèses paramétrées à l'aide d'un seul paramètre scalaire θ. On a une famille d'hypothèses $\{H_\theta, \theta \in \Theta\}$ où Θ est un ensemble de valeurs possibles pour θ. Si l'on note $\beta(\theta)$ le risque de deuxième espèce pour l'hypothèse alternative H_θ, la courbe indiquant la puissance du test en fonction du paramètre θ est appelée courbe caractéristique d'efficacité du test. C'est la courbe représentative de la fonction puissance du test : $\theta \mapsto 1 - \beta(\theta)$ qui indique donc la probabilité de rejeter l'hypothèse privillégiée quand c'est l'alternative H_θ qui est vraie.

La figure 4.1 montre des courbes d'efficacité selon le nombre d'observations effectués pour le problème de test d'égalité d'une valeur attendue m à la valeur de référence 25. Par exemple, Il peut s'agit de vérifier si, dans un atelier de remplissage, l'unique machine est bien réglée sur la contenance attendue 25cl, l'hypothèse alternative étant qu'elle est déréglée. Le problème de test peut s'exprimer ainsi : $\ll H_0 : m = m_0$ contre $m \neq m_0 \gg$. Pour l'illustration dans cette figure, le niveau du test est fixé à $\alpha = 0,05$ et l'écart-type est supposé connu et égal à 2cl. On voit que le test est d'autant plus efficace que le nombre d'observations est élevé, et que la puissance augmente avec l'écart entre la contenance attendue m et la contenance nominale m_0.

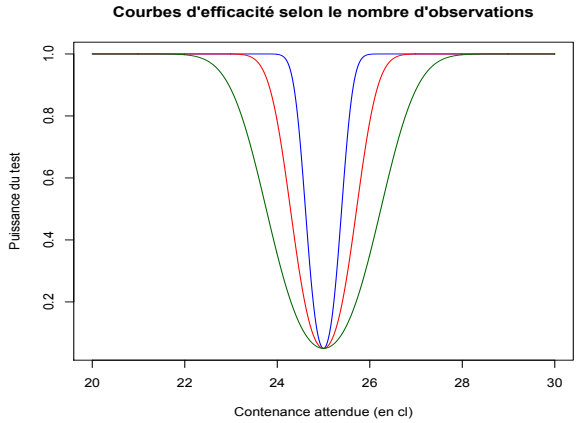

FIGURE 4.1 – Test de confirmité d'une espérance à 25cl de niveau 5% ; les observations sont gaussiennes d'écart-type σ=2cl ; la taille d'échantillon n vaut 10 (vert), 30 (rouge) et 100 (bleu)

Les principaux éléments d'identification d'un test de signification sont donc : le problème de test, la statistique de test et la règle de décision. Les propriétés primordiales d'un test sont sa puissance et sa robustesse.

Echantillons indépendants : Pour comparer les moyennes, les variances ou les autres pa-

ramètres estimés de deux échantillons, il faut prendre en considération la technique conduisant à la constitution des deux échantillons. Si la sélection des éléments est aléatoire, et si le choix des éléments du premier échantillon n'a aucune influence sur le choix des éléments du second, alors les deux échantillons sont dits indépendants.

Comparaison d'une proportion à une valeur de référence p_0 :

On considère un nombre d'observations $n \geq 30$.

Problème de test : $p = p_0$ contre $p \neq p_0$

Rejet de l'hypothèse $H_0 : p = p_0$ si $\dfrac{\mid \hat{p} - p_0 \mid}{\sqrt{\frac{p_0(1-p_0)}{n}}} > u_{1-\frac{\alpha}{2}}$

Non rejet de l'hypothèse H_0 sinon.

Problème de test : $p \geq p_0$ contre $p < p_0$

Rejet de l'hypothèse $H_0 : p \geq p_0$ si $\dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -u_{1-\alpha}$

Non rejet de l'hypothèse H_0 sinon.

Problème de test : $p \leq p_0$ contre $p > p_0$

Rejet de l'hypothèse $H_0 : p \leq p_0$ si $\dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > u_{1-\alpha}$

Non rejet de l'hypothèse H_0 sinon.

Comparaison d'une espérance à une valeur de référence μ_0 :

Cas 1 : Les observations suivent la loi gaussienne

$\mu = \mu_0$ contre $\mu \neq \mu_0$

Rejet de l'hypothèse $H_0 : \mu = \mu_0$ si $\dfrac{\mid \overline{x} - \mu_0 \mid}{\hat{\sigma}} \sqrt{n} > t_{1-\frac{\alpha}{2}}$

Non rejet de l'hypothèse H_0 sinon.

$\mu \geq \mu_0$ contre $\mu < \mu_0$

Rejet de l'hypothèse $H_0 : \mu \geq \mu_0$ si $\dfrac{\overline{x} - \mu_0}{\hat{\sigma}} \sqrt{n} < -t_{1-\alpha}$

Non rejet de l'hypothèse H_0 sinon.

$\mu \leq \mu_0$ contre $\mu > \mu_0$

Rejet de l'hypothèse $H_0 : \mu \leq \mu_0$ si $\quad \dfrac{\overline{x} - \mu_0}{\hat{\sigma}}\sqrt{n} > t_{1-\alpha}$

Non rejet de l'hypothèse H_0 sinon.

Cas 2 : Les observations suivent une loi quelconque (il est nécessaire alors que $n > 30$)

On remplace dans les règles ci-dessus le fractile $t_{1-\frac{\alpha}{2}}$ (respectivement $t_{1-\alpha}$) de la loi de Student par le fractile $u_{1-\frac{\alpha}{2}}$ (respectivement $u_{1-\alpha}$) de la loi Normale centrée réduite.

Test d'égalité de deux proportions :

On considère que n_1 et n_2 sont supérieurs à 30. On pose $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$.

$p_1 = p_2$ contre $p_1 \neq p_2$

Rejet de l'hypothèse $H_0 : p_1 = p_2$ si $\quad \dfrac{\mid \hat{p}_1 - \hat{p}_2 \mid}{\sqrt{\hat{p}(1-\hat{p})\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}} > u_{1-\frac{\alpha}{2}}$

Non rejet de H_0 sinon.

Test d'égalité de deux espérances :

On observe deux séries statistiques indépendantes de longueurs respectives n_1 et n_2 de moyennes respectives $\overline{x}_1$ et $\overline{x}_2$, et de variances corrigées respectives $\hat{\sigma}_1^2$ et $\hat{\sigma}_1^2$.

On considère que n_1 et n_2 sont supérieurs à 30.

$\mu_1 = \mu_2$ contre $\mu_1 \neq \mu_2$

Rejet de l'hypothèse $H_0 : \mu_1 = \mu_2$ si $\quad \dfrac{\mid \overline{x}_1 - \overline{x}_2 \mid}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}} > u_{1-\frac{\alpha}{2}}$

Non rejet de H_0 sinon.

Tests concernant une variance (observations gaussiennes) :

$\sigma^2 = \sigma_0^2$ contre $\sigma^2 \neq \sigma_0^2$

Rejet de l'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ si $\quad \dfrac{1}{\sigma_0^2}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2 \notin [\chi^2_{\frac{\alpha}{2}}\ ,\ \chi^2_{1-\frac{\alpha}{2}}]$

Non rejet de l'hypothèse H_0 sinon.

$\sigma^2 \geq \sigma_0^2$ contre $\sigma^2 < \sigma_0^2$

Rejet de l'hypothèse $H_0 : \sigma^2 \geq \sigma_0^2$ si $\quad \dfrac{1}{\sigma_0^2} \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2 \; < \; \chi^2_{\alpha}$

Non rejet de l'hypothèse H_0 sinon.

$\sigma^2 \leq \sigma_0^2$ contre $\sigma^2 > \sigma_0^2$

Rejet de l'hypothèse $H_0 : \sigma^2 \geq \sigma_0^2$ si $\quad \dfrac{1}{\sigma_0^2} \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2 \; > \; \chi^2_{1-\alpha}$

Non rejet de l'hypothèse H_0 sinon.

Test d'ajustement du χ^2 :

Rejet du modèle considéré si $\quad \displaystyle\sum_{i=1}^{I} \frac{(n_i - np_i)^2}{np_i} \; > \; \chi^2_{1-\alpha}$

Non rejet du modèle sinon.

4.3 Inférence pour la régression linéaire simple

Tests concernant les paramètres de régression :

Soient $\hat{a}$ et $\hat{b}$ les estimateurs respectifs de la pente a de la droite de régression et de l'interception b ; soient $\hat{\sigma}^2$ l'estimateur sans biais de σ^2, et $t_{n-2,1-\alpha/2}$ le fractile de la loi de Student à $n-2$ degrés de liberté.

$a = a_0$ contre $a \neq a_0$

Rejet de l'hypothèse $H_0 : a = a_0$ si $$\frac{\mid \hat{a} - a_0 \mid}{\sqrt{\frac{\hat{\sigma}^2}{n s_x^2}}} > t_{n-2,1-\alpha/2}$$

Acceptation de l'hypothèse H_0 sinon.

$b = b_0$ contre $b \neq b_0$

Rejet de l'hypothèse $H_0 : b = b_0$ si $$\frac{\mid \hat{b} - b_0 \mid}{\sqrt{\frac{\hat{\sigma}^2}{n}\Big(1 + \frac{\overline{x}^2}{s_x^2}\Big)}} > t_{n-2,1-\alpha/2}$$

Acceptation de l'hypothèse H_0 sinon.

Intervalles de confiance de sécurité $1 - \alpha$:

Pour a, on a
$$[\hat{a} \pm t_{n-2,1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n s_x^2}}].$$

Pour b, on a
$$[\hat{b} \pm t_{n-2,1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}\Big(1 + \frac{\overline{x}^2}{s_x^2}\Big)}].$$

Pour la prédiction associée à une valeur x_0 de la variable explicative, on a
$$[\hat{y}_0 \pm t_{n-2,1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}\Big(1 + \frac{(\overline{x} - x_0)^2}{s_x^2}\Big)}].$$

On rappelle que $\hat{\sigma}^2 = \dfrac{n}{n-2}(s_y^2 - \dfrac{s_{xy}^2}{s_x^2})$ où s_{xy} est la covariance empirique entre les x_i et les y_i.

4.4 Tests concernant l'indice de dispersion I_d

Cas où $n-1 \leq 30$. On utilise la statistique $T = (n-1)\widehat{I}_d$

$\underline{H_0 : I_d = 1 \text{ contre } I_d \neq 1}$

Rejet de H_0 si $T \notin [\chi^2_{n-1,\alpha/2}; \chi^2_{n-1,1-\alpha/2}]$

Non rejet de H_0 si $T \in [\chi^2_{n-1,\alpha/2}; \chi^2_{n-1,1-\alpha/2}]$

$\underline{H_0 : I_d = 1 \text{ contre } I_d > 1}$

Rejet de H_0 si $T > \chi^2_{n-1,1-\alpha}$

Non rejet de H_0 si $T \leq \chi^2_{n-1,1-\alpha}$

$\underline{H_0 : I_d = 1 \text{ contre } I_d < 1}$

Rejet de H_0 si $T < \chi^2_{n-1,\alpha}$

Non rejet de H_0 si $T \geq \chi^2_{n-1,\alpha}$

Cas où $n-1 > 30$. On utilise la statistique $U = \sqrt{2(n-1)\widehat{I}_d} - \sqrt{2(n-1)-1}$

$\underline{H_0 : I_d = 1 \text{ contre } I_d \neq 1}$

Rejet de H_0 si $|\,U\,| > u_{1-\alpha/2}$

Non rejet de H_0 si $|\,U\,| \leq u_{1-\alpha/2}$

$\underline{H_0 : I_d = 1 \text{ contre } I_d > 1}$

Rejet de H_0 si $U > u_{1-\alpha}$

Non rejet de H_0 si $U \leq u_{1-\alpha}$

$\underline{H_0 : I_d = 1 \text{ contre } I_d < 1}$

Rejet de H_0 si $U < -u_{1-\alpha}$

Non rejet de H_0 si $U \geq -u_{1-\alpha}$

Chapitre 5

Eléments de planification expérimentale

5.1 Vocabulaire de base des plans d'expérience

Voici de façon non exhausitive les termes techniques souvent utilisés en planification expérimentale.

Dispositif expérimental (= Plan d'expérience) : c'est la répartition des traitements sur les unités expérimentales.

Expérience : Epreuve, essai effectué pour étudier un phénomène.

Expérimentation : action d'expérimenter ; essai d'application, expérience.

Expérimenter : soumettre à des expériences

Facteur ou facteur de variation : caractère, variable concourant au résultat observé sur une unité expérimentale. Il est supposé influencer la variable principale (ou réponse). Il peut être qualitatif ou quantitatif.

Facteur contrôlé, facteur non contrôlé : dans certaines situations, le facteur peut être contrôlé c'est-à-dire qu'on peut imposer la modalité ou la valeur possible de ce facteur sur chaque unité expérimentale. Autrement, le facteur est dit non contrôlé.

Matériel expérimental : on désigne ici l'ensemble des unités expérimentales (et non les instruments techniques ou l'outillage permettant de réaliser l'expérience).

Niveau de facteur : modalité (en cas de facteur qualitatif) ou valeur possible (en cas de facteur

quantitatif) du facteur. Les différents niveaux d'un même facteur s'excluent donc mutuellement.

Objectif scientifique : formulation précise du problème scientifique résoudre. Il s'agit de préciser le but à atteindre relativement à la science ou une science. La définition claire et précise du but de l'expérience est toujours un élément primordial du protocole expérimental.

Plan d'expérience (=Dispositif expérimental) : c'est la répartition des traitements sur les unités expérimentales.

Planification expérimentale : mise en oeuvre, organisation du plan d'expérience.

Protocole expérimental : description détaillée d'une expérience permettant de la réaliser en totalité ou de la reproduire de manière identique sans indication supplémentaire. Il comprend : le dispositif expérimental ou plan d'expérience, la description complète du matériel expérimental, de la préparation et de la conduite de l'expérience, les différents protocoles d'observation et d'analyse statistique des résultats.

Répétitions : affectation d'un même traitement à plusieurs unités expérimentales. Les répétitions ont pour but, non seulement d'augmenter la précision des résultats, mais également de rendre possible l'estimation de cette précision par exemple sous forme de variances, écarts-types ou coefficients de variation.

Réponse (ou variable principale) : variable caractérisant le comportement du phénomène étudié.

Traitement : toute combinaison de niveaux des facteurs étudiés. S'il n'y a qu'un facteur considéré dans l'expérience, un traitement correspond à un niveau de facteur.

Unité expérimentale : unité de base de l'expérience. Elle est traitée individuellement dès le départ et fait l'objet d'au moins une observation.

Variable principale (ou réponse) : variable caractérisant le comportement du phénomène étudié.

Illustration 1 : Soit un essai de fertilisation azotée sur canne sucre avec deux facteurs étudiés : la fumure et la variété de canne sucre. Chacun des facteurs est étudié à plusieurs niveaux : 3 niveaux pour la fumure azotée : 40 ; 60 ; 80 u/ha et 4 niveaux pour les variétés de canne sucre : M1394/86, M1400/86, M3035/66 et la R579. Si on combine tous les niveaux des facteurs fumure et variété, on aura 12 traitements à affecter à des unités expérimentales, en l'occurrence des parcelles d'un domaine d'expérimentation agronomique. Pour cela, il nous faut disposer d'au moins 12 parcelles, sachant en outre qu'il y a nécessité de répétitions pour prendre en compte la variabilité.

Il est à noter que de nombreuses considérations pourraient être prises en compte dans l'élaboration du protocole expérimental. Citons quelques éléments de discussion concernant le facteur variété : une variété est mieux adaptée aux zones humides et peut être récoltée dès la mi-août ; une autre est mieux adaptée aux régions de moyenne et basse pluviométrie et est

particulièrement recommandée pour l'exploitation industrielle dans les sols francs de moyenne altitude. Une autre variété est issue d'un croisement entre deux variétés ayant la particularité de mieux s'adapter dans les sols rocheux ; certaines variétés ont des rendements qui se sont avérés supérieurs des variétés de référence en milieu et en fin de campagne. La sensibilité de ces variétés aux maladies qui sévissent dans la région de référence n'est pas connue. Le potentiel à l'échelle industrielle de certaines variétés reste à confirmer.

Illustration 2 : Une expérience vise à étudier l'effet de cinq doses distinctes d'un même produit sur le gain de poids journaliers chez des porcins d'un même élevage. On envisage d'appliquer chacune des cinq doses respectivement à cinq groupes d'animaux de même effectif. Cependant d'autres sources de variation sont connues (age, espèce, ascendance). Un dispositif permettant un contrôle de l'hétérogénéité est envisagé.

5.2 Plan en randomisation totale - Completely randomized design

Dans les paragraphes suivants, le nombre total d'unités expérimentales utilisées dans le plan est noté n ; la moyenne et la variance de la série statistique des n observations sont notées respectivement $\bar{x}$ et s^2.

5.2.1 Plan à un facteur avec répétitions

On étudie un facteur A ayant I niveaux. On a n_i répétitions pour le niveau i.

Données : x_{ij} est la valeur observée pour la $j$ème répétition du niveau i de A ; $\bar{x}_{i.}$ est la moyenne observée pour le niveau i de A.

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F de Fisher
Totale	SCT	$n-1$	CMT=$\dfrac{SCT}{n-1}$	
Facteur A	SCF_A	$I-1$	$CMF_A = \dfrac{SCF_A}{I-1}$	$F_A = \dfrac{CMF_A}{CMR}$
Résiduelle	SCR	$n-I$	$CMR = \dfrac{SCR}{n-I}$	

avec

$$SCT = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2 = ns^2;\ SCF_A = \sum_{i=1}^{I} n_i(\bar{x}_{i.}-\bar{x})^2;\ SCR = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2$$

Equation d'ANOVA : $SCT = SCF_A + SCR$

Modèle sous-jacent : $x_{ij} = \mu + \alpha_i + e_{ij}$ où les e_{ij} sont i.i.d. selon $\mathcal{N}(0,\sigma^2)$.

5.2.2 Plan à deux facteurs sans répétitions

On étudie deux facteurs A et B ayant respectivement I et J niveaux.

Données : x_{ij} est la valeur observée pour le traitement (i, j) ; $\bar{x}_{i.}$ est la moyenne observée pour le niveau i de A ; $\bar{x}_{.j}$ est la moyenne observée pour le niveau j de B.

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F de Fisher
Totale	SCT	$IJ-1$	$CMT = \dfrac{SCT}{n-1}$	
Facteur A	SCF_A	$I-1$	$CMF_A = \dfrac{SCF_A}{I-1}$	$F_A = \dfrac{CMF_A}{CMR}$
Facteur B	SCF_B	$J-1$	$CMF_B = \dfrac{SCF_B}{J-1}$	$F_B = \dfrac{CMF_B}{CMR}$
Résiduelle	SCR	$(I-1)(J-1)$	$CMR = \dfrac{SCR}{(I-1)(J-1)}$	

avec

$$SCT = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \bar{x})^2; \quad SCF_A = \sum_{i=1}^{I} J(\bar{x}_{i.} - \bar{x})^2; \quad SCF_B = \sum_{j=1}^{J} I(\bar{x}_{.j} - \bar{x})^2$$

et

$$SCR = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$$

Equation d'ANOVA : $SCT = SCF_A + SCF_B + SCR$

Modèle sous-jacent : $x_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$ où les e_{ij} sont i.i.d. selon $\mathcal{N}(0, \sigma^2)$.

5.2.3 Plan à deux facteurs avec répétitions

On étudie deux facteurs A et B ayant respectivement I et J niveaux.

Données : x_{ijk} est la valeur observée pour la $k$ème répétition du traitement (i,j) ; $\bar{x}_{i..}$ est la moyenne observée pour le niveau i de A ; $\bar{x}_{.j.}$ celle pour le niveau j de B ; $\bar{x}_{ij.}$ celle pour le traitement (i,j) ; n_{ij} est le nombre de répétitions du traitement (i,j) ; n_{i+} est le nombre de répétitions du niveau i de A ; n_{+j} est celui du niveau j de B.

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F de Fisher
Totale	SCT	$n-1$	$CMT = \dfrac{SCT}{n-1}$	
Facteur A	SCF_A	$I-1$	$CMF_A = \dfrac{SCF_A}{I-1}$	$F_A = \dfrac{CMF_A}{CMR}$
Facteur B	SCF_B	$J-1$	$CMF_B = \dfrac{SCF_B}{J-1}$	$F_B = \dfrac{CMF_B}{CMR}$
Interaction (A,B)	SCF_{AB}	$(I-1)(J-1)$	$CMF_{AB} = \dfrac{SCF_{AB}}{(I-1)(J-1)}$	$F_{AB} = \dfrac{CMF_{AB}}{CMR}$
Résiduelle	SCR	$n-IJ$	$CMR = \dfrac{SCR}{n-IJ}$	

avec $SCT = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk}-\bar{x})^2$; $SCF_A = \sum_{i=1}^{I} n_{i+}(\bar{x}_{i..}-\bar{x})^2$; $SCF_B = \sum_{j=1}^{J} n_{+j}(\bar{x}_{.j.}-\bar{x})^2$

$$SCF_{AB} = \sum_{i=1}^{I}\sum_{i=1}^{J} n_{ij}(\bar{x}_{ij.}-\bar{x}_{i..}-\bar{x}_{.j.}+\bar{x})^2 \text{ et } SCR = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk}-\bar{x}_{ij.})^2$$

Equation d'ANOVA : $SCT = SCF_A + SCF_B + SCF_{AB} + SCR$

Modèle sous-jacent : $x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ où les e_{ijk} sont i.i.d. selon $\mathcal{N}(0,\sigma^2)$.

5.3 Plan en blocs randomisés - Randomized block design

5.3.1 Plan à un facteur en blocs complets sans répétition

Données : b est le nombre de blocs, I est le nombre de niveaux du facteur A ; puisqu'il n'y a pas de répétitions dans les blocs, le nombre total d'unités expérimentales est $n = Ib$; x_{ij} est la valeur observée pour le niveau i de A et le bloc j ; $\bar{x}_{i.}$ est la moyenne observée pour le niveau i de A ; $\bar{x}_{.j}$ est la moyenne observée pour le bloc j.

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F de Fisher
Totale	SCT	$Ib-1$	CMT=$\dfrac{SCT}{Ib-1}$	
Facteur A	SCF_A	$I-1$	$CMF_A = \dfrac{SCF_A}{I-1}$	$F_A = \dfrac{CMF_A}{CMR}$
Facteur bloc	SCF_{bloc}	$b-1$	$CMF_{\text{bloc}} = \dfrac{SCF_{\text{bloc}}}{b-1}$	$F_{\text{bloc}} = \dfrac{CMF_{\text{bloc}}}{CMR}$
Résiduelle	SCR	$(I-1)(b-1)$	$CMR = \dfrac{SCR}{(I-1)(b-1)}$	

$$\text{avec } SCT = \sum_{i=1}^{I}\sum_{j=1}^{b}(x_{ij}-\bar{x})^2; SCF_A = \sum_{i=1}^{I} b(\bar{x}_{i.}-\bar{x})^2 ; SCF_{\text{bloc}} = \sum_{j=1}^{b} I(\bar{x}_{.j}-\bar{x})^2$$

Les sommes de carrés sont calculées comme dans le cas du plan à deux facteurs sans répétitions (voir I.2.) en considérant l'effet bloc comme second facteur.

Equation d'ANOVA : $SCT = SCF_A + SCF_{\text{bloc}} + SCR$

Modèle sous-jacent : $x_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$ où les e_{ij} sont i.i.d. selon $\mathcal{N}(0, \sigma^2)$.

5.3.2 Plan à un facteur en blocs complets équilibrés

On étudie un facteur A ayant I niveaux. Les b blocs étant complets équilibrés, on note K le nombre de répétitions d'un niveau donné de A à l'intérieur d'un bloc.

Données : x_{ijk} est la valeur observée pour la $k^{\text{ème}}$ répétition du niveau i de A dans le bloc j ; $\bar{x}_{i..}$ est la moyenne observée pour le niveau i de A ; $\bar{x}_{.j.}$ est celle pour le bloc j ; $\bar{x}_{ij.}$ est celle pour le niveau i dans le bloc j.

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F de Fisher
Totale	SCT	$IbK-1$	$CMT = \dfrac{SCT}{IbK-1}$	
Facteur A	SCF_A	$I-1$	$CMF_A = \dfrac{SCF_A}{I-1}$	$F_A = \dfrac{CMF_A}{CMR}$
Bloc	SCF_{bl}	$b-1$	$CMF_{\text{bl}} = \dfrac{SCF_{\text{bl}}}{b-1}$	$F_{\text{bl}} = \dfrac{CMF_{\text{bl}}}{CMR}$
Interaction (A,Bloc)	$SCF_{A,\text{bl}}$	$(I-1)(b-1)$	$CMF_{A,\text{bl}} = \dfrac{SCF_{A,\text{bl}}}{(I-1)(b-1)}$	$F_{A,\text{bl}} = \dfrac{CMF_{A,\text{bl}}}{CMR}$
Résiduelle	SCR	$Ib(K-1)$	$CMR = \dfrac{SCR}{Ib(K-1)}$	

avec $SCT = \sum_{i=1}^{I}\sum_{j=1}^{b}\sum_{k=1}^{K}(x_{ijk}-\bar{x})^2$; $SCF_A = \sum_{i=1}^{I} bK(\bar{x}_{i..}-\bar{x})^2$; $SCF_{\text{bl}} = \sum_{j=1}^{b} IK(\bar{x}_{.j.}-\bar{x})^2$;

$SCF_{A,\text{bl}} = \sum_{i=1}^{I}\sum_{i=1}^{b} K(\bar{x}_{ij.}-\bar{x}_{i..}-\bar{x}_{.j.}+\bar{x})^2$ et $SCR = \sum_{i=1}^{I}\sum_{j=1}^{b}\sum_{k=1}^{K}(x_{ijk}-\bar{x}_{ij.})^2$

Equation d'ANOVA : $SCT = SCF_A + SCF_{\text{bl}} + SCF_{A,\text{bl}} + SCR$

Modèle sous-jacent : $x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ où les e_{ijk} sont i.i.d. selon $\mathcal{N}(0,\, \sigma^2)$.

5.4 Plan avec plusieurs contrôles d'hétérogénéité - design with several controls of heterogeneity

5.4.1 Carré latin t x t

On étudie un facteur A qui a t niveaux ($t > 2$). On a deux facteurs de contrôle à t niveaux chacun. Le premier facteur de contrôle est appelé *ligne*, et le second, *colonne*.

Données : x_{ijk} est la valeur observée pour le niveau i de A, la ligne j et la colonne k du carré ; $\bar{x}_{i..}$ est la moyenne observée pour le niveau i de A ; $\bar{x}_{.j.}$ est la moyenne observée pour la ligne j ; $\bar{x}_{..k}$ est celle pour la colonne k.

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F de Fisher
Totale	SCT	t^2-1	$CMT = \dfrac{SCT}{t^2-1}$	
Facteur A	SCF_A	$t-1$	$CMF_A = \dfrac{SCF_A}{t-1}$	$F_A = \dfrac{CMF_A}{CMR}$
Facteur ligne	SCF_L	$t-1$	$CMF_L = \dfrac{SCF_L}{t-1}$	$F_L = \dfrac{CMF_L}{CMR}$
Facteur colonne	SCF_C	$t-1$	$CMF_C = \dfrac{SCF_C}{t-1}$	$F_C = \dfrac{CMF_C}{CMR}$
Résiduelle	SCR	$(t-1)(t-2)$	$CMR = \dfrac{SCR}{(t-1)(t-2)}$	

avec $SCT = t^2 s^2$; $SCF_A = \sum_{i=1}^{t} t(\bar{x}_{i..} - \bar{x})^2$ $SCF_L = \sum_{j=1}^{t} t(\bar{x}_{.j.} - \bar{x})^2$

$SCF_C = \sum_{k=1}^{t} t(\bar{x}_{..k} - \bar{x})^2$ et $SCR = \sum_{i=1}^{t}\sum_{j=1}^{t}\sum_{k=1}^{t}(x_{ijk} - \bar{x}_{i..} - \bar{x}_{.j.} - \bar{x}_{..k} + 2\bar{x})^2$

Equation d'ANOVA : $SCT = SCF_A + SCF_L + SCF_C + SCR$

Modèle sous-jacent : $x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}$ où les e_{ijk} sont i.i.d. selon $\mathcal{N}(0, \sigma^2)$.

5.4.2 Carré gréco-latin t x t

Deux facteurs A et B sont étudiés, chacun ayant t niveaux ($t > 3$). On a en outre deux facteurs de contrôle à t niveaux chacun. Le premier facteur de contrôle est appelé *ligne*, et le second *colonne*.

Données : x_{ijkl} est la valeur observée pour le niveau i de A, le niveau j de B, la ligne k et la colonne l du carré ; $\bar{x}_{i...}$ est la moyenne observée pour le niveau i de A ; $\bar{x}_{.j..}$ est celle pour le niveau j de B ; $\bar{x}_{..k.}$ est celle pour la ligne k ; $\bar{x}_{...l}$ est celle pour la colonne l.

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	F de Fisher
Totale	SCT	$t^2 - 1$	$CMT = \dfrac{SCT}{t^2-1}$	
Facteur A	SCF_A	$t - 1$	$CMF_A = \dfrac{SCF_A}{t-1}$	$F_A = \dfrac{CMF_A}{CMR}$
Facteur B	SCF_B	$t - 1$	$CMF_B = \dfrac{SCF_B}{t-1}$	$F_B = \dfrac{CMF_B}{CMR}$
Facteur ligne	SCF_L	$t - 1$	$CMF_L = \dfrac{SCF_L}{t-1}$	$F_L = \dfrac{CMF_L}{CMR}$
Facteur colonne	SCF_C	$t - 1$	$CMF_C = \dfrac{SCF_C}{t-1}$	$F_C = \dfrac{CMF_C}{CMR}$
Résiduelle	SCR	$(t-1)(t-3)$	$CMR = \dfrac{SCR}{(t-1)(t-3)}$	

avec $SCT = t^2 s^2$, $SCF_A = \sum_{i=1}^{t} t(\bar{x}_{i...} - \bar{x})^2$, $SCF_B = \sum_{j=1}^{t} t(\bar{x}_{.j..} - \bar{x})^2$, $SCF_L = \sum_{j=1}^{t} t(\bar{x}_{..k.} - \bar{x})^2$

$$SCF_C = \sum_{k=1}^{t} t(\bar{x}_{...l} - \bar{x})^2,\ SCR = \sum_{i=1}^{t}\sum_{j=1}^{t}\sum_{k=1}^{t}\sum_{l=1}^{t}(x_{ijkl} - \bar{x}_{i...} - \bar{x}_{.j..} - \bar{x}_{..k.} - \bar{x}_{...l} + 3\bar{x})^2$$

Equation d'ANOVA : $SCT = SCF_A + SCF_B + SCF_L + SCF_C + SCR$

Modèle sous-jacent : $x_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + e_{ijkl}$ où les e_{ijkl} sont i.i.d. selon $\mathcal{N}(0, \sigma^2)$.

IV. Plan en parcelles divisées, plan en bandes croisées - Split-plot design, criss-cross design

Les ANOVAs associées à ces plans ne sont pas au programme.

5.4.3 Split-plot

Ce plan est utilisé quand les plus petites unités auxquelles on peut affecter un niveau de facteur sont de taille bien plus élevée que la taille requise pour l'application des niveaux des autres facteurs. Ainsi, en expérimentation agronomique, le facteur pratique culturale ne peut être appliqué qu' à une échelle suffisamment grande par rapport au facteur variété ou au facteur espacement des rangées de culture.

Un autre exemple concerne l'étude de la production laitière par rapport aux facteurs pâturage et méthode de traite. L'unité expérimentale idéale pour le facteur méthode de traite est la vache mais pour le facteur pâturage, il ne peut s'agir que d'un groupe d'animaux.

Le split-plot est un plan en blocs équilibrés qui consiste à subdiviser les blocs en sous-blocs et à affecter à toutes les unités expérimentales d'un même sous-bloc le même niveau de l'un des facteurs étudiés. Dans le cas d'un split-plot à deux facteurs étudiés A et B, on procède de la façon suivante : dans chaque bloc, les niveaux du facteur A sont attribués par tirages aléatoires aux sous-blocs (randomisation intra-bloc), puis dans chaque sous-bloc, l'attribution des niveaux du facteur B se fait par tirages aléatoires (randomisation intra-sous-bloc). B est appelé facteur split-plot.

Exemple : On dispose de 48 parcelles expérimentales en 2 blocs ayant chacun 4 sous-blocs. Chaque sous-bloc contient 6 parcelles. On attribue aléatoirement un des 4 niveaux c_1, c_2, c_3 et c_4 du facteur *Procédé cultural* à chaque sous-bloc d'un bloc. A l'intérieur d'un sous-bloc, on attribue aléatoirement un des 6 niveaux du facteur *Variété* à chaque parcelle.

Bloc 1

c_3v_2	c_3v_1	c_1v_5	c_1v_6	c_4v_1	c_4v_4	c_2v_3	c_2v_4
c_3v_5	c_3v_3	c_1v_2	c_1v_1	c_4v_5	c_4v_6	c_2v_1	c_2v_5
c_3v_4	c_3v_6	c_1v_3	c_1v_4	c_4v_2	c_4v_3	c_2v_6	c_2v_2

Bloc 2

c_2v_3	c_2v_5	c_3v_6	c_3v_3	c_1v_6	c_1v_4	c_4v_4	c_4v_5
c_2v_2	c_2v_4	c_3v_4	c_3v_1	c_1v_3	c_1v_2	c_4v_3	c_4v_2
c_2v_6	c_2v_1	c_3v_2	c_3v_5	c_1v_1	c_1v_5	c_4v_1	c_4v_6

5.4.4 Criss-cross

Le plan criss-cross est un plan utilisé essentiellement en expérimentation agronomique et appliqué cause de certaines contraintes techniques. C'est un plan à deux facteurs étudiés qui sont appliqués en bandes perpendiculaires.

Exemple : On désire étudier l'influence sur le rendement du facteur *Technique de travail du sol* (3 niveaux a_1, a_2 et a_3) et du facteur *Irrigation* (deux niveaux b_1=irrigué ; b_2=non irrigué). Le seul plan que l'on puisse pratiquement mettre en place est :

	a_1	a_2	a_3
	$\downarrow$	$\downarrow$	$\downarrow$
$b_1 \rightarrow$	a_1b_1	a_2b_1	a_3b_1
$b_2 \rightarrow$	a_1b_2	a_2b_2	a_3b_2

Le criss-cross est un plan relativement peu utilisé : il faut vraiment que, pratiquement, l'expérimentateur ne puisse disposer ses parcelles d'une autre façon.

5.5 Test d'hypothèse concernant l'influence de facteur

A partir du tableau d'ANOVA, on peut tester l'influence des différents facteurs étudiés ou de leurs interactions éventuelles. En effet, chaque statistique F de Fisher, obtenue dans la dernière colonne de ce tableau, correspond à une source de variation précisée dans la ligne concernée en première colonne. L'hypothèse privilégiée H_0 que l'on veut tester est alors

H_0 : "la source de variation est sans effet" contre l'hypothèse alternative

H_1 : "Il y a un effet dû à la source de variation"

L'hypothèse H_0 peut être formulée également de la façon suivante :

"les valeurs espérées (moyennes théoriques) sont identiques pour la source de variation".

Règle de décison au niveau α :

Rejet de H_0 si $F > f_{\nu_1;\nu_2;1-\alpha}$

Non rejet de H_0 si $F \leq f_{\nu_1;\nu_2;1-\alpha}$

où $f_{\nu_1;\nu_2;1-\alpha}$ est un fractile de la loi de Fisher-Snedecor à ν_1 et ν_2 degrés de liberté,

ν_1 est le nombre de degrés de liberté associé à la source de variation testée,

ν_2 est le nombre de degrés de liberté associé à la variation résiduelle.

Conclusion du test :

Le rejet de l'hypothèse nulle H_0 signifie que l'expérience met en évidence un effet significatif de la source de variation étudiée sur la variable principale : au moins une moyenne théorique est différente des autres.

Le non rejet de l'hypothèse nulle H_0 signifie que l'expérience n'a pu mettre en évidence une influence significative de la source de variation étudiée sur la variable principale : il n'y a pas d'écart significatif entre les différentes moyennes théoriques.

Utilisation des probabilités critiques (p-values)

Les logiciels modernes fournissent les résultats des règles de décision sous forme de probabilités critiques. Par définition, la probabilité critique p_c (en anglais p-value) du test de Fisher est la probabilité qu'une réalisation de la loi de Fisher-Snedecor soit plus élevée que la valeur de la statistique de Fisher F sous l'hypothèse H_0. Ainsi, une très faible valeur pour p_c indique que l'hypothèse privilégiée H_0 est vraisemblablement fausse. Plus p_c est faible, plus les données témoignent de l'effet significatif de la source de variation étudiée. Elles nous conduisent alors à rejeter H_0.

Remarque :

Avec MS EXCEL ou OpenOffice.org Calc, la fonction SOMME.CARRES.ECARTS permet de calculer la somme des carrés d'écart pour une série de données (utiliser le bouton fx puis chercher dans la catégorie "Statistique").

Chapitre 6

Tables statistiques

Les fractiles et les fréquences espérées cumulées de certaines lois de probabilité classiques peuvent être obtenus à l'aide d'Excel ou R. Les tables proposées par la suite concernent la loi gaussienne centrée réduite, la loi de Student, la loi de Pearson et loi de Fisher-Snedecor. Ces lois interviennent dans les règles de décision de divers tests d'hypothèses. Les logiciels statistiques actuels peuvent effectuer le calcul de probabilité critique correspondant à la probabilité de dépassement de seuil par le critère de test observé. Ceci permet, quand c'est le cas, de se passer de ces tables.

6.1 Fonction de répartition de la loi normale centrée réduite

Si U suit la loi normale centrée réduite, pour $x \geq 0$, la table donne la valeur $\phi(x) = P(U \leq x)$ avec $x = x_1 + x_2$ où x_1 et x_2 sont indiqués en marge. Pour $x < 0$, on utilise $\phi(x) = 1 - \phi(-x)$.

	x_1									
x_2	**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
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

6.2 Fractiles de la loi normale centrée réduite

u_p est le fractile d'ordre p de la loi normale centrée réduite. Donc $\phi(u_p) = p$.
La table donne la valeur u_p pour $p = p_1 + p_2$ avec p_1 et p_2 indiqués en marge.
Pour les valeurs $p < 0,5$, on utilise la relation $u_p = -u_{1-p}$.

	p_1									
p_2	**0.000**	**0.001**	**0.002**	**0.003**	**0.004**	**0.005**	**0.006**	**0.007**	**0.008**	**0.009**
0.50	0.0000	0.0025	0.0050	0.0075	0.0100	0.0125	0.0150	0.0175	0.0201	0.0226
0.51	0.0251	0.0276	0.0301	0.0326	0.0351	0.0376	0.0401	0.0426	0.0451	0.0476
0.52	0.0502	0.0527	0.0552	0.0577	0.0602	0.0627	0.0652	0.0677	0.0702	0.0728
0.53	0.0753	0.0778	0.0803	0.0828	0.0853	0.0878	0.0904	0.0929	0.0954	0.0979
0.54	0.1004	0.1030	0.1055	0.1080	0.1105	0.1130	0.1156	0.1181	0.1206	0.1231
0.55	0.1257	0.1282	0.1307	0.1332	0.1358	0.1383	0.1408	0.1434	0.1459	0.1484
0.56	0.1510	0.1535	0.1560	0.1586	0.1611	0.1637	0.1662	0.1687	0.1713	0.1738
0.57	0.1764	0.1789	0.1815	0.1840	0.1866	0.1891	0.1917	0.1942	0.1968	0.1993
0.58	0.2019	0.2045	0.2070	0.2096	0.2121	0.2147	0.2173	0.2198	0.2224	0.2250
0.59	0.2275	0.2301	0.2327	0.2353	0.2378	0.2404	0.2430	0.2456	0.2482	0.2508
0.60	0.2533	0.2559	0.2585	0.2611	0.2637	0.2663	0.2689	0.2715	0.2741	0.2767
0.61	0.2793	0.2819	0.2845	0.2871	0.2898	0.2924	0.2950	0.2976	0.3002	0.3029
0.62	0.3055	0.3081	0.3107	0.3134	0.3160	0.3186	0.3213	0.3239	0.3266	0.3292
0.63	0.3319	0.3345	0.3372	0.3398	0.3425	0.3451	0.3478	0.3505	0.3531	0.3558
0.64	0.3585	0.3611	0.3638	0.3665	0.3692	0.3719	0.3745	0.3772	0.3799	0.3826
0.65	0.3853	0.3880	0.3907	0.3934	0.3961	0.3989	0.4016	0.4043	0.4070	0.4097
0.66	0.4125	0.4152	0.4179	0.4207	0.4234	0.4261	0.4289	0.4316	0.4344	0.4372
0.67	0.4399	0.4427	0.4454	0.4482	0.4510	0.4538	0.4565	0.4593	0.4621	0.4649
0.68	0.4677	0.4705	0.4733	0.4761	0.4789	0.4817	0.4845	0.4874	0.4902	0.4930
0.69	0.4959	0.4987	0.5015	0.5044	0.5072	0.5101	0.5129	0.5158	0.5187	0.5215
0.70	0.5244	0.5273	0.5302	0.5330	0.5359	0.5388	0.5417	0.5446	0.5476	0.5505
0.71	0.5534	0.5563	0.5592	0.5622	0.5651	0.5681	0.5710	0.5740	0.5769	0.5799
0.72	0.5828	0.5858	0.5888	0.5918	0.5948	0.5978	0.6008	0.6038	0.6068	0.6098
0.73	0.6128	0.6158	0.6189	0.6219	0.6250	0.6280	0.6311	0.6341	0.6372	0.6403
0.74	0.6433	0.6464	0.6495	0.6526	0.6557	0.6588	0.6620	0.6651	0.6682	0.6713
0.75	0.6745	0.6776	0.6808	0.6840	0.6871	0.6903	0.6935	0.6967	0.6999	0.7031
0.76	0.7063	0.7095	0.7128	0.7160	0.7192	0.7225	0.7257	0.7290	0.7323	0.7356
0.77	0.7388	0.7421	0.7454	0.7488	0.7521	0.7554	0.7588	0.7621	0.7655	0.7688
0.78	0.7722	0.7756	0.7790	0.7824	0.7858	0.7892	0.7926	0.7961	0.7995	0.8030
0.79	0.8064	0.8099	0.8134	0.8169	0.8204	0.8239	0.8274	0.8310	0.8345	0.8381
0.80	0.8416	0.8452	0.8488	0.8524	0.8560	0.8596	0.8633	0.8669	0.8705	0.8742
0.81	0.8779	0.8816	0.8853	0.8890	0.8927	0.8965	0.9002	0.9040	0.9078	0.9116
0.82	0.9154	0.9192	0.9230	0.9269	0.9307	0.9346	0.9385	0.9424	0.9463	0.9502
0.83	0.9542	0.9581	0.9621	0.9661	0.9701	0.9741	0.9782	0.9822	0.9863	0.9904
0.84	0.9945	0.9986	1.0027	1.0069	1.0110	1.0152	1.0194	1.0237	1.0279	1.0322
0.85	1.0364	1.0407	1.0450	1.0494	1.0537	1.0581	1.0625	1.0669	1.0714	1.0758
0.86	1.0803	1.0848	1.0893	1.0939	1.0985	1.1031	1.1077	1.1123	1.1170	1.1217
0.87	1.1264	1.1311	1.1359	1.1407	1.1455	1.1503	1.1552	1.1601	1.1650	1.1700
0.88	1.1750	1.1800	1.1850	1.1901	1.1952	1.2004	1.2055	1.2107	1.2160	1.2212
0.89	1.2265	1.2319	1.2372	1.2426	1.2481	1.2536	1.2591	1.2646	1.2702	1.2759
0.90	1.2816	1.2873	1.2930	1.2988	1.3047	1.3106	1.3165	1.3225	1.3285	1.3346
0.91	1.3408	1.3469	1.3532	1.3595	1.3658	1.3722	1.3787	1.3852	1.3917	1.3984
0.92	1.4051	1.4118	1.4187	1.4255	1.4325	1.4395	1.4466	1.4538	1.4611	1.4684
0.93	1.4758	1.4833	1.4909	1.4985	1.5063	1.5141	1.5220	1.5301	1.5382	1.5464
0.94	1.5548	1.5632	1.5718	1.5805	1.5893	1.5982	1.6072	1.6164	1.6258	1.6352
0.95	1.6449	1.6546	1.6646	1.6747	1.6849	1.6954	1.7060	1.7169	1.7279	1.7392
0.96	1.7507	1.7624	1.7744	1.7866	1.7991	1.8119	1.8250	1.8384	1.8522	1.8663
0.97	1.8808	1.8957	1.9110	1.9268	1.9431	1.9600	1.9774	1.9954	2.0141	2.0335
0.98	2.0537	2.0749	2.0969	2.1201	2.1444	2.1701	2.1973	2.2262	2.2571	2.2904
0.99	2.3263	2.3656	2.4089	2.4573	2.5121	2.5758	2.6521	2.7478	2.8782	3.0902

6.3 Fractiles de la loi de Student

$t_{\nu,p}$ est le fractile d'ordre p de la loi de Student à ν degrés de liberté.
Pour les valeurs de $p \leq 0,5$, on utilise la relation $t_{\nu,p} = -t_{\nu,1-p}$.
Lorsque $\nu > 50$, on utilise l'approximation de la loi de Student par la loi normale $\mathcal{N}(0,1)$, ce qui revient à : $t_{\nu,p} \approx u_p$.

					p					
ν	**0.60**	**0.70**	**0.80**	**0.90**	**0.95**	**0.9750**	**0.9900**	**0.9950**	**0.9990**	**0.9995**
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
31	0.256	0.530	0.853	1.309	1.696	2.040	2.453	2.744	3.375	3.633
32	0.255	0.530	0.853	1.309	1.694	2.037	2.449	2.738	3.365	3.622
33	0.255	0.530	0.853	1.308	1.692	2.035	2.445	2.733	3.356	3.611
34	0.255	0.529	0.852	1.307	1.691	2.032	2.441	2.728	3.348	3.601
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724	3.340	3.591
36	0.255	0.529	0.852	1.306	1.688	2.028	2.434	2.719	3.333	3.582
37	0.255	0.529	0.851	1.305	1.687	2.026	2.431	2.715	3.326	3.574
38	0.255	0.529	0.851	1.304	1.686	2.024	2.429	2.712	3.319	3.566
39	0.255	0.529	0.851	1.304	1.685	2.023	2.426	2.708	3.313	3.558
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
41	0.255	0.529	0.850	1.303	1.683	2.020	2.421	2.701	3.301	3.544
42	0.255	0.528	0.850	1.302	1.682	2.018	2.418	2.698	3.296	3.538
43	0.255	0.528	0.850	1.302	1.681	2.017	2.416	2.695	3.291	3.532
44	0.255	0.528	0.850	1.301	1.680	2.015	2.414	2.692	3.286	3.526
45	0.255	0.528	0.850	1.301	1.679	2.014	2.412	2.690	3.281	3.520
46	0.255	0.528	0.850	1.300	1.679	2.013	2.410	2.687	3.277	3.515
47	0.255	0.528	0.849	1.300	1.678	2.012	2.408	2.685	3.273	3.510
48	0.255	0.528	0.849	1.299	1.677	2.011	2.407	2.682	3.269	3.505
49	0.255	0.528	0.849	1.299	1.677	2.010	2.405	2.680	3.265	3.500
50	0.255	0.528	0.849	1.299	1.676	2.009	2.403	2.678	3.261	3.496

6.4 Fractiles de la loi du χ^2

$\chi^2_{\nu,p}$ est le fractile d'ordre p de la loi du χ^2.

Pour les valeurs de $\nu > 50$, on utilise l'approximation $\chi^2_{\nu,p} \approx \frac{(u_p + \sqrt{2\nu - 1})^2}{2}$

	p												
ν	**0.001**	**0.005**	**0.010**	**0.025**	**0.05**	**0.1000**	**0.5000**	**0.9000**	**0.9500**	**0.9750**	**0.9900**	**0.9950**	**0.9990**
1	0.000	0.000	0.000	0.001	0.004	0.016	0.455	2.706	3.841	5.024	6.635	7.879	10.828
2	0.002	0.010	0.020	0.051	0.103	0.211	1.386	4.605	5.991	7.378	9.210	10.597	13.816
3	0.024	0.072	0.115	0.216	0.352	0.584	2.366	6.251	7.815	9.348	11.345	12.838	16.266
4	0.091	0.207	0.297	0.484	0.711	1.064	3.357	7.779	9.488	11.143	13.277	14.860	18.467
5	0.210	0.412	0.554	0.831	1.145	1.610	4.351	9.236	11.070	12.833	15.086	16.750	20.515
6	0.381	0.676	0.872	1.237	1.635	2.204	5.348	10.645	12.592	14.449	16.812	18.548	22.458
7	0.598	0.989	1.239	1.690	2.167	2.833	6.346	12.017	14.067	16.013	18.475	20.278	24.322
8	0.857	1.344	1.646	2.180	2.733	3.490	7.344	13.362	15.507	17.535	20.090	21.955	26.124
9	1.152	1.735	2.088	2.700	3.325	4.168	8.343	14.684	16.919	19.023	21.666	23.589	27.877
10	1.479	2.156	2.558	3.247	3.940	4.865	9.342	15.987	18.307	20.483	23.209	25.188	29.588
11	1.834	2.603	3.053	3.816	4.575	5.578	10.341	17.275	19.675	21.920	24.725	26.757	31.264
12	2.214	3.074	3.571	4.404	5.226	6.304	11.340	18.549	21.026	23.337	26.217	28.300	32.909
13	2.617	3.565	4.107	5.009	5.892	7.042	12.340	19.812	22.362	24.736	27.688	29.819	34.528
14	3.041	4.075	4.660	5.629	6.571	7.790	13.339	21.064	23.685	26.119	29.141	31.319	36.123
15	3.483	4.601	5.229	6.262	7.261	8.547	14.339	22.307	24.996	27.488	30.578	32.801	37.697
16	3.942	5.142	5.812	6.908	7.962	9.312	15.338	23.542	26.296	28.845	32.000	34.267	39.252
17	4.416	5.697	6.408	7.564	8.672	10.085	16.338	24.769	27.587	30.191	33.409	35.718	40.790
18	4.905	6.265	7.015	8.231	9.390	10.865	17.338	25.989	28.869	31.526	34.805	37.156	42.312
19	5.407	6.844	7.633	8.907	10.117	11.651	18.338	27.204	30.144	32.852	36.191	38.582	43.820
20	5.921	7.434	8.260	9.591	10.851	12.443	19.337	28.412	31.410	34.170	37.566	39.997	45.315
21	6.447	8.034	8.897	10.283	11.591	13.240	20.337	29.615	32.671	35.479	38.932	41.401	46.797
22	6.983	8.643	9.542	10.982	12.338	14.041	21.337	30.813	33.924	36.781	40.289	42.796	48.268
23	7.529	9.260	10.196	11.689	13.091	14.848	22.337	32.007	35.172	38.076	41.638	44.181	49.728
24	8.085	9.886	10.856	12.401	13.848	15.659	23.337	33.196	36.415	39.364	42.980	45.559	51.179
25	8.649	10.520	11.524	13.120	14.611	16.473	24.337	34.382	37.652	40.646	44.314	46.928	52.620
26	9.222	11.160	12.198	13.844	15.379	17.292	25.336	35.563	38.885	41.923	45.642	48.290	54.052
27	9.803	11.808	12.879	14.573	16.151	18.114	26.336	36.741	40.113	43.195	46.963	49.645	55.476
28	10.391	12.461	13.565	15.308	16.928	18.939	27.336	37.916	41.337	44.461	48.278	50.993	56.892
29	10.986	13.121	14.256	16.047	17.708	19.768	28.336	39.087	42.557	45.722	49.588	52.336	58.301
30	11.588	13.787	14.953	16.791	18.493	20.599	29.336	40.256	43.773	46.979	50.892	53.672	59.703
31	12.196	14.458	15.655	17.539	19.281	21.434	30.336	41.422	44.985	48.232	52.191	55.003	61.098
32	12.811	15.134	16.362	18.291	20.072	22.271	31.336	42.585	46.194	49.480	53.486	56.328	62.487
33	13.431	15.815	17.074	19.047	20.867	23.110	32.336	43.745	47.400	50.725	54.776	57.648	63.870
34	14.057	16.501	17.789	19.806	21.664	23.952	33.336	44.903	48.602	51.966	56.061	58.964	65.247
35	14.688	17.192	18.509	20.569	22.465	24.797	34.336	46.059	49.802	53.203	57.342	60.275	66.619
36	15.324	17.887	19.233	21.336	23.269	25.643	35.336	47.212	50.998	54.437	58.619	61.581	67.985
37	15.965	18.586	19.960	22.106	24.075	26.492	36.336	48.363	52.192	55.668	59.893	62.883	69.346
38	16.611	19.289	20.691	22.878	24.884	27.343	37.335	49.513	53.384	56.896	61.162	64.181	70.703
39	17.262	19.996	21.426	23.654	25.695	28.196	38.335	50.660	54.572	58.120	62.428	65.476	72.055
40	17.916	20.707	22.164	24.433	26.509	29.051	39.335	51.805	55.758	59.342	63.691	66.766	73.402
41	18.575	21.421	22.906	25.215	27.326	29.907	40.335	52.949	56.942	60.561	64.950	68.053	74.745
42	19.239	22.138	23.650	25.999	28.144	30.765	41.335	54.090	58.124	61.777	66.206	69.336	76.084
43	19.906	22.859	24.398	26.785	28.965	31.625	42.335	55.230	59.304	62.990	67.459	70.616	77.419
44	20.576	23.584	25.148	27.575	29.787	32.487	43.335	56.369	60.481	64.201	68.710	71.893	78.750
45	21.251	24.311	25.901	28.366	30.612	33.350	44.335	57.505	61.656	65.410	69.957	73.166	80.077
46	21.929	25.041	26.657	29.160	31.439	34.215	45.335	58.641	62.830	66.617	71.201	74.437	81.400
47	22.610	25.775	27.416	29.956	32.268	35.081	46.335	59.774	64.001	67.821	72.443	75.704	82.720
48	23.295	26.511	28.177	30.755	33.098	35.949	47.335	60.907	65.171	69.023	73.683	76.969	84.037
49	23.983	27.249	28.941	31.555	33.930	36.818	48.335	62.038	66.339	70.222	74.919	78.231	85.351
50	24.674	27.991	29.707	32.357	34.764	37.689	49.335	63.167	67.505	71.420	76.154	79.490	86.661

6.5 Fractiles de la loi de Fisher-Snédécor

$f_{\nu_1,\nu_2,p}$ est le fractile d'ordre p de la loi de Fisher-Snédécor à ν_1 et ν_2 degrés de liberté.
Les tables statistiques qui suivent donnent les valeurs de $f_{\nu_1,\nu_2,p}$ pour $p \in \{0,90; 0,95; 0,975; 0,99\}$.
Pour $p \in \{0,01; 0,025; 0,05; 0,10\}$, on utilise la relation $f_{\nu_1,\nu_2,p} = 1/f_{\nu_2,\nu_1,1-p}$.

ν_2	$\nu_1 \rightarrow$	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
$\downarrow$	p														
1	0.900	49.5	53.6	55.8	57.2	58.2	59.1	59.7	60.5	61.0	61.5	62.0	62.6	63.0	63.3
	0.950	199.	216.	225.	230.	234.	237.	239.	242.	244.	246.	248.	250.	252.	254.
	0.975	800.	864.	900.	922.	937.	948.	957.	969.	977.	985.	993.			
	0.990														
	0.999														
2	0.900	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.39	9.41	9.43	9.44	9.46	9.47	9.49
	0.950	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
	0.975	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5
	0.990	99.0	99.2	99.2	99.3	99.3	99.4	100.	100.	100.	100.	100.	100.	100.	99.5
	0.999	999.	999.												
3	0.900	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.23	5.22	5.20	5.18	5.17	5.15	5.13
	0.950	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.70	8.66	8.62	8.58	8.53
	0.975	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.4	14.3	14.3	14.2	14.1	14.0	13.9
	0.990	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.2	27.1	26.9	26.7	26.5	26.4	26.1
	0.999	149.	141.	137.	135.	133.	132.	131.	129.	128.	127.	126.	125.	125.	123.
4	0.900	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.92	3.90	3.87	3.84	3.82	3.79	3.76
	0.950	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.86	5.80	5.75	5.70	5.63
	0.975	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.84	8.75	8.66	8.56	8.46	8.38	8.26
	0.990	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.5	14.4	14.2	14.0	13.8	13.7	13.5
	0.999	61.2	56.2	53.4	51.7	50.5	49.7	49.0	48.0	47.4	46.8	46.1	45.4	44.9	44.1
5	0.900	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.30	3.27	3.24	3.21	3.17	3.15	3.10
	0.950	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.62	4.56	4.50	4.44	4.36
	0.975	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.62	6.52	6.43	6.33	6.23	6.14	6.02
	0.990	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.1	9.89	9.72	9.55	9.38	9.24	9.02
	0.999	37.1	33.2	31.1	29.8	28.8	28.2	27.6	26.9	26.4	25.9	25.4	24.9	24.4	23.8
6	0.900	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.94	2.90	2.87	2.84	2.80	2.77	2.72
	0.950	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.94	3.87	3.81	3.75	3.67
	0.975	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.46	5.37	5.27	5.17	5.07	4.98	4.85
	0.990	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.87	7.72	7.56	7.40	7.23	7.09	6.88
	0.999	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.4	18.0	17.6	17.1	16.7	16.3	15.7
7	0.900	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.70	2.67	2.63	2.59	2.56	2.52	2.47
	0.950	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.51	3.44	3.38	3.32	3.23
	0.975	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.76	4.67	4.57	4.47	4.36	4.28	4.14
	0.990	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.62	6.47	6.31	6.16	5.99	5.86	5.65
	0.999	21.7	18.8	17.2	16.2	15.5	15.0	14.6	14.1	13.7	13.3	12.9	12.5	12.2	11.7
8	0.900	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.54	2.50	2.46	2.42	2.38	2.35	2.29
	0.950	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.22	3.15	3.08	3.02	2.93
	0.975	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.29	4.20	4.10	4.00	3.89	3.81	3.67
	0.990	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.81	5.67	5.52	5.36	5.20	5.07	4.86
	0.999	18.5	15.8	14.4	13.5	12.9	12.4	12.0	11.5	11.2	10.8	10.5	10.1	9.80	9.33

Fractiles $f_{\nu_1,\nu_2,p}$ de la loi de Fisher-Snédécor

ν_2	$\nu_1 \rightarrow$	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
$\downarrow$	p														
9	0.900	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.42	2.38	2.34	2.30	2.25	2.22	2.16
	0.950	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	3.01	2.94	2.86	2.80	2.71
	0.975	5.71	5.08	4.72	4.48	4.32	4.20	4.10	3.96	3.87	3.77	3.67	3.56	3.47	3.33
	0.990	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.26	5.11	4.96	4.81	4.65	4.52	4.31
	0.999	16.4	13.9	12.6	11.7	11.1	10.7	10.4	9.89	9.57	9.24	8.90	8.55	8.26	7.81
10	0.900	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.32	2.28	2.24	2.20	2.16	2.12	2.06
	0.950	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.84	2.77	2.70	2.64	2.54
	0.975	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.72	3.62	3.52	3.42	3.31	3.22	3.08
	0.990	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.85	4.71	4.56	4.41	4.25	4.11	3.91
	0.999	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.75	8.45	8.13	7.80	7.47	7.19	6.76
11	0.900	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.25	2.21	2.17	2.12	2.08	2.04	1.97
	0.950	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.72	2.65	2.57	2.51	2.40
	0.975	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.53	3.43	3.33	3.23	3.12	3.03	2.88
	0.990	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.54	4.40	4.25	4.10	3.94	3.81	3.60
	0.999	13.8	11.6	10.3	9.58	9.05	8.66	8.35	7.92	7.63	7.32	7.01	6.68	6.42	6.00
12	0.900	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.19	2.15	2.10	2.06	2.01	1.97	1.90
	0.950	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.62	2.54	2.47	2.40	2.30
	0.975	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.37	3.28	3.18	3.07	2.96	2.87	2.72
	0.990	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.30	4.16	4.01	3.86	3.70	3.57	3.36
	0.999	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.29	7.00	6.71	6.40	6.09	5.83	5.42
13	0.900	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.14	2.10	2.05	2.01	1.96	1.92	1.85
	0.950	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.53	2.46	2.38	2.31	2.21
	0.975	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.25	3.15	3.05	2.95	2.84	2.74	2.60
	0.990	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.10	3.96	3.82	3.66	3.51	3.37	3.17
	0.999	12.3	10.2	9.07	8.35	7.86	7.49	7.21	6.80	6.52	6.23	5.93	5.63	5.37	4.97
14	0.900	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.10	2.05	2.01	1.96	1.91	1.87	1.80
	0.950	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.46	2.39	2.31	2.24	2.13
	0.975	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.15	3.05	2.95	2.84	2.73	2.64	2.49
	0.990	6.51	5.56	5.04	4.69	4.46	4.28	4.14	3.94	3.80	3.66	3.51	3.35	3.22	3.00
	0.999	11.8	9.73	8.62	7.92	7.44	7.08	6.80	6.40	6.13	5.85	5.56	5.25	5.00	4.60
15	0.900	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.06	2.02	1.97	1.92	1.87	1.83	1.76
	0.950	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.40	2.33	2.25	2.18	2.07
	0.975	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.06	2.96	2.86	2.76	2.64	2.55	2.40
	0.990	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.80	3.67	3.52	3.37	3.21	3.08	2.87
	0.999	11.3	9.34	8.25	7.57	7.09	6.74	6.47	6.08	5.81	5.53	5.25	4.95	4.70	4.31
16	0.900	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.03	1.99	1.94	1.89	1.84	1.79	1.72
	0.950	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.35	2.28	2.19	2.12	2.01
	0.975	4.69	4.08	3.73	3.50	3.34	3.22	3.12	2.99	2.89	2.79	2.68	2.57	2.47	2.32
	0.990	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.69	3.55	3.41	3.26	3.10	2.97	2.75
	0.999	11.0	9.01	7.94	7.27	6.80	6.46	6.19	5.81	5.55	5.27	4.99	4.70	4.45	4.06
17	0.900	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.00	1.96	1.91	1.86	1.81	1.76	1.69
	0.950	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.31	2.23	2.15	2.08	1.96
	0.975	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.92	2.82	2.72	2.62	2.50	2.41	2.25
	0.990	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.59	3.46	3.31	3.16	3.00	2.87	2.65
	0.999	10.7	8.73	7.68	7.02	6.56	6.22	5.96	5.58	5.32	5.05	4.77	4.48	4.24	3.85

Fractiles $f_{\nu_1,\nu_2,p}$ de la loi de Fisher-Snédécor

ν_2	$\nu_1 \rightarrow$	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
$\downarrow$	p														
18	0.900	2.62	2.42	2.29	2.20	2.13	2.08	2.04	1.98	1.93	1.89	1.84	1.78	1.74	1.66
	0.950	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.41	2.34	2.27	2.19	2.11	2.04	1.92
	0.975	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.87	2.77	2.67	2.56	2.44	2.35	2.19
	0.990	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.51	3.37	3.23	3.08	2.92	2.78	2.57
	0.999	10.4	8.49	7.46	6.81	6.35	6.02	5.76	5.39	5.13	4.87	4.59	4.30	4.06	3.67
19	0.900	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.96	1.91	1.86	1.81	1.76	1.71	1.63
	0.950	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.23	2.16	2.07	2.00	1.88
	0.975	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.82	2.72	2.62	2.51	2.39	2.30	2.13
	0.990	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.43	3.30	3.15	3.00	2.84	2.71	2.49
	0.999	10.2	8.28	7.27	6.62	6.18	5.85	5.59	5.22	4.97	4.70	4.43	4.14	3.90	3.51
20	0.900	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.94	1.89	1.84	1.79	1.74	1.69	1.61
	0.950	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.20	2.12	2.04	1.97	1.84
	0.975	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.77	2.68	2.57	2.46	2.35	2.25	2.09
	0.990	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.37	3.23	3.09	2.94	2.78	2.64	2.42
	0.999	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.08	4.82	4.56	4.29	4.00	3.76	3.38
21	0.900	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.92	1.87	1.83	1.78	1.72	1.67	1.59
	0.950	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.32	2.25	2.18	2.10	2.01	1.94	1.81
	0.975	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.73	2.64	2.53	2.42	2.31	2.21	2.04
	0.990	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.31	3.17	3.03	2.88	2.72	2.58	2.36
	0.999	9.77	7.94	6.95	6.32	5.88	5.56	5.31	4.95	4.70	4.44	4.17	3.88	3.64	3.26
22	0.900	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.90	1.86	1.81	1.76	1.70	1.65	1.57
	0.950	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.30	2.23	2.15	2.07	1.98	1.91	1.78
	0.975	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.70	2.60	2.50	2.39	2.27	2.17	2.00
	0.990	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.26	3.12	2.98	2.83	2.67	2.53	2.31
	0.999	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.83	4.58	4.33	4.06	3.78	3.54	3.15
23	0.900	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.89	1.84	1.80	1.74	1.69	1.64	1.55
	0.950	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.27	2.20	2.13	2.05	1.96	1.88	1.76
	0.975	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.67	2.57	2.47	2.36	2.24	2.14	1.97
	0.990	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.21	3.07	2.93	2.78	2.62	2.48	2.26
	0.999	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.73	4.48	4.23	3.96	3.68	3.44	3.05
24	0.900	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.88	1.83	1.78	1.73	1.67	1.62	1.53
	0.950	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.25	2.18	2.11	2.03	1.94	1.86	1.73
	0.975	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.64	2.54	2.44	2.33	2.21	2.11	1.94
	0.990	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.17	3.03	2.89	2.74	2.58	2.44	2.21
	0.999	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.64	4.39	4.14	3.87	3.59	3.36	2.97
25	0.900	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.87	1.82	1.77	1.72	1.66	1.61	1.52
	0.950	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.24	2.16	2.09	2.01	1.92	1.84	1.71
	0.975	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.61	2.51	2.41	2.30	2.18	2.08	1.91
	0.990	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.13	2.99	2.85	2.70	2.54	2.40	2.17
	0.999	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.56	4.31	4.06	3.79	3.52	3.28	2.89
26	0.900	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.86	1.81	1.76	1.71	1.65	1.59	1.50
	0.950	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.22	2.15	2.07	1.99	1.90	1.82	1.69
	0.975	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.59	2.49	2.39	2.28	2.16	2.05	1.88
	0.990	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.09	2.96	2.81	2.66	2.50	2.36	2.13
	0.999	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.48	4.24	3.99	3.72	3.44	3.21	2.82

Fractiles $f_{\nu_1,\nu_2,p}$ de la loi de Fisher-Snédécor

ν_2	$\nu_1 \rightarrow$	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
$\downarrow$	p														
27	0.900	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.85	1.80	1.75	1.70	1.64	1.58	1.49
	0.950	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.20	2.13	2.06	1.97	1.88	1.81	1.67
	0.975	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.57	2.47	2.36	2.25	2.13	2.03	1.85
	0.990	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.06	2.93	2.78	2.63	2.47	2.33	2.10
	0.999	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.41	4.17	3.92	3.66	3.38	3.14	2.75
28	0.900	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.84	1.79	1.74	1.69	1.63	1.57	1.48
	0.950	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.19	2.12	2.04	1.96	1.87	1.79	1.65
	0.975	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.55	2.45	2.34	2.23	2.11	2.01	1.83
	0.990	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.03	2.90	2.75	2.60	2.44	2.30	2.06
	0.999	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.35	4.11	3.86	3.60	3.32	3.09	2.69
29	0.900	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.83	1.78	1.73	1.68	1.62	1.56	1.47
	0.950	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.18	2.10	2.03	1.94	1.85	1.77	1.64
	0.975	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.53	2.43	2.32	2.21	2.09	1.99	1.81
	0.990	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.00	2.87	2.73	2.57	2.41	2.27	2.03
	0.999	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.29	4.05	3.80	3.54	3.27	3.03	2.64
30	0.900	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.82	1.77	1.72	1.67	1.61	1.55	1.46
	0.950	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	2.01	1.93	1.84	1.76	1.62
	0.975	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.51	2.41	2.31	2.20	2.07	1.97	1.79
	0.990	5.39	4.51	4.02	3.70	3.47	3.30	3.17	2.98	2.84	2.70	2.55	2.39	2.25	2.01
	0.999	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.24	4.00	3.75	3.49	3.22	2.98	2.59
60	0.900	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.71	1.66	1.60	1.54	1.48	1.41	1.29
	0.950	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.84	1.75	1.65	1.56	1.39
	0.975	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.27	2.17	2.06	1.94	1.82	1.70	1.48
	0.990	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.63	2.50	2.35	2.20	2.03	1.88	1.60
	0.999	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.54	3.32	3.08	2.83	2.55	2.32	1.89
80	0.900	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.44	1.38	1.24
	0.950	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.95	1.88	1.79	1.70	1.60	1.51	1.32
	0.975	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.21	2.11	2.00	1.88	1.75	1.63	1.40
	0.990	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.55	2.42	2.27	2.12	1.94	1.79	1.49
	0.999	7.54	5.97	5.12	4.58	4.20	3.92	3.70	3.39	3.16	2.93	2.68	2.41	2.16	1.72
100	0.900	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.66	1.61	1.56	1.49	1.42	1.35	1.21
	0.950	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.93	1.85	1.77	1.68	1.57	1.48	1.28
	0.975	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.18	2.08	1.97	1.85	1.71	1.59	1.35
	0.990	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.50	2.37	2.22	2.07	1.89	1.74	1.43
	0.999	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.30	3.07	2.84	2.59	2.32	2.08	1.62
120	0.900	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.65	1.60	1.54	1.48	1.41	1.34	1.19
	0.950	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.91	1.83	1.75	1.66	1.55	1.46	1.25
	0.975	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.16	2.05	1.94	1.82	1.69	1.56	1.31
	0.990	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.47	2.34	2.19	2.03	1.86	1.70	1.38
	0.999	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.24	3.02	2.78	2.53	2.26	2.02	1.54
∞	0.900	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.60	1.55	1.49	1.42	1.34	1.26	1.00
	0.950	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.67	1.57	1.46	1.35	1.00
	0.975	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.05	1.94	1.83	1.71	1.57	1.43	1.00
	0.990	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.32	2.18	2.04	1.88	1.70	1.52	1.00
	0.999	6.91	5.42	4.62	4.10	3.74	3.47	3.27	2.96	2.74	2.51	2.27	1.99	1.73	1.00