STATISTIQUES INFÉRENTIELLES

Julien JACQUES
http://eric.univ-lyon2.fr/~jjacques/

Table des matières

Chapitre 1

Échantillonnage et statistiques descriptives

La problématique de l'inférence statistique consiste, à partir d'un **échantillon** de données provenant d'une population de loi de probabilité inconnue, à déduire des propriétés sur cette population : quelle est sa loi (problème d'**estimation**, chapitre 2), comment prendre une décision en contrôlant au mieux le risque de se tromper (problème de **test**, chapitre 3).

1.1 Échantillon

Un échantillonnage correspond à des tirages *indépendants* et *équiprobables* d'individus au sein de la population. On associe alors à chaque individu i une variable aléatoire X_i, dont on observe une seule réalisation x_i.

Définition 1.1.1. *Un **échantillon** $X_1, \ldots, X_n$ est un n-uplet $(X_1, \ldots, X_n)$ de variables aléatoires X_i indépendantes et identiquement distribuées (même loi).*

Par simplicité nous utiliserons régulièrement le terme échantillon pour signifier à la fois l'échantillon d'observations $x_1, \ldots, x_n$ et le n-uplet aléatoire $(X_1, \ldots, X_n)$.
Il est fréquent de caractériser un échantillon par des quantités telle que la moyenne, variance, etc. Ces quantités sont elles-mêmes des variables aléatoires fonction de $X_1, \ldots, X_n$.

Définition 1.1.2. *Une **statistique** T est une variable aléatoire fonction (mesurable) de $X_1, \ldots, X_n$.*

1.2 Exemple introductif

Le jeu de données `GermanCredit.data`, disponible en ligne[1], comporte des renseignements sur 1000 clients d'une banque allemande, chaque client étant décrit par 20 variables. Ce jeu de données sera utilisé pour illustrer les notions de ce chapitre. Le tableau 1.2 contient la description des 20 variables.

1.3 Description d'une variable

1.3.1 Les différents types de variables

Les variables que l'on rencontre en statistique peuvent être de différentes natures :

Définition 1.3.1.
- *une variable est **quantitative** si ses valeurs sont mesurables. Elle peut être continue ($\mathbb{R}$) ou discrète ($\mathbb{N}$).*
- *une variable est **qualitative** si ses valeurs ne sont pas des valeurs numériques, mais des caractéristiques, appelées **modalités**.*
- *une variable qualitative est dite **ordinale** si ses valeurs sont naturellement ordonnées (mention au bac, appréciation, classe d'âge...). Dans le cas contraire elle est dite **nominale** (sexe, couleur des cheveux...).*

Exercice. Définir le type de chacune des variables dans l'exemple `GermanCredit.data`.

1. `http://labomath.univ-lille1.fr/~jacques/`

numero	nom de la variable	valeur
1	état du compte chèque (en DM)	A11 : < 0 A12 : $\in [0, 200[$ A13 : ≥ 200 ou versement des salaires pendant au moins un an A14 : pas de compte chèque
2	durée en mois du crédit	$\in \mathbb{N}$
3	historique des crédits	A30 : pas de crédit / tous remboursés A31 : tous les crédits dans la banque remboursés A32 : crédits en cours A33 : retard de paiement dans le passé A34 : compte critique / crédit existant dans d'autre banque
4	but du crédit	A40 : voiture neuve A41 : voiture occasion A42 : équipement / fourniture A43 : radio / télévision A44 : appareils ménagers A45 : réparation A46 : éducation A47 : vacances A48 : recyclage A49 : professionnel A410 : autre
5	montant du crédit (en DM)	$\in \mathbb{R}$
6	montant de l'épargne (en DM)	A61 : < 100 A62 : $\in [100, 500[$ A63 : $\in [500, 100[$ A64 : ≥ 1000 A65 : inconnu
7	ancienneté dans le travail actuel (an)	A71 : sans emploi A72 : < 1 A73 : $\in [1, 4[$ A74 : $\in [4, 7[$ A75 : ≥ 7
8	taux d'apport	$\in \mathbb{R}$
9	état marital	A91 : homme divorcé / séparé A92 : femme divorcé / séparé / mariée A93 : homme célibataire A94 : homme marié / veuf A95 : femme célibataire
10	autre demandeurs / garants	A101 : aucun A102 : co-demandeur A103 : garant
11	durée d'habitation dans la résidence actuelle (an)	$\in \mathbb{N}$
12	biens	A121 : immobilier A122 : si pas A121 : placement (assurance vie ou part dans la banque) A123 : si pas A121 et A122 : voiture ou autre, non compris dans la variable 6 A124 : inconnu
13	âge (an)	$\in \mathbb{N}$
14	autre demande de crédits	A141 : banque A142 : magasins A143 : aucun
15	situation dans la résidence actuelle	A151 : locataire A152 : propriétaire A153 : occupant à titre gratuit
16	nombre de crédits dans la banque	$\in \mathbb{N}$
17	emploi	A171 : sans emploi / non qualifié - étranger A172 : non qualifié - non étranger A173 : emploi qualifié / fonctionnaire A174 : gestion / indépendant / emploi hautement qualifié / haut fonctionnaire
18	nombre de personnes pouvant rembourser le crédit	$\in \mathbb{N}$
19	téléphone	A191 : aucun A192 : oui, enregistré au nom du client
20	travailleur étranger	A201 : oui A202 : non

TABLE 1.1 – Variables du jeu de données `GermanCredit.data`

1.3.2 Résumés numériques d'une variable quantitative

Soit $X_1, \ldots, X_n$ un échantillon d'une variable aléatoire quantitative, de fonction de répartition F.

1.3.2.1 Caractéristiques de tendance centrale

La **moyenne empirique** exprime la valeur moyenne de l'échantillon :

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

Attention, cette quantité est très sensible aux valeurs extrêmes.
Beaucoup moins sensible aux extrêmes, la **médiane** M est la valeur qui partage l'échantillon, rangé dans l'ordre croissant $X_1 \leq X_2 \leq \ldots \leq X_n$ (ou décroissant), en deux parties égales. Si n est impair la médiane sera $X_{\frac{n+1}{2}}$, sinon ce sera par convention $\frac{X_{\frac{n}{2}}+X_{\frac{n}{2}+1}}{2}$. La fonction de répartition vaut 0.5 en la médiane : $F(M) = 0.5$.
Lorsque les données sont entières, on utilise parfois le **mode** qui est la valeur la plus fréquente.

1.3.2.2 Caractéristiques de dispersion

L'**étendue**, ou intervalle de variation est la différence entre les deux valeurs extrêmes : $X_{max} - X_{min}$. Attention, les variables X_{min} et X_{max} n'ont plus la même distribution que les variables $X_1, \ldots, X_n$ de l'échantillon. En effet, on montre (*exercice*) que leur fonction de répartition sont respectivement :

$$F_{min}(x) = F^n(x) \qquad \text{et} \qquad F_{max}(x) = 1 - (1 - F(x))^n.$$

Les 1er et 3ème quartiles Q_1 et Q_3 sont définis par $F(Q_1) = 0.25$ et $F(Q_3) = 0.75$. L'**intervalle inter-quartile** $[Q_1, Q_3]$ contient donc 50% des données.
Bien que l'intervalle inter-quartile soit moins sensible aux valeurs extrêmes que l'étendue, il n'est pas très souvent utilisé. On utilise plus souvent la **variance empirique** V^2 et sa racine carré V l'**écart-type** :

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

L'écart-type a l'avantage de s'exprimer dans la même unité que les données.
Le **coefficient de variation** exprime quant à lui le rapport $V/\bar{X}$.

1.3.2.3 Caractéristiques de forme

Elles permettent de situer la distribution observée par rapport à une distribution de référence qu'est la distribution gaussienne.
Le coefficient d'asymétrie γ_1 (*skewness*) indique la symétrie de la distribution :

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{(\sqrt{n/(n-1)}V)^3},$$

l'intérêt du facteur $\sqrt{n/(n-1)}$ au dénominateur sera précisé au chapitre 2. Il est nul pour une distribution symétrique.
Un γ_1 positif indique une distribution décalée vers la gauche avec une queue de distribution étendue vers la droite.
Le coefficient d'aplatissement γ_2 (*kurtosis*) renseigne sur la diffusion de la distribution :

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{(n/(n-1))^2 V^4}.$$

Il vaut 3 pour une distribution gaussienne. Si la distribution est plus aplatie qu'une gaussienne, le coefficient d'aplatissement sera supérieur à 3.
Attention : certains logiciels et/ou auteurs soustraient 3 à γ_2 pour le comparer directement à 0.

1.3.3 Représentation graphique d'une variable quantitative

1.3.3.1 Boîte à moustaches ou *box plot*

Une **boîte à moustaches** (figure 1.1) résume la série de données à l'aide des caractéristiques suivantes :

- la médiane est le trait centré au milieu de la boîte,
- la boîte est formée par les 1er quartile q_1 et 3ème quartile q_3,
- les moustaches sont définies par les valeurs observées les plus extrêmes dans l'intervalle $[q_1 - 1.5(q_3 - q_1), q_3 + 1.5(q_3 - q_1)]$,
- les $\circ$ représentent les valeurs extrêmes non contenues dans l'intervalle précédent.

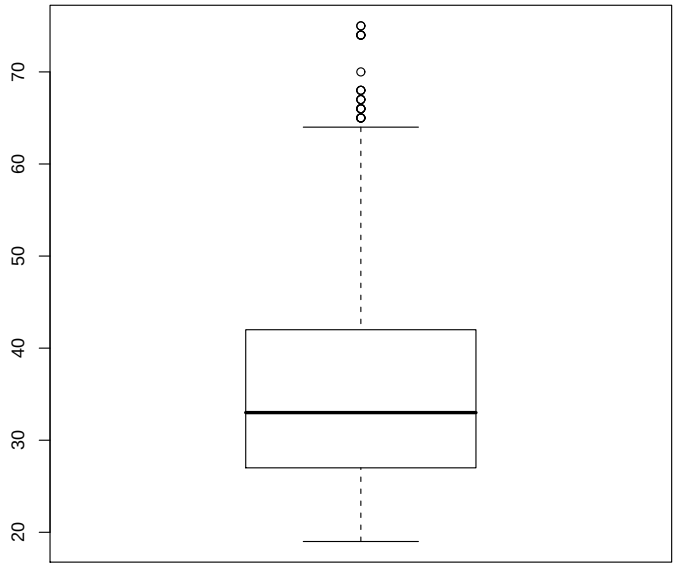

FIGURE 1.1 – Boîte à moustaches illustrant la distribution des âges des clients.

Cette représentation permet également de comparer facilement la distribution de différentes variables, ou encore de la même variable pour différentes modalités d'une variable qualitative (figure 1.2). On remarque ainsi que parmi les clients de la banque allemande les femmes divorcées, séparées ou mariées ainsi que les hommes mariés ou veufs sont généralement moins âgés que les hommes célibataires, divorcés ou séparés.

1.3.3.2 Histogramme

Un **histogramme** est un graphique en barres verticales accolées obtenu après découpage en classes de l'intervalle de variation des données. La surface de chaque barre est proportionnelle à la fréquence de la classe. Pour des classes de même largeur (souvent utilisées dans les logiciels), c'est donc la hauteur de la barre qui est proportionnelle à la fréquence de la classe. La surface de l'ensemble des barres vaut 1.
L'histogramme d'une série de données peut être vue comme une version discontinue empirique de la courbe de densité d'une variable aléatoire. Ainsi, sa visualisation permet d'avoir un avis sur la nature de la distribution des données. Par exemple (figure 1.3), la variable âge ne semble pas suivre une loi normale.
Attention : sur un histogramme figurent en ordonnées des fréquences et non pas des effectifs, comme ont tendance à le faire beaucoup de logiciels !

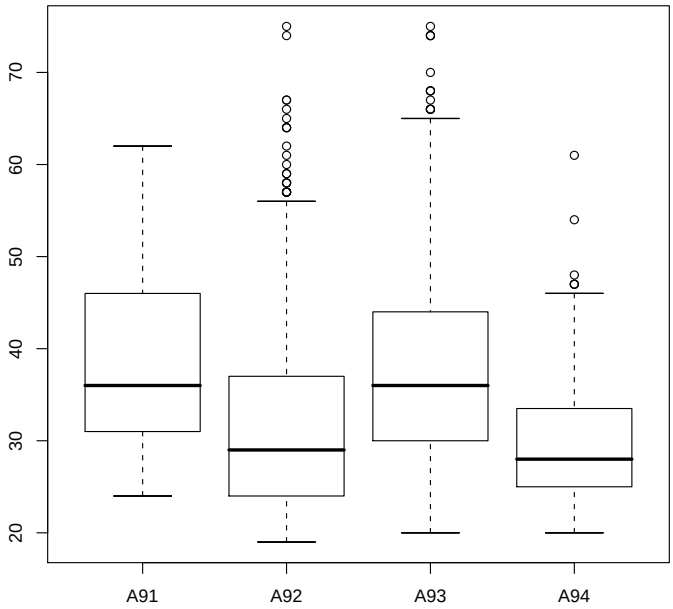

FIGURE 1.2 – Boîte à moustaches illustrant la distribution des âges des clients suivant les différents statut maritaux.

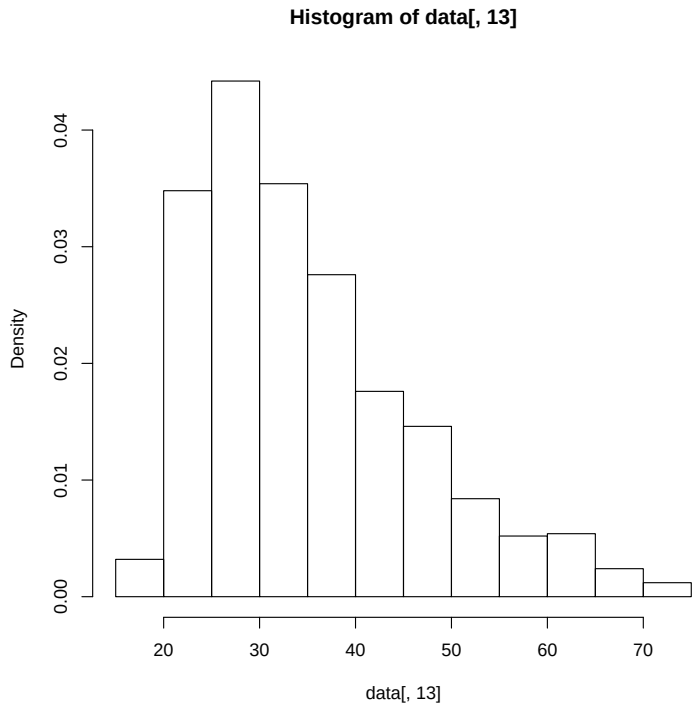

FIGURE 1.3 – Histogramme des âges des clients.

1.3.3.3 La fonction de répartition empirique

La fonction de répartition empirique d'une série de données est définie par :

$$F_n(x) = \frac{N_x}{n}$$

où $N_x = \#\{X_i : X_i \leq x, 1 \leq i \leq n\}$ est le nombre de données inférieures ou égales à x. En tant que fonction de l'échantillon, la fonction de répartition empirique est une variable aléatoire. Voir un exemple de fonction de répartition empirique sur la figure 1.4, calculée et représentée à l'aide de la fonction `ecdf` sous le logiciel **R**.

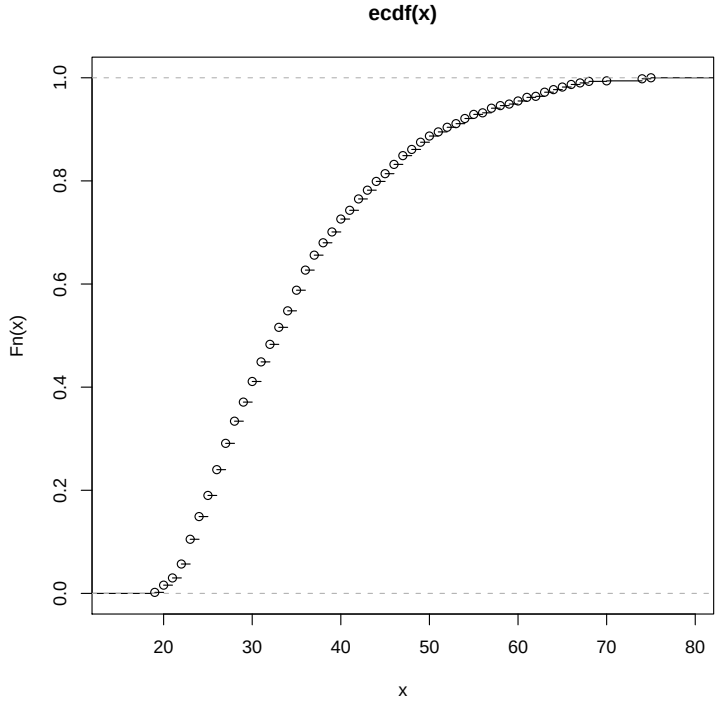

FIGURE 1.4 – Fonction de répartition empirique des âges des clients.

1.3.4 Résumé numérique d'une variable qualitative

Soit X une variable aléatoire qualitative prenant ses valeurs dans l'espace des modalités $\{m_1, \ldots, m_p\}$. Plutôt que de s'intéresser directement à l'échantillon $X_1, \ldots, X_n$, on s'intéresse généralement aux fréquences d'observation de chaque modalité dans cet échantillon. Pour chaque modalité m_j de la variable qualitative ($1 \leq j \leq p$), on note

$$N_j = \#\{X_i : X_i = m_j, 1 \leq i \leq n\}$$

le nombre d'occurrences (effectif) de la modalité m_j dans l'échantillon ($\sum_j^p N_j = n$), et F_i la **fréquence** correspondante :

$$F_j = \frac{N_j}{n}.$$

1.3.5 Représentation graphique d'une variable qualitative

Les variables qualitatives nominales sont généralement représentées sous la forme de camemberts (*pie-chart*, figure 1.5) ou diagramme en barres horizontales (figure 1.6). On utilisera des diagrammes en barres verticales lorsque les variables sont qualitatives ordinales.

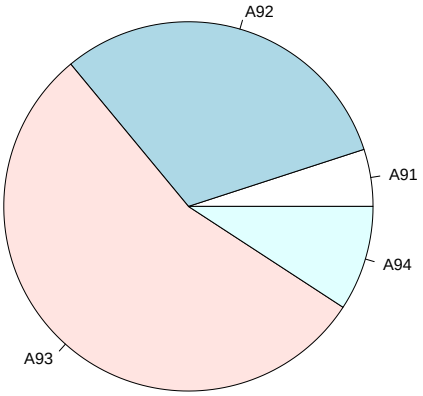

FIGURE 1.5 – Diagrammes en camenbert des situations maritales des clients.

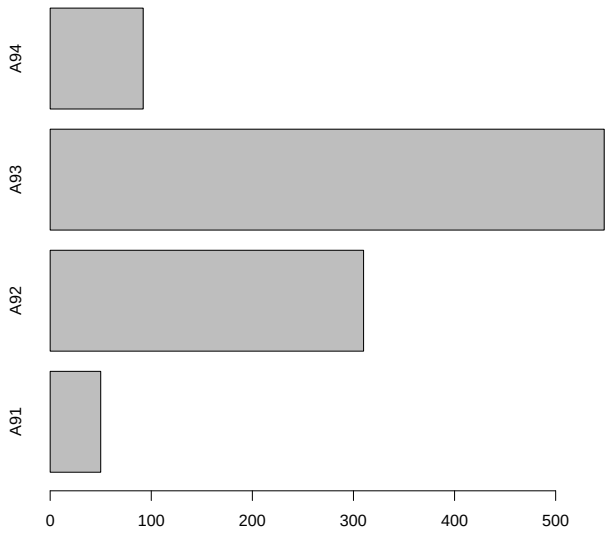

FIGURE 1.6 – Diagrammes en barres des situations maritales des clients.

1.4 Description de plusieurs variables

Nous nous intéressons dans cette section à l'étude simultanée de deux variables, avec comme objectif de mettre en évidence une évolution simultanée de ces deux variables.

1.4.1 Liaison entre deux variables quantitatives

Nuage de points. L'étude graphique du **nuage de points** représentant les deux variables X et Y d'intérêts permet de mettre en évidence un certain lien entre les variables :

- une liaison linéaire positive ou négative,
- une liaison non linéaire,
- une absence de liaison,
- ou encore des structures de liaison plus particulières (absence de liaison en moyenne mais pas en dispersion).

On devine sur l'exemple bancaire (figure 1.7) une liaison linéaire linéaire positive entre la durée et le montant du crédit.

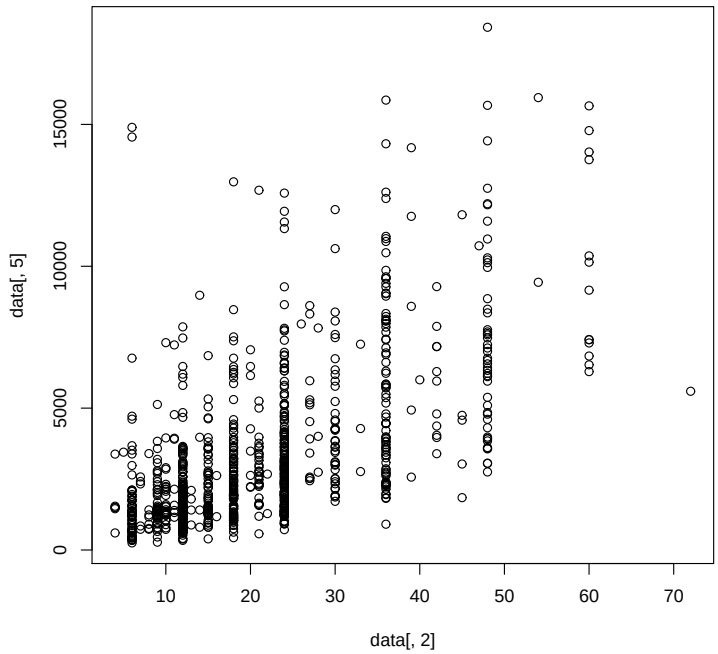

FIGURE 1.7 – Représentations du montant du crédit en fonction de sa durée.

Coefficient de corrélation linéaire L'indice de liaison utilisé est le **coefficient de corrélation linéaire**, défini par :

$$\rho_{XY} = \frac{V_{XY}}{V_X V_Y}$$

où V_X et V_Y sont les écart-types des variables X et Y, et V_{XY} est la covariance empirique entre X et Y, définie par :

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

Le coefficient de corrélation (comme la covariance) est symétrique ($\rho_{XY} = \rho_{YX}$) et prend ses valeurs entre -1 et $+1$.

Attention : si les variables X et Y sont indépendantes, leur covariance est nulle et donc leur coefficient de corrélation linéaire également. Mais la réciproque est fausse !

Coefficient de corrélation partielle Il arrive parfois que l'on constate une corrélation étonnante entre deux variables. Ce phénomène arrive lorsque la corrélation est en fait due à une troisième variable. On cite souvent l'exemple du nombre de maladies mentales (X) corrélé positivement avec le nombre de postes de radio (Y), corrélation purement fictive étant en fait due à une troisième variable non aléatoire, le temps (T). Pour remédier à ce phénomène on utilise le **coefficient de corrélation partielle** (ou conditionnel) de X et Y conditionnellement à T :

$$\rho_{XY \cdot T} = \frac{\rho_{XY} - \rho_{XT}\rho_{YT}}{\sqrt{(1-\rho_{XT}^2)(1-\rho_{YT}^2)}}$$

1.4.2 Liaison entre une variable quantitative et une variable qualitative

On a déjà vu sur la figure 1.2 comment il est possible d'illustrer la liaison entre une variable qualitative et une variable quantitative en représentant côte à côte des boîtes à moustaches pour chaque modalité de la variable qualitative.
Soit X la variable qualitative à R modalités, et Y la variable quantitative. Notons $N_1, \ldots, N_R$ les effectifs de chaque modalité au sein de l'échantillon, $\bar{Y}_1, \ldots, \bar{Y}_R$ et $V_1^2, \ldots, V_R^2$ les moyennes et variances de Y pour chaque modalité de X, et $\bar{Y}$ et V^2 les moyenne et variance globales de Y.
On montre alors que la variance de Y peut se décomposer suivant la **formule d'analyse de variance** suivante :

$$V^2 = \underbrace{\frac{1}{n}\sum_{j=1}^{R} N_j(\bar{Y}_j - \bar{Y})^2}_{V_X^2 : \text{variance inter (between) ou expliquee par} X} + \underbrace{\frac{1}{n}\sum_{j=1}^{R} N_j V_j^2}_{\text{variance intra (within) ou résiduelle}} .$$

Cette formule d'analyse de variance est l'analogue empirique, dans le cas où X est une variable aléatoire qualitative, de la formule vue en probabilité :

$$V(Y) = V(E[Y|X]) + E[V(Y|X)].$$

On peut alors définir comme indice de liaison le **rapport de corrélation** :

$$R_{Y|X} = \sqrt{\frac{V_X^2}{V^2}}.$$

Le carré de ce rapport est appelé **coefficient de détermination**, et est également utilisé par la suite pour exprimer le degré de liaison entre deux variables quantitatives.

1.4.3 Liaisons entre deux variables qualitatives

Soient deux variables aléatoires qualitatives pouvant prendre respectivement R et C modalités : $m_1, \ldots, m_R$ et $o_1, \ldots, o_C$. Les données de ce type sont présentées dans un tableau dans lequel les modalités de X figurent en ligne et celles de Y en colonne, contenant dans chaque case les effectifs conjoints N_{rc}. Un tel tableau est appelé **table de contingence** :
Les $N_{r\cdot}$ et $N_{\cdot c}$ sont les **marges**, ou effectifs marginaux, en lignes et en colonnes.
On appelle r-ème **profil-ligne** l'ensemble des fréquences de la variables Y conditionnelles à la modalités m_r de X :

$$\{\frac{N_{r1}}{N_{r\cdot}}, \cdots, \frac{N_{rc}}{N_{r\cdot}}, \cdots, \frac{N_{rC}}{N_{r\cdot}}\}.$$

De même on définit le c-ème **profil-colonne** :

$$\{\frac{N_{1c}}{N_{\cdot c}}, \cdots, \frac{N_{rc}}{N_{\cdot c}}, \cdots, \frac{N_{Rc}}{N_{\cdot c}}\}.$$

	o_1	$\cdots$	o_c	$\cdots$	o_C	sommes
m_1	N_{11}	$\cdots$	N_{1c}	$\cdots$	N_{1C}	$N_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
m_r	N_{r1}	$\cdots$	N_{rc}	$\cdots$	N_{rC}	$N_{r\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
m_R	N_{R1}	$\cdots$	N_{Rc}	$\cdots$	N_{RC}	$N_{R\cdot}$
sommes	$N_{\cdot 1}$	$\cdots$	$N_{\cdot c}$	$\cdots$	$N_{\cdot C}$	n

TABLE 1.2 – Table de contingence

Lorsque aucune liaison n'existe entre les deux variables qualitatives, tous les profils-lignes sont égaux entre eux, ainsi que tous les profils-colonnes. On a ainsi

$$N_{rc} = \frac{N_{r\cdot}N_{\cdot c}}{n} \qquad \forall 1 \leq r \leq R, 1 \leq c \leq C.$$

Une mesure de la liaison entre les deux variables peut être faite en évaluant l'écart à cette situation de non liaison, par l'indice suivant :

$$\chi^2 = \sum_{r=1}^{R}\sum_{c=1}^{C} \frac{\left(N_{rc} - \frac{N_{r\cdot}N_{\cdot c}}{n}\right)^2}{\frac{N_{r\cdot}N_{\cdot c}}{n}} = n\left[\sum_{r=1}^{R}\sum_{c=1}^{C}\frac{N_{rc}^2}{N_{r\cdot}N_{\cdot c}} - 1\right]$$

Le χ^2 est toujours positif ou nul, et il est d'autant plus grand que la liaison est forte. Malheureusement cet indice dépend des dimensions R et C ainsi que de l'effectif total n. D'autres indicateurs sont alors utilisés comme :

- le $\Phi^2 = \frac{\chi^2}{n}$ qui dépend encore de C et de R,
- le V de Cramer

$$V_{Cramer} = \sqrt{\frac{\Phi^2}{inf(R,C) - 1}}$$

  qui est compris entre 0 et 1,
- le T de Tschuprow

$$T_{Tschuprow} = \sqrt{\frac{\Phi^2}{(R-1)(C-1)}}$$

  qui est compris entre 0 et 1 et est inférieur au V de Cramer.

1.4.3.1 Cas des variables ordinales

Lorsque les variables aléatoires sont ordinales, beaucoup d'utilisateurs des statistiques ont tendances à considérer les variables comme si elles étaient quantitatives. Or ceci est très abusif, et peut amener à des conclusions erronées, notamment lorsque les modalités ne sont pas équiréparties. Une solution plus correcte consiste à travailler sur les rangs associés (cf. section 3.2.1.2). L'échantillon $X_1, \ldots, X_n$ est remplacé par les rangs associés $R_1, \ldots, R_n$, où R_i est le rang de la variable X_i dans le classement par ordre croissant des variables de l'échantillon.
On utilise alors simplement comme indice de liaison entre deux variables ordinales le coefficient de corrélation linéaire entre leurs rangs, appelé **coefficient de corrélation des rangs de Spearman**.

1.4.4 Vers le cas multidimensionnel

Considérons désormais un échantillon $X_1, \ldots, X_n$ de variables aléatoires quantitatives p-dimensionnelles ($X_i = (X_i^1, \ldots, X_i^p) \in \mathbb{R}^p$). On note généralement cet échantillon sous la forme d'une matrice (ou d'un tableau) $n \times p$: $X = (X_i^j)_{1 \leq i \leq n, 1 \leq j \leq p}$.
Les covariances entre les variables sont exprimées par la **matrice de variance** Σ, de taille $p \times p$, composées des variances sur la diagonale et des covariances en dehors de la diagonale :

$$\Sigma = \frac{1}{n} Y^t Y$$

où Y est le tableau des données centrées, obtenu par $Y = AX$ avec A la matrice $n \times n$ de terme général a_{ij} vérifiant $a_{ij} = \mathbb{1}_{i=j} - 1/n$.

Propriétés de la matrice de variance :

- Σ est symétrique : $\Sigma^t = \Sigma$,
- Les valeurs propres de Σ sont positives ou nulles. Lorsqu'il n'existe aucune relation affine presque sûre entre les composantes du vecteur aléatoire, la matrice Σ est à valeurs propres strictement positives : elle est définie positive.

Chapitre 2

Estimation

Soit un échantillon $X_1, \ldots, X_n$ de variables aléatoires indépendantes et identiquement distribuées, d'espérance μ et de variance σ^2.
L'estimation statistique consiste à donner une valeur approchée à une caractéristique d'une population, à partir d'un échantillon d'observations issus de cette population. Nous nous intéressons dans un premier temps à l'estimation de paramètres de la population (espérance, variance, proportion...). Dans un second temps, nous chercherons à décrire de façon encore plus fine le comportement d'une population statistique en estimant la fonction de répartition et la densité de probabilité d'une variable aléatoire quantitative.

2.1 Préambule : étude des statistiques $\bar{X}$ et V^2

Nous avons vu dans le chapitre précédent l'intérêt des statistiques $\bar{X}$ et V^2 pour décrire la tendance centrale et la variabilité d'un échantillon $X_1, \ldots, X_n$. Nous étudions dans cette section les propriétés de ces deux statistiques.

2.1.1 Etude de la statistique $\bar{X}$

On montre facilement (*exercice*) que :

$$E[\bar{X}] = \mu \qquad \text{et} \qquad V(\bar{X}) = \frac{\sigma^2}{n}.$$

Nous verrons plus tard que la première propriété fait de $\bar{X}$ un estimateur *sans biais* de l'espérance μ de la population.
On peut montrer également que les coefficients d'asymétrie (*skewness*) et d'aplatissement (*kurtosis*) de $\bar{X}$ sont respectivement

$$\gamma_1(\bar{X}) = \frac{\gamma_1}{\sqrt{n}} \qquad \text{et} \qquad \gamma_2(\bar{X}) = 3 + \frac{\gamma_2 - 3}{n}$$

où γ_1 et γ_2 sont les coefficients d'asymétrie [1] et d'aplatissement [2] de la loi de l'échantillon.
On remarque que :

- comme $V(\bar{X}) \overset{n\to\infty}{\longrightarrow} 0$ on a $E[(\bar{X}-\mu)^2] \to 0$ et donc $\bar{X}$ converge en moyenne quadratique vers μ l'espérance de la loi de l'échantillon,
- $\gamma_1(\bar{X}) \overset{n\to\infty}{\longrightarrow} 0$ et $\gamma_2(\bar{X}) \overset{n\to\infty}{\longrightarrow} 3$ ce qui tend à penser à la normalité asymptotique de $\bar{X}$.

Enfin, l'application de la loi forte des grands nombres au cas d'un échantillon (i.i.d.) assure que

$$\bar{X} \overset{p.s.}{\longrightarrow} \mu$$

Remarque : la loi faible assure la convergence en probabilité.
Finalement, le théorème central-limite assure la la normalité asymptotique de $\bar{X}$:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\mathcal{L}}{\longrightarrow} \mathcal{N}(0,1)$$

1. le coefficient d'asymétrie ou *skewness* est définit pour une variable aléatoire X de moyenne μ et de variance σ^2 par $\gamma_1 = \frac{E[(X-\mu)^3]}{\sigma^3}$, et est nul si la loi de X est symétrique

2. le coefficient d'aplatissement ou *kurtosis* est définit par $\gamma_2 = \frac{E[(X-\mu)^4]}{\sigma^4}$, vaut 3 si la loi de X est normale et est supérieur à 3 si sa densité est plus aplatie qu'une gaussienne

Application 1 : sondage électoral

Considérons le sondage d'une population visant à déterminer la proportion p d'électeurs votant pour un certain candidat C. Nous supposons (ce qui n'est généralement pas le cas dans la réalité) que les différents sondeurs agissent indépendamment, aléatoirement et ne relève pas l'identité des personnes sondées.
Soit X_i la variable aléatoire qui vaut 1 si le sondé i déclare voter pour C et 0 sinon. Soit n le nombre de personnes interrogées. Avec ces notations, la fréquence empirique des personnes déclarant voter pour C, définie par $F = \frac{1}{n}\sum_{i=1}^{n} X_i$, n'est autre que $\bar{X}$.
Les variables $(X_1, \ldots, X_n)$ constituent un échantillon de loi de Bernoulli de paramètre p. Ainsi, si n est grand, le théorème central limite nous permet de considérer que F suit une loi normale de moyenne p et de variance $\frac{p(1-p)}{n}$.

Exercice. *On suppose que 1000 personnes sondées, 300 ont déclaré voter pour C.*
Sachant que la probabilité pour qu'une variable aléatoire de loi normale centrée réduite appartienne à $[-1.96, 1.96]$ *est de 0.95, donner un intervalle (de confiance) auquel la variable aléatoire* $\bar{X}$ *a* 95% *de chance d'appartenir.*

Réponse : $IC(p)_{95\%} = [0.2716, 0.3284]$

2.1.2 Etude de la statistique V^2

On peut montrer en *exercice* que la statistique V^2 peut s'écrire sous la forme suivante

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

La loi des grands nombres nous assure que

$$V^2 \xrightarrow{p.s.} \sigma^2,$$

mais

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

La preuve de cette dernière égalité est un exercice intéressant.
Contrairement à la statistique $\bar{X}$, V^2 sera un estimateur *biaisé* de la variance de la population : il la sous-estime légèrement. La variance de V^2 est :

$$V(V^2) = \frac{n-1}{n^3}[(n-1)\mu_4 - (n-3)\sigma^4].$$

Enfin, un théorème limite nous assure que la statistique V^2 converge en loi vers une loi normale :

$$\frac{V^2 - \frac{n-1}{n}\sigma^2}{\sqrt{V(V^2)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

A noter que lorsque $n \to \infty$, on a l'équivalence $V(V^2) \sim \frac{\mu_4 - \sigma^4}{n}$, d'où l'approximation suivante :

$$\frac{V^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Propriété 2.1.1. *La corrélation entre* $\bar{X}$ *et* V^2 *est :*

$$\rho(\bar{X}, V^2) = \frac{\mu_3}{\sigma\sqrt{\mu_4 - \frac{n-3}{n-1}\sigma^4}}$$

Démonstration en exercice (indication : on supposera sans perte de généralité que $\mu = 0$).
Ainsi, la corrélation entre $\bar{X}$ et V^2 est nulle si et seulement si $\mu_3 = 0$, ce qui est le cas des distributions symétriques. Attention, cela n'implique nécessairement pas leur indépendance.
Afin de corriger le fait que $E[V^2] \neq \sigma^2$ on utilise la statistique

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

pour exprimer la variance de l'échantillon. Ainsi, $E[S^2] = E[\frac{n}{n-1}V^2] = \sigma^2$

2.1.3 Définition des lois du χ^2, de Student et de Fisher-Snedecor

Définition 2.1.1. *Soient $U_1, \ldots, U_n$ une suite de variables aléatoires normales centrées réduites indépendantes. On appelle loi du **khi-deux** à n degrés de liberté χ^2_n la loi de la variable aléatoire $\sum_{i=1}^n U_i^2$*

L'espérance et la variance d'une variable aléatoire de loi χ^2_n sont :

$$E[\chi^2_n] = n \qquad \text{et} \qquad V(\chi^2_n) = 2n$$

La densité d'une variable aléatoire de loi χ^2_n est :

$$f(x) = \frac{x^{\frac{n}{2}-1}}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} e^{-\frac{x}{2}} \mathbb{1}_{\{x>0\}}$$

où $\Gamma(a) = \int_0^\infty e^{-x} x^{a-1}\, dx$

Définition 2.1.2. *Soient X et Y deux variables aléatoires indépendantes de lois du χ^2_n et χ^2_p. On appelle loi de **Fisher** de paramètres n et p, notée $F_{n,p}$, la loi de la variable*

$$F = \frac{\frac{X}{n}}{\frac{Y}{p}}.$$

L'espérance et la variance d'une variable aléatoire de loi $F_{n,p}$ sont :

$$E[F] = \frac{p}{p-2} \text{ pour tout } p > 2 \qquad \text{et} \qquad V(F) = \frac{2p^2(n+p-2)}{n(p-2)^2(p-4)} \text{ pour tout } p > 4.$$

Définition 2.1.3. *Soient U une variable aléatoire normale centrée réduite et X une variable aléatoire de loi du χ^2_n, indépendante de U. On appelle loi de **Student** à n degrés de liberté, notée t_n, la loi de la variable aléatoire $T_n = \frac{U}{\sqrt{\frac{X}{n}}}$*

L'espérance et la variance d'une variable aléatoire de loi t_n sont :

$$E[T_n] = 0 \text{ si } n > 1 \qquad \text{et} \qquad V(T_n) = \frac{n}{n-2} \text{ si } n > 2.$$

2.1.4 Cas des échantillons gaussiens

Lorsque l'échantillon $(X_1, \ldots, X_n)$ est issu d'une loi normale, la statistique $\bar{X}$ suit alors une loi normale en tant que combinaison linéaire de variables normales (plus besoin de théorème asymptotique).
En partant de l'égalité $X_i - \bar{X} = X_i - \mu + \mu - \bar{X}$, on peut décomposer V^2 sous la forme :

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

d'où, en multipliant par $\frac{n}{\sigma^2}$:

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

En appliquant le théorème de Cochran sur les formes quadratiques à cette décomposition, on en déduit les deux théorèmes suivants.

Théorème 2.1.1. *$(X_1, \ldots, X_n)$ est un échantillon gaussien $\Longrightarrow \frac{n}{\sigma^2} V^2 \sim \chi^2_{n-1}$.*

Théorème 2.1.2. *$\bar{X}$ et V^2 sont indépendants $\Longleftrightarrow$ $(X_1, \ldots, X_n)$ est un échantillon gaussien.*

2.2 Notion d'estimateur

Nous avons étudié au paragraphe précédent les deux statistiques $\bar{X}$ et V^2. Les lois des grands nombres nous assure que les valeurs $\bar{x}$ et v^2 de ces statistiques pour un échantillon donné sont de bonnes estimations de la moyenne μ et la variance σ^2 de la population :

$$\bar{X} \xrightarrow{p.s.} \mu \qquad \text{et} \qquad V^2 \xrightarrow{p.s.} \sigma^2$$

De même la fréquence empirique f d'un événement est une bonne estimation de sa probabilité p.
Les variables aléatoires $\bar{X}$, V^2 et F sont des **estimateurs** de μ, σ^2 et p.

Définition 2.2.1. *On appelle **estimateur** d'un paramètre θ d'une population, toute fonction*

$$T_n = f(X_1, \ldots, X_n)$$

Un estimateur est une variable aléatoire (c'est une fonction de variable aléatoire).

Il est cependant possible d'utiliser plusieurs estimateurs pour une même quantité (pour une distribution symétrique, la médiane est également un estimateur de μ). Nous allons donc présenter dans le paragraphe suivant les différentes qualités d'un estimateur qui nous guideront dans son choix.

2.3 Qualité d'un estimateur

La première qualité que l'on attend d'un estimateur est qu'il converge vers le paramètre qu'il estime, lorsque la taille de l'échantillon tend vers l'infini.

Définition 2.3.1. *Un estimateur T_n est **faiblement consistant** s'il converge en probabilité vers θ quand n tend vers l'infini*

$$\forall \epsilon > 0 \qquad \mathbb{P}(|T_n - \theta| \geq \epsilon) \xrightarrow{n \to \infty} 0$$

*Un estimateur T_n est **fortement consistant** s'il converge presque-sûrement vers θ quand n tend vers l'infini*

$$\mathbb{P}\left(\lim_{n \to \infty} T_n = \theta\right) = 1$$

Une seconde qualité est l'absence de biais d'un estimateur.

Définition 2.3.2. *On appelle **biais** d'un estimateur la quantité $E[T_n] - \theta$*

On parle alors d'estimateur sans biais, biaisé ou asymptotiquement sans biais.

Exemple. Que dire des estimateurs $\bar{X}$, V^2 et S^2 ?

On mesure également la précision d'un estimateur T_n par l'erreur quadratique moyenne $E[(T_n - \theta)^2]$, qui se décompose sous la forme

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

Ainsi, de deux estimateurs sans biais, le plus performant sera celui de variance minimale. Nous chercherons donc généralement à utiliser des estimateurs **sans biais** de **variance minimale**.

Exemple. On peut montrer que lorsque μ est connue, l'estimateur $V_\mu^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ est meilleur que S^2.

Exercice. Proposer 2 estimateurs pour le paramètre d'une loi de Poisson et déterminer le meilleur.

2.4 Estimateur exhaustif

Un échantillon $X_1, \ldots, X_n$ contient une certaine information vis-à-vis d'un paramètre inconnu θ de la population. Une statistique T_n résumant l'information contenue dans l'échantillon, il sera très important de ne pas perdre d'information : c'est cette qualité que l'on nomme l'**exhaustivité**.

Définition 2.4.1. *On appelle **vraisemblance** du paramètre θ la fonction*

$$L(x_1, \ldots, x_n, \theta) = \begin{cases} \prod_{i=1}^n f(x_i; \theta) & \text{si les } X_i \text{ sont continues} \\ \prod_{i=1}^n I\!P(X_i = x_i; \theta) & \text{si les } X_i \text{ sont discrètes} \end{cases}$$

où $f(.;\theta)$ est la densité de la variable aléatoire X_1 et $I\!P(X_i = x_i; \theta)$ est la probabilité de l'événement $\{X_i = x_i\}$ paramétrée par θ.

Soit T_n une statistique fonction de $X_1, \ldots, X_n$ de loi $g(t, \theta)$ (densité dans le cas continu, $P(T = t)$ dans le cas discret).

Définition 2.4.2. *La statistique T est exhaustive pour θ si*

$$L(x_1, \ldots, x_n, \theta) = g(t, \theta) h(x_1, \ldots, x_n).$$

En d'autre terme, elle est exhaustive si la loi de l'échantillon sachant $T = t$ ne dépend pas de θ

Ce qui signifie que si T est connue, l'échantillon n'apportera plus aucune autre information supplémentaire sur θ.

Exemple. Pour la loi normale de moyenne connue μ, la statistique $T = \sum_{i=1}^n (X_i - \mu)^2$ est exhaustive pour σ^2.

Théorème 2.4.1 (de Darmois)**.** *Soit $X_1, \ldots, X_n$ un échantillon dont le domaine de définition de la loi ne dépend pas de θ. Une condition nécessaire et suffisante pour que l'échantillon admette une statistique exhaustive est que la densité soit de la forme :*

$$f(x, \theta) = \exp[a(x)\alpha(\theta) + b(x) + \beta(\theta)]$$

Une telle densité est dite de la famille exponentielle.
Si de plus l'application $x_1 \to \sum_{i=1}^n a(x_i)$ est bijective et $\mathcal{C}^1$ alors $T = \sum_{i=1}^n a(X_i)$ est une statistique exhaustive particulière.

Exemple. Montrer que $T = ln \prod_{i=1}^n X_i$ est une statistique exhaustive pour une loi Gamma de paramètre θ inconnu, dont la densité est

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

Exercice. Donner des statistiques exhaustives pour les lois de Bernoulli, exponentielle et normale (avec soit la variance connue, soit la moyenne).

La notion d'exhaustivité renseigne sur le pouvoir d'une statistique à véhiculer l'information contenue dans un échantillon vis-à-vis d'un paramètre inconnu θ que l'on cherche à estimer. La quantité d'information sur le paramètre apportée par l'échantillon s'exprime elle par l'**information de Fisher**.

Définition 2.4.3. *On appelle quantité d'information de Fisher $I_n(\theta)$ apportée par un n-échantillon sur le paramètre θ la quantité suivante (si elle existe) :*

$$I_n(\theta) = E\left[\left(\frac{\partial ln L}{\partial \theta}\right)^2\right]$$

Théorème 2.4.2. *Si le domaine de définition de la loi de l'échantillon ne dépend pas de θ, on a :*

$$I_n(\theta) = -E\left[\frac{\partial^2 ln L}{\partial \theta^2}\right]$$

Propriété 2.4.1. *(i) Si le domaine de définition de la loi de l'échantillon ne dépend pas de θ, $I_n(\theta) = nI_1(\theta)$*

(ii) Si la loi de l'échantillon est une loi normale de variance connue, ($\theta = \mu$), alors $I_1(\theta) = \frac{1}{\sigma^2}$

(iii) en notant $I_T(\theta) = E\left[\left(\frac{\partial lng(t,\theta)}{\partial\theta}\right)^2\right]$ l'information de Fisher apportée par la statistique T, avec $g(t,\theta)$ la densité de T, on a $I_T(\theta) \leq I_n(\theta)$. On a égalité si T est exhaustive, et réciproquement si le domaine de définition de la loi de l'échantillon est indépendant de θ.

La propriété 1 dit que chaque observation a la même importance, ce qui n'est pas le cas lorsque le domaine de définition dépend de θ, comme pour une loi uniforme sur $[0, \theta]$, où la plus grande valeur de l'échantillon apporte plus d'information que les autres sur θ.
La propriété 2 nous assure l'information apportée par une observation est d'autant plus grande que la dispersion est petite.

2.5 Estimation sans biais de variance minimale

Nous avons vu précédemment que les deux qualités les plus importantes pour un estimateur étaient d'être sans biais, et de variance minimale. Il existe un certain nombre de théorèmes facilitant la recherche d'un tel estimateur.

Théorème 2.5.1 (Unicité)**.** *S'il existe un estimateur de θ sans biais de variance minimale, il est unique presque sûrement.*

Théorème 2.5.2 (Rao-Blackwell)**.** *Soit T un estimateur sans biais de θ et U une statistique exhaustive pour θ. Alors $T^* = E[T|U]$ est un estimateur sans biais de θ au moins aussi bon que T (d'un point de vue variance).*

Théorème 2.5.3. *S'il existe une statistique exhaustive U, alors l'unique estimateur T de θ sans biais de variance minimale ne dépend que de U.*

Définition 2.5.1. *Une statistique U est **complète** si $E[h(U) = 0] \quad \forall\theta \Rightarrow h = 0$p.s.*

Théorème 2.5.4 (Lehmann-Scheffé)**.** *Si T^* est un estimateur sans biais de θ dépendant d'une statistique exhaustive complète U alors T^* est l'unique estimateur sans biais de variance minimale. En particulier si l'on dispose d'un estimateur T sans biais de θ, $T^* = E[T|U]$.*

***Exemple*.** Le nombre de bug informatique par semaine d'un logiciel donné suit une loi de Poisson de paramètre λ. On cherche à évaluer la probabilité de n'avoir aucune panne pendant une semaine $P(X = 0) = e^{-\lambda}$. Que proposez-vous ?

Le résultat suivant nous indique une borne à laquelle ne peut être inférieure la variance d'un estimateur.

Théorème 2.5.5 (Inégalité de Fréchet-Darmois-Cramer-Rao)**.** *Si le domaine de définition de la loi de l'échantillon ne dépend pas de θ, tout estimateur T vérifie*

$$V(T) \geq \frac{1}{I_n(\theta)}$$

et si T est un estimateur sans biais de $h(\theta)$

$$V(T) \geq \frac{[h'(\theta)]^2}{I_n(\theta)}$$

Définition 2.5.2. *Un estimateur qui atteint la borne de Cramer-Rao est dit **efficace**. Autrement dit, un estimateur est efficace s'il n'est pas possible de trouver un estimateur sans biais de variance plus faible.*

Théorème 2.5.6 (efficacité)**.** *– la borne de Cramer-Rao ne peut être atteinte que si la loi de l'échantillon est de la famille exponentielle :*

$$f(x,\theta) = \exp[a(x)\alpha(\theta) + b(x) + \beta(\theta)]$$

– *dans ce cas il n'existe qu'une seule fonction du paramètre θ (à une transformation linéaire près) qui puisse être estimée efficacement, c'est*

$$h(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)}$$

L'estimateur de $h(\theta)$ est alors

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

et la variance minimale est

$$V(T) = \frac{h'(\theta)}{n\alpha'(\theta)}$$

Exemple. Donner un estimateur de l'écart-type d'une loi normale de moyenne connue.

La recherche d'estimateur sans biais de variance minimale passe donc par la recherche d'estimateur exhaustif. Or cette recherche peut ne pas aboutir, et elle est de plus assez lourde. La méthode du maximum de vraisemblance est une méthode systématique permettant de trouver des estimateurs.

2.6 Méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance permet de trouver des estimateurs dans toutes les situations, même les plus compliquées. C'est une des méthodes d'estimation les plus utilisées.
Cette méthode consiste à recherche le paramètre θ qui maximise la fonction de vraisemblance $L(x_1, \ldots, x_n, \theta)$, c'est-à-dire pour lequel la densité de l'échantillon est la plus grande.
L'estimateur du maximum de vraisemblance (EMV) est donc une solution de l'équation de vraisemblance

$$\frac{\partial}{\partial\theta} lnL(X_1, \ldots, X_n, \theta) = 0$$

vérifiant $\frac{\partial^2}{\partial\theta^2} lnL(X_1, \ldots, X_n, \hat{\theta}) < 0$. Un certain nombre de propriété nous prouve l'intérêt de cette estimateur.

Propriété 2.6.1. *(i) S'il existe une statistique exhaustive U, alors l'EMV en dépend.*

(ii) Si $\hat{\theta}$ est l'EMV, $f(\hat{\theta})$ est l'EMV de $f(\theta)$

(iii) Il existe une suite $\hat{\theta}_n$ de racines de l'équation de vraisemblance qui converge presque sûrement vers θ. de plus, il existe un rang à partir duquel le maximum est atteint.

(iv) $\hat{\theta}_n \xrightarrow{\mathcal{L}} \mathcal{N}(\theta, \frac{1}{I_n(\theta)})$.

La dernière propriété nous assure que l'EMV est asymptotiquement efficace. Il est donc important d'avoir un échantillon important pour utiliser cet estimateur.
Lorsque le modèle comporte plusieurs paramètres $\theta_1, \ldots, \theta_p$, il sera nécessaire de résoudre le système d'équation simultanées

$$\frac{\partial}{\partial\theta_i} lnL = 0 \qquad \forall 1 \leq i \leq p$$

Remarque 2.6.1.
– *L'équation de vraisemblance n'a pas nécessairement une unique racine.*
– *La solution de l'équation de vraisemblance n'est pas toujours calculable analytiquement. Dans ce cas, des algorithmes de recherche de maximum (de type Newton) peuvent être utilisés.*

2.7 Estimation par intervalles

Il est souvent plus intéressant de donner une estimation d'un paramètre d'intérêt sous la forme d'un intervalle, associé à une certaine probabilité d'être dans cet intervalle, plutôt que de donner une estimation ponctuelle de ce paramètre.

Exemple. Sondages électoraux.

Considérons un estimateur T de θ dont on connait la loi de probabilité. On prendra bien entendu le meilleur estimateur possible, dès lors que sa loi est connue. Connaissant la loi de T qui dépend de θ, pour une valeur estimée t de θ il est possible de déterminer un intervalle tel que :

$$P(\theta \in [t_1(t,\alpha), t_2(t,\alpha)]) = 1 - \alpha.$$

Ainsi, la vraie valeur (inconnue) du paramètre θ sera dans l'intervalle $[t_1(t,\alpha), t_2(t,\alpha)]$ avec une probabilité $1-\alpha$. On dit que $[t_1(t,\alpha), t_2(t,\alpha)]$ est un **intervalle de confiance de niveau** $1-\alpha$, que l'on note $IC_{1-\alpha}(\theta)$.
A contrario, le **risque** α est la probabilité pour que l'intervalle de confiance ne comprenne pas θ.

Remarque 2.7.1. *(i) l'intervalle de confiance est fonction de l'estimation t de θ,*

(ii) l'intervalle de confiance est également fonction de α. Plus α est petit, plus le niveau de confiance est grand, et donc plus l'intervalle s'élargit.

(iii) lorsque la taille de l'échantillon grandit, l'estimateur T étant convergeant la variance $V(T)$ diminue, et l'intervalle se rétrécit.

Soit a et b les bornes d'un intervalle de confiance $IC_{1-\alpha}(\theta)$ de niveau de confiance $1-\alpha$ pour le paramètre θ. On a :

$$p(a \leq \theta \leq b) = 1 - \alpha \text{ et donc } p(\theta < a) + p(\theta > b) = \alpha$$

En posant $\alpha = \alpha_1 + \alpha_2$, il existe une infinité de choix possibles pour α_1 et α_2, et donc de choix pour a et b. Nous ne considérerons que le cas d'un intervalle bilatéral à risques symétriques, pour lesquels le risque est partagé en deux parts égales $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$. Néanmoins, il arrive en pratique que l'on s'intéresse à des risque unilatéraux, mais nous en parlerons plus en détail dans le chapitre 3 sur les tests statistiques.

Dans la suite de ce chapitre, nous décrivons les intervalles de confiance les plus classiques. Mais il faut garder à l'esprit que ce ne sont pas les seuls, et que dès lors que l'on connait la loi de l'estimateur, il est possible de donner un intervalle de confiance.

2.7.1 Intervalle de confiance sur l'espérance

2.7.1.1 Intervalle de confiance sur l'espérance d'une loi normale avec variance connue

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ avec σ connu. Le meilleur estimateur de μ est $\bar{X}$. Comme X est de loi normale,

$$T = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$

En prenant des risques symétriques, on peut lire dans les tables les **quantiles** $u_{\frac{\alpha}{2}}$ et $u_{1-\frac{\alpha}{2}}$ de la loi normale centrée réduite d'ordres respectifs $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$, tels que :

$$\mathbf{P}(u_{\frac{\alpha}{2}} \leq T \leq u_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

ou encore

$$\mathbf{P}(T \leq u_{\frac{\alpha}{2}}) = p(T \geq u_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

La notion de quantile est définie de la façon suivante :

Définition 2.7.1. *pour une variable aléatoire continue X, le nombre q_α tel que*

$$\boldsymbol{P}(X \leq q_\alpha) = \alpha,$$

est le quantile d'ordre α de la loi de X.

Ces quantiles sont notés de différentes façons : u_α pour la loi normale, t^n_α pour la loi de Student à n degrés de liberté, χ^n_α pour la loi du χ^2_n, etc.
La figure 2.1 illustre la définition de ces quantiles.
Comme la loi normale est symétrique, on a la propriété suivante :

$$u_{1-\frac{\alpha}{2}} = -u_{\frac{\alpha}{2}}. \tag{2.1}$$

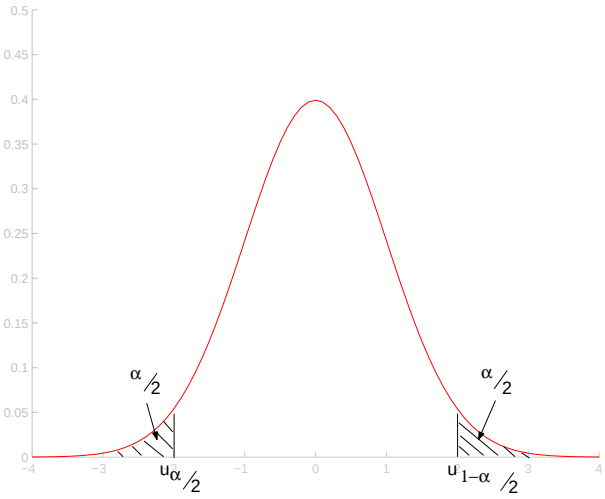

FIGURE 2.1 – quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi normale centrée réduite

Ces quantiles sont donnés par les tables statistiques. Par exemple, pour $\alpha = 0.05$, pour lequel on obtient $u_{\frac{\alpha}{2}} = -1.96$.
D'après (2.1),

$$\mathbf{P}(u_{\frac{\alpha}{2}} \leq T \leq u_{1-\frac{\alpha}{2}}) = 1-\alpha,$$

peut s'écrire

$$\mathbf{P}(u_{\frac{\alpha}{2}} \leq T \leq -u_{\frac{\alpha}{2}}) = 1-\alpha,$$

d'où on tire

$$\mathbf{P}(\bar{X} + u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha,$$

d'où l'intervalle de confiance :

$$IC_{1-\alpha}(\mu) = [\bar{X} + u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}].$$

Pour une réalisation numérique $x_1, ..., x_n$ du n-échantillon $X_1, ..., X_n$, on obtient l'intervalle de confiance sur m au niveau de confiance $1-\alpha$:

$$IC_{1-\alpha}(\mu) = [\bar{x} + u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]. \tag{2.2}$$

qui donne pour $\alpha = 0.05$:

$$[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$$

2.7.1.2 Intervalle de confiance sur l'espérance d'une loi normale avec variance inconnue

Si la variance σ^2 est inconnue, on utilise a sa place son meilleur estimateur S^2.
Comme on sait que $\frac{n}{\sigma^2}V^2$ suit une loi du χ^2 à $n-1$ degrés de liberté, $\frac{n-1}{\sigma^2}S^2$ aussi.
La statistique que l'on utilise est donc

$$T_{n-1} = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}.$$

En remarquant qu'elle s'écrit

$$T_{n-1} = \frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{n-1}{\sigma^2}S^2}{n-1}}}$$

on trouve qu'elle suit une loi de Student à $n-1$ degrés de liberté, comme rapport d'une loi normale centrée réduite sur la racine d'un χ^2 divisé par son degré de liberté.
Comme précédemment, on obtient l'intervalle de confiance :

$$IC_{1-\alpha}(\mu) = [\bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}],$$

où $t_{n-1,\frac{\alpha}{2}}$ est le quantile d'ordre $\frac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté.

2.7.1.3 Si la loi de X n'est pas une loi normale

Dans ce cas, lorsque la taille de l'échantillon n est supérieure ou égale à 30, le théorème central limite nous permet d'utiliser le fait que $\bar{X}$ suit une loi normale, et donc les résultats précédents sont applicables.

2.7.2 Intervalle de confiance sur la variance d'une loi normale

2.7.2.1 Intervalle de confiance sur la variance d'une loi normale lorsque μ est connue

Comme μ est connue, le meilleur estimateur de la variance est la statistique :

$$V_\mu^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n}.$$

Or, $\frac{\sum (X_i-\mu)^2}{\sigma^2} = \frac{n}{\sigma^2} V_\mu^2$ suit une loi du χ^2 à n degrés de liberté en tant que somme de n carrés de loi normale centrée réduite indépendantes.
Il est possible d'obtenir un intervalle de confiance sur σ^2, en fixant le niveau de confiance $1-\alpha$ dans l'inégalité :

$$\mathbf{P}(\chi^2_{n,\frac{\alpha}{2}} \leq \frac{n}{\sigma^2} V_\mu^2 \leq \chi^2_{n,1-\frac{\alpha}{2}}) = 1-\alpha,$$

où $\chi^2_{n,\frac{\alpha}{2}}$ et $\chi^2_{n,1-\frac{\alpha}{2}}$ les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi du χ^2 à n degrés de liberté.
L'intervalle est alors :

$$IC_{1-\alpha}(\sigma^2) = [\frac{nV_\mu^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \frac{nV_\mu^2}{\chi^2_{n,\frac{\alpha}{2}}}]$$

On obtient une estimation numérique de cet intervalle en remplaçant V_μ^2 par sa valeur sur le n-échantillon de X obtenu par expérience.

2.7.2.2 Intervalle de confiance sur la variance d'une loi normale lorsque μ est inconnue

Si μ est inconnue, on utilise l'estimateur de σ^2 :

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

La propriété qui nous assure que $\frac{n-1}{\sigma^2} S^2$ suit un loi du χ^2_{n-1} nous permet de construire l'intervalle de confiance :

$$IC_{1-\alpha}(\sigma^2) = [\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}],$$

et donc, en remplaçant S^2 par sa valeur s^2 sur le n-échantillon obtenu par expérience :

$$IC_{1-\alpha}(\sigma^2) = [\frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}].$$

Remarque 2.7.2. *Ces intervalles de confiance ne sont valables que pour une loi normale. Il n'est pas possible d'étendre ces résultats au cas d'autre loi comme pour les intervalles de confiance sur la moyenne.*

2.7.3 Intervalle de confiance sur une proportion

Nous supposons que la proportion p d'individus présentant un certain caractère C au sein d'une population est inconnue. Le meilleur estimateur de p est la fréquence empirique F, que l'on peut définir par :

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

où X_i est une v.a. de Bernoulli de paramètre p, définie par :

$$X_i = \begin{cases} 1 \text{ si l'individu i possède la caractère C} \\ 0 \text{ sinon.} \end{cases}$$

Comme X suit une loi de Bernoulli $\mathcal{B}(p)$, $nF = \sum_{i=1}^{n} X_i$ suit une loi binomiale $\mathcal{B}(n, p)$.
Si n est faible, on utilisera les tables de la loi binomiale (ou des *abaques*).
Si n est suffisamment grand, de sorte que $np > 5$ et $n(1-p) > 5$, on peut considérer (loi des grands nombres) que $\sum_{i=1}^{n} X_i$ suit une loi normale $\mathcal{N}(np, np(1-p))$, d'où F suit une loi normale $\mathcal{N}(p, \frac{p(1-p)}{n})$, et donc $T = \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}}$ suit une loi $\mathcal{N}(0, 1)$.
On obtient alors, en fonction des quantiles $p(u_{\frac{\alpha}{2}} \leq T \leq -u_{\frac{\alpha}{2}}) = 1 - \alpha$, l'intervalle de confiance sur p :

$$IC_{1-\alpha}(p) = [F + u_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, F - u_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}].$$

Cet intervalle recouvre p avec la probabilité $1 - \alpha$, mais il est toutefois inopérant puisque ses bornes dépendent de p. En pratique, il existe trois façons d'obtenir l'intervalle de confiance. Nous retiendrons celle qui remplace p par son estimateur F.
Ainsi, on obtient l'intervalle de confiance sur la proportion p en fonction de la valeur f de F sur notre échantillon :

$$IC_{1-\alpha}(p) = [f + u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}, f - u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}].$$

2.7.4 Récapitulatif

Intervalle de confiance d'une moyenne

$IC_{1-\alpha}(\mu)$

loi normale ou $n \geq 30$

σ^2 connue	σ^2 inconnue
$[\bar{x} + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$	$[\bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}}]$

Intervalle de confiance d'une variance

$IC_{1-\alpha}(\sigma^2)$

loi normale

μ connue	μ inconnue
$[\frac{nV_\mu^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \frac{nV_\mu^2}{\chi^{2n,\frac{\alpha}{2}}}]$	$[\frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}]$

Intervalle de confiance d'une proportion

$IC_{1-\alpha}(p)$

$np > 5$ et $n(1-p) > 5$

$[f + u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}, f - u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}]$

2.8 Plus d'estimation statistique

2.8.1 Estimation bayésienne

Le point de vue bayésien suppose que les paramètres θ de la loi des observations $X_1, \ldots, X_n$ sont également des variables aléatoires.
La densité $g(\theta)$ de θ est la loi *a priori* de θ.
La densité *conditionnelle* des observations X_i sachant θ est $f(x_i|\theta)$.
La vraisemblance (conditionnelle) est $L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^n f(x_i|\theta)$.
La loi conjointe des observations et du paramètre $(X_1, \ldots, X_n, \theta)$ est

$$f(x_1, \ldots, x_n, \theta) = L(x_1, \ldots, x_n, \theta)g(\theta).$$

On définit également la loi *a posteriori* du paramètre θ connaissant les observations :

$$g(\theta|X_1 = x_1, \ldots, X_n = x_n) = \frac{L(x_1, \ldots, x_n, \theta)g(\theta)}{\int_{\mathbb{R}} L(x_1, \ldots, x_n, \theta)g(\theta)d\theta}.$$

2.8.1.1 Application : estimation bayésienne de la moyenne d'une loi normale de variance connue

On suppose que la loi de l'échantillon conditionnellement à μ est $\mathcal{N}(\mu, \sigma^2)$, et que la loi a priori de μ est également une loi normale $\mathcal{N}(\mu_0, \sigma_0^2)$.
Le calcul de la loi a posteriori donne une loi normale d'espérance et de variance :

$$E[\theta|X_1, \ldots, X_n] = \frac{\frac{\sigma^2}{n}\mu_0 + \sigma_0^2\bar{X}}{\frac{\sigma^2}{n} + \sigma_0^2} \quad \text{et} \quad V(\theta|X_1, \ldots, X_n) = \frac{\frac{\sigma^2\sigma_0^2}{n}}{\frac{\sigma^2}{n} + \sigma_0^2}$$

L'estimateur bayésien de μ, qui est l'espérance a posteriori est donc une moyenne pondérée de l'espérance a priori et de la moyenne empirique des observations.
Introduisons le concept de **précision**, comme l'inverse de la variance. La précision a priori sur μ est $\eta_1 = \frac{1}{\sigma_0^2}$ et sur la moyenne empirique elle est $\eta_2 = \frac{n}{\sigma^2}$. On voit alors que $E[\theta|X_1, \ldots, X_n] = \frac{\eta_1\mu_0 + \eta_2\bar{X}}{\eta_1 + \eta_2}$ et $\frac{1}{V(\theta|X_1, \ldots, X_n)} = \eta_1 + \eta_2$.
L'estimateur bayésien de μ est donc la moyenne pondérée des deux estimations (a priori et empirique) pondérées par leur précision. Si l'information a priori est très précise, les observations n'auront que peu d'influence dans l'estimateur bayésien. Au contraire si la précision a priori tend vers 0 ou si n tend vers l'infini, l'estimateur bayésien est l'estimateur classique $\bar{X}$.

Cette application fonctionne très bien car la loi a posteriori se calcule facilement. Mais pour des lois quelconques, les calculs sont généralement beaucoup plus compliqués, et la loi a posteriori doit être estimée par des algorithmes spécifiques.
La statistique bayésienne peut être vu comme un raffinement de la statistique classique, mais le choix de la loi a priori peut être très problématique et reste toujours subjectif. Néanmoins, pour les problèmes statistique dans lesquels on dispose de peu de données (fiabilité de systèmes très rarement défaillant par exemple), l'incorporation d'une information a priori (≪ jugement d'expert ≫) peut s'avérer très intéressante.

2.8.2 Estimation robuste : cas de la valeur centrale d'une distribution symétrique

L'estimation $\bar{x}$ de l'espérance μ d'une distribution symétrique est très sensibles à des valeurs extrêmes ≪ aberrantes ≫.
Lorsque des valeurs aberrantes sont présentes (ou soupçonnées), un estimateur robuste de l'espérance peut être utilisé : la **moyenne tronquée** d'ordre α, qui est la moyenne arithmétique obtenue en éliminant de l'échantillon les αn plus grandes et plus petites valeurs. Une valeur généralement recommandée est $\alpha = 15\%$.
La médiane est le cas extrême de cet estimateur pour $\alpha = 50\%$, et est très robuste.
Au lieu d'éliminer les αn plus grandes valeurs, il est également possible de toutes les fixer à la plus grande valeur conservées : c'est ce qu'on appelle la ≪ winzorization ≫.

D'autres approches existent également, comme celle des M-estimateurs, qui consistent à chercher une estimation μ qui minimise une fonction du type

$$\sum_{i=1}^n h\left(\frac{x_i - \mu}{s}\right)$$

où s est une estimation robuste de la dispersion. Toute une famille d'estimateur est ainsi définie en fonction du choix de h. Pour $h(x) = -lnf(x)$, avec f la densité des données, on retrouve les estimateurs du maximum de vraisemblance.

2.9 Estimation fonctionnelle

2.9.1 Estimation de la fonction de répartition

La fonction de répartition empirique, introduite section 1.3.3.3 et définie comme la proportion des n variables $X_1, \ldots, X_n$ inférieures ou égales à x :

$$F_n(x) = \frac{\#\{X_i : X_i \leq x, 1 \leq i \leq n\}}{n} \tag{2.3}$$

est un estimateur de la fonction de répartition $F(x) = p(X \leq t)$.
C'est une variable aléatoire, en tant que fonction des variables aléatoires $X_1, \ldots, X_n$. A un échantillon d'observations $x_1, \ldots, x_n$ correspond une réalisation de cette fonction aléatoire, qui est une fonction en escalier de sauts $1/n$.

Théorème 2.9.1 (Glivenko-Cantelli)**.** *Soit F_n la fonction de répartition empirique d'un échantillon $(X_1, \ldots, X_n)$ où les X_i ont pour fonction de répartition F. Alors*
- $\forall x \in \mathbb{R}, \quad F_n(x) \xrightarrow{p.s.} F(x)$
- $||F_n - F||_\infty \xrightarrow{p.s.} 0$

Preuve. Le premier point est démontré en cours, le second point est admis. Pour un rappel sur les différents modes de convergence d'une suite de variables aléatoires, se reporter à l'annexe 4.1.

Le second point de ce théorème nous assure que pour une taille assez grande d'échantillon, la fonction de répartition théorique peut être approximée par la fonction de répartition empirique.

2.9.2 Estimation non paramétrique de la densité

Pour aller plus loin se référer à [1].
La connaissance de la densité d'une variable aléatoire donne une information très importante. Nous avons vu qu'un premier estimateur de la densité de probabilité pouvait être l'histogramme (section 1.3.3.2). L'histogramme est un graphique en bâtons, dont la hauteur pour une classe j est proportionnelle à la proportion de point observé dans cette classe $\frac{n_j}{n}$ (où n_j est le nombre de points dans la classe et n est le nombre de points total). Si la longueur de l'intervalle vaut h, la hauteur est alors $\frac{n_j}{n}\frac{1}{h}$, de sorte à ce que l'air totale des bâtons soit égale à 1. Cet estimateur discontinue s'améliore lorsque l'on fait tendre vers 0 la largeur h de chaque intervalle, et que l'on fait tendre vers l'infini le nombre de points par classe. Mais en pratique le nombre de points est fini, et cet estimateur discontinu n'est pas le meilleur estimateur pour une fonction continue.
Nous présentons ici une méthode d'estimation fonctionnelle plus évoluée, qui permet, en l'absence de toute hypothèse de modèle paramétrique donné, une estimation point par point de la densité de probabilité.
On cherche une estimation $\hat{f}_n$ de la densité f minimisant l'erreur quadratique moyenne intégrée :

$$MISE = E\left[\int_{\mathbb{R}} (\hat{f}_n(x) - f(x))^2 \, dx\right].$$

Soit $X_1 \leq \ldots \leq X_n$ un échantillon, rangé dans l'ordre croissant, de la variable aléatoire dont on cherche à estimer la densité. Sachant que la fonction de densité est la dérivée de la fonction de répartition, on a

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h},$$

et on peut donc approcher f, pour de petite valeur de h par

$$f_n(x) \simeq \frac{F(x+h) - F(x-h)}{2h} \simeq \frac{F_n(x+h) - F_n(x-h)}{2h}$$

où F_n est la fonction de répartition empirique. En remplaçant F_n par son expression (2.3), on obtient l'estimateur par **fenêtre mobile** de la densité

$$f_n(x) \simeq \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} \mathbb{1}_{[-1,1]} \left(\frac{X_i - x}{h} \right).$$

Cet estimateur se généralise à l'estimateur par la **méthode du noyau** de Parzen

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x - x_i}{h} \right)$$

où K est une fonction noyau, définie de $\mathbb{R} \to \mathbb{R}^+$ et d'intégrale égale à 1.
Il existe différents types de noyau, parmi lesquels :

- uniforme (ci-dessus) : $K(x) = \frac{1}{2} \mathbb{1}_{[-1,1]}(x)$,
- gaussien : $K(x) = \frac{1}{\sqrt{2\pi}} \exp^{-x^2/2}$,
- triangulaire : $K(x) = (|x| + 1) \mathbb{1}_{[-1,1]}$,
- Epanechnikov : $K(x) = 3/4(1 - x^2) \mathbb{1}_{[-1,1]}$.

Le choix du noyau n'est pas d'une importance capitale, au contraire du choix de la taille de la fenêtre h : plus h est petit, plus les fluctuations sont importantes, plus h est grand, plus le lissage est important. Tout l'intérêt sera de trouver le meilleur compromis. On recommande généralement le choix de $h = s_n n^{-1/5}$ où s_n est l'écart-type estimé des observations.

Propriétés des estimateurs à noyau $\hat{f}_n$

- estimateur asymptotiquement sans biais : $\lim_{n\to\infty} E[\hat{f}_n(x)] = f(x)$ pour tout $x \in \mathbb{R}$
- $V(\hat{f}_n(x)) \to 0$ si $h \to 0$ et $hn \to \infty$ (h tend vers 0 moins vite que $1/n$)
- vitesse de convergence en $n^{-4/5}$:

$$E[(\hat{f}_n(x) - f(x))^2] \leq cste \times n^{-4/5},$$

qui est la vitesse optimale pour les estimateurs non-paramétriques, mais qui est plus faible que la vitesse typique des méthodes paramétriques, généralement n^{-1}.

*Logiciel : l'estimation par noyau se fait sous le logiciel **R** à l'aide de la fonction* `density`.

Chapitre 3

Tests statistiques

On distingue différentes catégories de tests :
- les tests **paramétriques** ont pour objet de tester une certaine hypothèse relative à un ou plusieurs paramètres d'une variable aléatoire de loi spécifiée (généralement supposée normale). Lorsque le test est toujours valide pour des variables non gaussiennes, on dit que le test est robuste (à la loi).
- les tests **non paramétriques** qui portent généralement sur la fonction de répartition de la variable aléatoire, sa densité...
- les tests **libres** (*distributions free*) qui ne supposent rien sur la loi de probabilité de la variable aléatoire étudiée (et qui sont donc robuste). Ces tests sont souvent non paramétriques, mais pas toujours.

Dans ce cours, nous classons les tests en fonction de leur fonctionnalité :
- Tests sur une population :
  - test sur le caractère centrale d'une population,
  - test sur la variance,
  - test sur une proportion,
  - test de l'aléatoire d'un échantillon,
  - test d'ajustement à une loi spécifiée,
  - test de liaison entre variables (quantitatives, qualitatives, mixtes)
- Tests de comparaison de deux populations

3.1 Théorie des tests paramétriques

3.1.1 Introduction : test sur l'espérance d'une loi normale de variance connue

Soit un échantillon $(X_1, ..., X_n)$ de loi $\mathcal{N}(\mu, \sigma^2)$, avec μ inconnue et σ^2 connue. On cherche à tester si l'espérance μ est égale ou non à une valeur de référence μ_0 :

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

Sous l'hypothèse H_0, la statistique suivante suit une loi $\mathcal{N}(0, 1)$

$$T = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

Ainsi, si H_0 est vraie, la valeur de cette statistique pour l'échantillon observé devrait appartenir à l'intervalle $[u_{\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}]$ avec la probabilité $1 - \alpha$. Ce qui revient à dire que la réalisation de $\bar{X}$ appartient à l'intervalle

$$[\mu_0 + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \mu_0 + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

avec une probabilité de $1 - \alpha$.
Ainsi, si l'observation $\bar{x}$ de $\bar{X}$ n'est pas dans cet intervalle on peut décider de rejeter l'hypothèse H_0. Le risque de se tromper en rejetant H_0 est α.

3.1.2 Vocabulaire des tests

Un test est un procédé qui permet de trancher entre deux hypothèses, au vu des résultats d'un échantillon : on teste une **hypothèse nulle** contre une **hypothèse alternative**. L'hypothèse nulle H_0 est l'hypothèse que l'on veut contrôler. Elle est toujours de forme **simple**

$$H_0 : \theta = \theta_0$$

où θ_0 est une valeur donnée du paramètre. Le choix de cette hypothèse est fait de manière *conservative* : si on test un médicament, on prendra H_0 l'hypothèse où le médicament n'a pas d'effet. C'est également souvent la plus importante des deux hypothèses puisque c'est celle dont on contrôle le risque. L'hypothèse alternative H_1 est quant à elle généralement **composite** :

$$H_1 : \theta \in \Theta_1$$

où Θ_1 est une partie de $\mathbb{R}$ non nécessairement réduite à un élément. Cette hypothèse se ramène souvent à un des cas suivants : $\theta < \theta_0$, $\theta > \theta_0$ (test unilatéraux) ou $\theta \neq \theta_0$ (test bilatéral).

Suivant la justesse de la décision prise à l'issue du test, on est en présence de 4 cas de figure (tableau 3.1).

Décision \ Vérité	H_0	H_1
H_0	conclusion correcte	erreur de deuxième espèce
H_1	erreur de première espèce	conclusion correcte

TABLE 3.1 – Erreurs associés à un test

Exemple (Importance du choix des hypothèses)**.** Considérons le test des hypothèses suivantes :
- hypothèse H_0 : le patient doit être hospitalisé,
- hypothèse alternative H_1 : le patient ne doit pas être hospitalisé.

L'erreur de première espèce consiste à ne pas hospitaliser un patient qui en avait besoin. Cette erreur est très grave, puisqu'elle peut conduire au décès du patient. Le risque de deuxième espèce, qui consiste à hospitaliser un patient qui n'en avait pas besoin peut s'avérer moins grave.
Pour l'exemple du médicament, l'erreur de première espèce consiste à mettre sur le marché un médicament qui n'a pas d'effet.

3.1.3 Probabilité d'erreur et risque, puissance de test

On associe aux erreurs de première et deuxième espèces les probabilités (**risques**) associées (tableau 3.2). Le **niveau de confiance** du test est la probabilité $1 - \alpha$ de ne pas rejeter à raison H_0. Le risque de première espèce α est le risque de rejeter H_0 à tort. Le risque de deuxième espèce β est le risque de conserver H_0 à tort.

Décision \ Vérité	H_0	H_1
H_0	niveau de confiance $1 - \alpha$	risque β
H_1	risque α	$1 - \beta$

TABLE 3.2 – Risques associés à un test

En pratique il est d'usage de **fixer le risque** α : 5%, 1%, 10%. Ainsi, on contrôle le risque associé à l'erreur de première espèce, qui nous l'avons vu est l'erreur la plus grave. Choisir un risque α trop petit va conduire à ne rejeter que très rarement H_0 (si on ne la rejette pas on ne risque pas de la rejeter à tort !). Au contraire, choisir un risque trop grand va conduire à n'accepter que très rarement α.
Le risque β se déduit alors par le calcul, si la loi sous H_1 est connue. Il varie en sens contraire de α. Ainsi, en diminuant le risque α, on augmente le risque β. On définit alors la **puissance du test** par $1 - \beta$, qui correspond à la probabilité de rejeter H_0 à raison.
Le choix d'un test sera donc le résultat d'un compromis entre risque de premier espèce et puissance du test.

Une fois que l'on a fixé raisonnablement α, il faut choisir une **variable de décision**, qui doit apporté le maximum d'information sur le problème posé, et dont la loi sera différente selon que H_0 ou H_1 est vraie. La loi sous H_0 doit être connue. On définit alors la **région critique** W qui est l'ensemble des valeurs de la variable de décision qui conduisent à rejeter H_0 au profit de H_1. Sa forme est déterminée par la nature de H_1, et sa détermination exacte est donnée par $p(W|H_0) = \alpha$. La **région d'acceptation** est son complémentaire $\bar{W}$.

3.1.4 Choix optimal de la statistique de test et de la région de rejet

Le choix de la statistique de test et de la région de rejet est fait de sorte à maximiser la puissance du test $1 - \beta$ pour un risque de première espèce α fixé.
Plaçons nous dans le cadre d'un test entre hypothèses simples :

$$H_0 : \theta = \theta_0 \text{ contre } H_1 : \theta = \theta_1$$

Neyman et Pearson (1933) ont montré que le **test du rapport de vraisemblance** est le test le plus puissant au niveau de confiance α.

Théorème 3.1.1 (Neyman et Pearson)**.** *La région critique optimale est définie par les points* $\mathbf{x} = (x_1, \ldots, x_n)$ *vérifiant*

$$W = \{\mathbf{x} : \frac{L(\mathbf{x}, \theta_1)}{L(\mathbf{x}, \theta_0)} > c_\alpha\}$$

La constante c_α, qui dépend de α, est déterminée par $\alpha = \mathbf{P}_{\theta_0}(\mathbf{x} \in W)$.

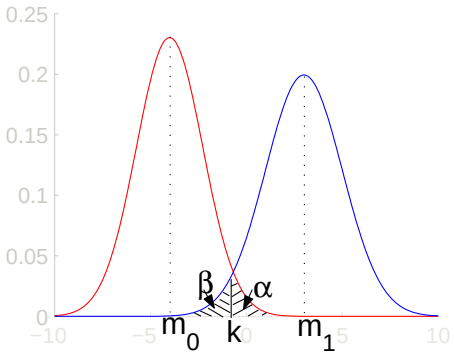

FIGURE 3.1 – illustration de la règle de décision

Exemple. Reprenons le test d'introduction, où $(X_1, ..., X_n)$ est de loi normale de variance σ^2 connue et d'espérance μ inconnue, avec cette fois une hypothèse alternative simple :

$$H_0 : \mu = \mu_0 \quad \text{contre} \quad H_1 : \mu = \mu_1.$$

On suppose $\mu_0 < \mu_1$. La vraisemblance de l'échantillon gaussien s'écrit

$$L(\mathbf{x}, \mu) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2}$$

d'où le rapport de vraisemblance

$$\frac{L(\mathbf{x}, \theta_1)}{L(\mathbf{x}, \theta_0)} = \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^n 2(\mu_1 - \mu_0)x_i - \frac{n}{2\sigma^2}(\mu_1^2 - \mu_0^2)\right)$$

Ainsi, $\frac{L(\mathbf{x},\theta_1)}{L(\mathbf{x},\theta_0)} > c_\alpha$ est équivalent à $\bar{x} > \log(c_\alpha)\frac{\sigma^2}{n(\mu_1-\mu_0)} + \frac{\mu_1+\mu_0}{2} = C$, où la constante C est déterminée $\mathbf{P}_{\mu_0}(\mathbf{x} \in W) = \mathbf{P}_{\mu_0}(\bar{x} > C) = \alpha$. La région critique optimale du test de Neyman-Pearson est donc

$$W = \{\mathbf{x} : \bar{x} > \mu_0 + u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\}$$

et on retombe bien sur le test ≪ intuitif ≫de l'introduction.

Dans le cas où l'hypothèse alternative est composite ($\theta \in \Theta_1$), la puissance du test est fonction de $\theta : 1 - \beta(\theta)$ est appelée la **fonction puissance du test**.
Un test est dit **uniformément le plus puissant** (UPP) si quelque soit la valeur de θ appartenant à l'hypothèse alternative, sa puissance est supérieure à celle de tout autre test.

Exemple. On a vu précédemment pour le test $H_0 : \mu = \mu_0$ contre $H_1 : \mu = \mu_1 > \mu_0$ que la région critique ne dépend pas de μ_1, et qu'elle est donc la même pour tout $\mu_1 > \mu_0$. Le test est donc UPP pour $H_0 : \mu = \mu_0$ contre $H_1 : \mu > \mu_0$.
Si cette fois $\mu_1 < \mu_0$, on obtient encore un test UPP $H_0 : \mu = \mu_0$ contre $H_1 : \mu < \mu_0$, mais différent du précédent. Il n'existe donc pas de test UPP pour $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$.

3.1.5 Utilisation de la puissance de test

Dans le cas d'un test entre deux hypothèses simples avec variance σ^2 connue

$$H_0 : \mu = \mu_0 \quad \text{contre} \quad H_1 : \mu = \mu_0 + \delta,$$

(avec $\delta > 0$), nous avons vu que la région critique avait la forme

$$W = \{\mathbf{x} : \bar{x} > \mu_0 + u_{1-\alpha}\frac{\sigma}{\sqrt{n}}\}.$$

On peut calculer le risque de second espèce :

$$\beta = p(\text{décider } H_0 | H_1) = \Phi(u_{1-\alpha} - \frac{\delta\sqrt{n}}{\sigma}).$$

La puissance du test, $1 - \beta$, est donc fonction de α, n et δ. En considérant α et n fixés, on peut représenter la courbe de puissance du test par la Figure (3.2).

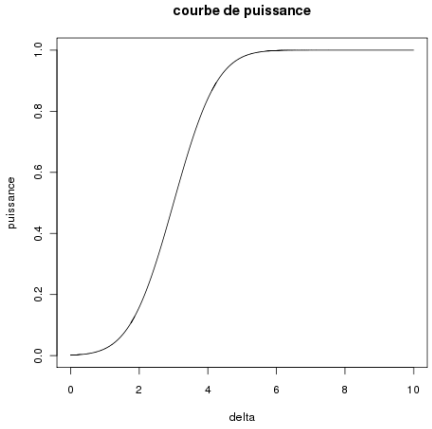

FIGURE 3.2 – Courbe de puissance d'un test

La courbe de puissance peut donc permettre
– de **choisir entre plusieurs tests** en fonction de leur courbes de puissance (que l'on veut la plus forte possible, i.e. proche de la droite d'ordonnée 1),
– pour un problème donné, dans lequel α et δ sont fixés, on pourra **choisir le nombre de sujets nécessaire** n pour atteindre une puissance donnée à l'aide de l'équation (3.1).

3.1.6 Résumé

La démarche de construction d'un test est la suivante :
– choix de H_0 et H_1,
– détermination de la variable de décision,

- allure de la région critique en fonction de H_1,
- calcul de la région critique en fonction de α,
- calcul de la valeur expérimentale de la variable de décision,
- conclusion : rejet ou acceptation de H_0.

3.1.7 p-value

En pratique, plutôt que de calculer la région critique en fonction de α, on préfère donner un seuil critique α^*, appelée **p-value**, qui est la plus grande valeur de α conduisant à ne pas rejeter H_0. Cette information permet au lecteur de conclure à l'acceptation de H_0 pour tout risque de première espèce $\alpha \leq \alpha^*$, et à son rejet pour tout $\alpha > \alpha^*$.

3.2 Tests sur une population

Nous pouvons maintenant présenter les différents tests statistiques classiques, obtenus par la méthode de Neyman-Pearson lorsque les échantillons sont gaussiens (voir de grandes tailles). Dans le cas de petits échantillons non gaussiens, des alternatives non paramétriques seront présentées.

3.2.1 Test sur le caractère central d'une population

3.2.1.1 Cas d'un échantillon grand ou gaussien

Soit un n-échantillon $(X_1, ..., X_n)$ issu d'une population de moyenne μ et de variance σ^2. Nous supposons que au moins l'une des deux conditions suivantes est satisfaite :

- la population est de loi normale,
- l'échantillon est de taille n suffisamment grande ($n \geq 30$).

Test $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$ lorsque σ^2 est connue La statistique de test est

$$U = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

Sous H_0, cette statistique suit une loi normale centrée réduite d'après les conditions précédentes (via le théorème centrale limite si seule la seconde condition est satisfaite).
La région critique, définie par $|U| > k$, se traduit par $|\bar{X} - \mu_0| > -u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, où $u_{\frac{\alpha}{2}}$ est le quantile de la loi normale centrée réduite d'ordre $\frac{\alpha}{2}$.
Ainsi,

$$\boxed{\text{on rejette } H_0 \text{ si } |\bar{x} - \mu_0| > -u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.}$$

Remarque 3.2.1 (Calcul de la **p-value**). *Pour ce test, on rejette H_0 dès que $\frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}} > -u_{\frac{\alpha}{2}}$. La p-value est la valeur critique α^* de α telle que $\frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}} = -u_{\frac{\alpha^*}{2}}$, d'où $\alpha^* = 2\Phi\left(-\frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}}\right)$ avec Φ la fonction de répartition de la loi normale centrée réduite. Ainsi, dès que l'on choisi un risque α plus grand que α^*, on a $-u_{\frac{\alpha^*}{2}} > -u_{\frac{\alpha}{2}}$ et donc on rejette H_0. Au contraire, si le risque est plus petit, on aura cette fois $\frac{|\bar{x}-\mu_0|}{\frac{\sigma}{\sqrt{n}}} = -u_{\frac{\alpha^*}{2}} < -u_{\frac{\alpha}{2}}$ et on conserve H_0.*

Remarque 3.2.2 (Tests unilatéraux). *Si le test est unilatéral, $H_0 : \mu = \mu_0$ contre $H_1 : \mu < \mu_0$, on rejette H_0 si la vraie valeur de μ est trop éloignée inférieurement de μ_0, ce qui se traduit par $\bar{x} < \mu_0 + u_\alpha \frac{\sigma}{\sqrt{n}}$.*
Si le test est $H_0 : \mu = \mu_0$ contre $H_1 : \mu > \mu_0$, on rejette H_0 si $\bar{x} > \mu_0 - u_\alpha \frac{\sigma}{\sqrt{n}}$.

Test $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$ lorsque σ^2 est inconnue Ce test est généralement connu sous le nom de **test de Student**.
Dans ce cas la variance σ^2 est estimée par son estimateur S^2. La statistique de test est

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

qui suit une loi de Student à $n-1$ degré de liberté.
La conclusion du test devient alors

$$\boxed{\text{on rejette } H_0 \text{ si } |\bar{x} - \mu_0| > -t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}},}$$

où $t_{n-1,\frac{\alpha}{2}}$ est le quantile d'ordre $\frac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté, et $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$.

*Logiciel **R** : les tests sur la moyenne s'effectuent à l'aide de la fonction* `t.test`.
*Logiciel **SAS** :* `proc ttest` *ou* `proc univariate`.
Attention : seul des test bilatéraux sont possibles sous SAS. Dans le cas d'un test unilatéral, il conviendra donc d'ajuster la p-value (en la divisant par deux), et de s'assurer avant de rejeter H_0 que la statistique de test est bien du bon coté de l'hypothèse nulle.

3.2.1.2 Cas d'un petit échantillon non gaussien

Le caractère central de la population sera testé cette fois, non plus en travaillant sur l'espérance de la loi comme précédemment, mais en testant la symétrie de la distribution par rapport à une valeur μ_0 d'intérêt. Nous supposons, sans perte de généralité, que $\mu_0 = 0$.
Les hypothèses que nous testons sont donc :
- $H_0 : F(x) = 1 - F(-x)$ la distribution est symétrique par rapport à 0
- contre $H_1 : F(x + \delta) = 1 - F(\delta - x)$ la distribution est symétrique par rapport à δ

où F est la fonction de répartition de la variable aléatoire testée.
Les tests que nous allons présenter dans cette section seront basés sur les rangs des observations et nécessitent quelques notions introduites dans le paragraphe suivant.

Statistique de rang
Rang et anti-rang. Soit $X = (X_1, \ldots, X_n)$ un échantillon. Soit R_i la variable aléatoire égale au **rang** de la variable X_i dans le classement dans l'ordre croissant des variables $X_1, \ldots, X_n$ (on ne suppose pas d'ex-æquo).
On appelle **anti-rang**, D_i l'indice de la la variable classée en $i$ème position.

Exemple : pour $X = (3.2, 6.4, 2.1, 4.5)$ on a $R = (2, 4, 1, 3)$ et $D = (3, 1, 4, 2)$.

Remarque : les vecteurs des rangs R et des anti-rangs D sont tous deux des permutations des n premiers entiers. De plus, R et D sont des permutations inverses : $R = D^{-1}$.

La suite des rangs $R = (R_1, \ldots, R_n)$ est donc une suite de variable aléatoire identiquement distribuées mais non indépendantes. On a pour tout $1 \leq i \leq n$:

$$E[R_i] = \frac{n+1}{2} \qquad V(R_i) = \frac{n^2 - 1}{12}$$

Cas des ex-æquo : lorsque plusieurs variables sont ex-æquo, on leur associe généralement le rang moyen des rangs partagés par ces variables. Par exemple, si on a 4 variables ex-æquo avec 5 autres variables plus petites et 4 plus grandes, elles partageront les rangs 6, 7, 8 et 9 et on leur associera donc le rang moyen 7.5.
Tous les test basés sur les statistiques de rangs présentés dans ce cours supposent l'absence d'ex-æquo. Dans le cas contraire, les tests doivent subir des modifications, qui ne seront pas abordées dans ce cours, sauf pour le test de Wilcoxon de comparaison de deux échantillons (cf. section 3.3.2.1).

Statistique de rangs signés. On appelle **rang signé** R_i^+ de la variable X_i le rang de $|X_i|$ dans le classement des $|X_1|, \ldots, |X_n|$ par ordre croissant.

Nous serons par la suite amenés à travailler avec différentes statistiques de test associées aux rangs signés, définie par

$$S = \sum_{i=1}^{n} a(R_i^+) \mathbb{1}_{X_i \geq 0}$$

où a est une fonction de $\{1, 2, \ldots, n\}$ dans $\mathbb{R}$.

Définition 3.2.1. *Une variable aléatoire a une distribution symétrique par rapport à μ_0 si pour tout $x \in \mathbb{R}$:*

$$p(X \leq \mu_0 + x) = p(X \geq \mu_0 - x)$$

Sous l'hypothèse d'une distribution symétrique par rapport à 0, on a

$$E[S] = \sum_{i=1}^{n} a(i)/2 \qquad V(S) = \sum_{i=1}^{n} a^2(i)/4.$$

Lorsque n est grand le théorème central limite nous permet de considérer que S est distribué suivant une loi normale.
Lorsque n est petit, la statistique S a été tabulée pour différentes fonctions a.

Nous présentons ci-après trois tests basés sur trois choix de la fonction a.

Test des rangs signés (Wilcoxon à un échantillon) Pour le test des rangs signés, il faut supprimer de l'échantillon les valeurs nulles. On choisit ensuite $a(i) = i$ et la statistique de test devient

$$W^+ = \sum_{i=1}^{n^*} R_i^+ \mathbb{1}_{X_i \geq 0}$$

ou n^* est le nombre de valeurs non nulles de l'échantillon. Cette statistique admet comme espérance et variance sous H_0 :

$$E_{H_0}[W^+] = n(n+1)/4 \qquad V_{H_0}(W^+) = n(n+1)(2n+1)/24.$$

A noter qu'en présence d'ex-æquo, l'espérance est identique mais la variance est différente.
Si la taille d'échantillon n est suffisamment grande, on rejetera H_0 si $\frac{|W^+ - E_{H_0}[W^+]|}{\sqrt{V_{H_0}(W^+)}} > u_{1-\frac{\alpha}{2}}$.
Si n est petit, on utilisera les tables statistiques dédiées à ce test (Annexe 4.2.1). Ces tables donne, pour un risque α de 5% et 1%, les quantiles de la statistique de Wilcoxon d'ordre $\alpha/2$ et $1 - \alpha/2$. Ces tables sont toujours valables en présence d'ex-æquo.
La même démarche sera appliquée pour les deux tests suivants.
*Logiciel **R** : fonction `wilcox.test`.*
*Logiciel **SAS** : **proc univariate**. Attention, SAS utilise une statistique de test W^+ centrée.*

Test du signe Pour le test du signe, il faut supprimer de l'échantillon les valeurs nulles. On choisit ensuite $a(i) = 1$ et la statistique de test devient

$$S^+ = \sum_{i=1}^{n^*} \mathbb{1}_{X_i > 0}$$

ou n^* est le nombre de valeurs non nulles de l'échantillon. La statistique S^+, qui est le nombre de valeurs positives dans l'échantillon, suit, sous l'hypothèse H_0 de symétrie par rapport à 0, une loi binomiale de paramètre n et $1/2$. On peut donc facilement déduire la p-value correspondant à la valeur observée sur l'échantillon de la statistique S^+. Ces p-values ont été tabulée et figurent en Annexe 4.2.2.
En outre, l'espérance et la variance de S^+ sous H_0 sont :

$$E_{H_0}[S^+] = n/2 \qquad V_{H_0}(S^+) = n/4.$$

Ce test est plus puissant que le test de Wilcoxon lorsque les queues de distributions sont très diffuses.
Remarquons enfin que la présence d'ex-æquo ne pose aucun problème pour ce test.

*Logiciel **R** : fonction `SIGN.test` du package `BSDA`.*
*Logiciel **SAS** : **proc univariate**. Attention, SAS utilise une statistique de test S^+ centrée.*

Test des scores normaux En choisissant $a(i) = \Phi^{-1}\left(\frac{i}{n+1}\right)$ la statistique de test devient

$$SN^+ = \sum_{i=1}^{n} \Phi^{-1}\left(R_i^+/(n+1)\right) \mathbb{1}_{X_i \geq 0}$$

qui admet comme espérance et variance sous H_0 :

$$E_{H_0}[SN^+] = \sum_{i=1}^{n} \Phi^{-1}\left(i/(n+1)\right)/2 \qquad V_{H_0}(SN^+) = \sum_{i=1}^{n} \left(\Phi^{-1}(i/(n+1))\right)^2/4.$$

Ce test est particulièrement intéressant pour les distributions très concentrées.

*Logiciel **R** : test à implémenter.*

3.2.2 Test sur la variance d'une population gaussienne

Soit un n-échantillon $(X_1, ..., X_n)$ issu d'une population de loi normale, de moyenne μ et de variance σ^2. La normalité est indispensable pour ce test sur la variance.

3.2.2.1 Test $H_0 : \sigma^2 = \sigma_0^2$ contre $H_1 : \sigma^2 \neq \sigma_0^2$, moyenne μ connue

Lorsque la moyenne est connue, la statistique V_μ^2 est la meilleure estimation de la variance (cf. exercice en TD) :

$$V_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Sous l'hypothèse H_0, comme l'échantillon est gaussien, $\frac{n}{\sigma_0^2}V_\mu^2$ suit une loi du χ_n^2 (en tant que somme de carrés de $\mathcal{N}(0,1)$). Ainsi,

$$\boxed{\text{on rejette } H_0 \text{ si } V_\mu^2 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2 < \tfrac{\sigma_0^2}{n}\chi^2_{n,\frac{\alpha}{2}} \text{ où si } V_\mu^2 > \tfrac{\sigma_0^2}{n}\chi^2_{n,1-\frac{\alpha}{2}},}$$

où $\chi^2_{n,\frac{\alpha}{2}}$ et $\chi^2_{n,1-\frac{\alpha}{2}}$ sont les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi de χ^2 à n degrés de liberté. Attention, contrairement à la loi de Student et à la loi normale, la loi du χ^2 n'est pas symétrique.

3.2.2.2 Test $H_0 : \sigma^2 = \sigma_0^2$ contre $H_1 : \sigma^2 \neq \sigma_0^2$, moyenne μ inconnue

Lorsque la moyenne est inconnue, on la remplace par son estimateur $\bar{X}$. La variance est alors estimée par $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2$ et la statistique du test

$$\frac{n-1}{\sigma_0^2}S^2$$

suit sous H_0 une loi du χ^2 à $n-1$ degrés de liberté.
La conclusion du test est alors la suivante :

$$\boxed{\text{on rejette } H_0 \text{ si } S^2 = \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2 < \tfrac{\sigma_0^2}{n-1}\chi^2_{n-1,\frac{\alpha}{2}} \text{ ou si } S^2 > \tfrac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\frac{\alpha}{2}}.}$$

3.2.2.3 Tests unilatéraux sur la variance

Test $H_0 : \sigma^2 = \sigma_0^2$ **contre** $H_1 : \sigma^2 > \sigma_0^2$
- si la moyenne μ est connue, on rejette H_0 si $V_\mu^2 > \frac{\sigma_0^2}{n}\chi^2_{n,1-\alpha}$.
- si la moyenne μ est inconnue, on rejette H_0 si $S^2 > \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\alpha}$.

Test $H_0 : \sigma^2 = \sigma_0^2$ **contre** $H_1 : \sigma^2 < \sigma_0^2$
- si la moyenne μ est connue, on rejette H_0 si $V_\mu^2 < \frac{\sigma_0^2}{n}\chi^2_{n,\alpha}$.
- si la moyenne μ est inconnue, on rejette H_0 si $S^2 < \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\alpha}$.

3.2.3 Test sur une proportion pour un grand échantillon

Dans la population étudiée, une proportion p des individus possèdent un certain caractère C. On se propose de comparer cette proportion p à une valeur de référence p_0.
On considère un échantillon d'individus de taille n de cette population. La variable aléatoire X_i égale à 1 si l'individu i possède le caractère C suit une loi de Bernoulli $\mathcal{B}(p)$, et le nombre d'individus $\sum_{i=1}^n X_i$ possédant ce caractère suit une loi binomiale $\mathcal{B}(n,p)$.
Si n est suffisamment grand, de sorte que $np > 5$ et $n(1-p) > 5$, on peut considérer (loi des grands nombres) que $\sum_{i=1}^n X_i$ suit une loi normale $\mathcal{N}(np, np(1-p))$, d'où la fréquence empirique $F = \frac{1}{n}\sum_{i=1}^n X_i$ suit une loi normale $\mathcal{N}(p, \frac{p(1-p)}{n})$. Si n est trop petit, le test est construit sur la loi binomiale, et on peut utiliser les *abaques*.

3.2.3.1 Test $H_0 : p = p_0$ contre $H_1 : p \neq p_0$

La statistique du test est donc la fréquence empirique F qui suit sous H_0 une loi $\mathcal{N}(p_0, \frac{p_0(1-p_0)}{n})$.

$$\boxed{\text{on rejette } H_0 \text{ si } |f - p_0| > u_{1-\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0)}{n}}.}$$

3.2.3.2 Tests unilatéraux sur une proportion

Test $H_0 : p = p_0$ **contre** $H_1 : p > p_0$ On rejette H_0 si $f > -u_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} + p_0$.
Test $H_0 : p = p_0$ **contre** $H_1 : p < p_0$ On rejette H_0 si $f < u_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} + p_0$.

***Exemple*.** Sur un échantillon de 200 individus d'une commune, 45% sont favorables à l'implantation d'un centre commercial. Ceci contredit-il l'hypothèse qu'un habitant sur deux y est favorable ?
On test $H_0 : p = 0.5$ contre $H_1 : p \neq 0.5$ avec un risque $\alpha = 0.05$, d'où $u_{1-\frac{\alpha}{2}} = 1.96$. On rejette H_0 si $|f - 0.5| > 1.96\sqrt{\frac{0.5^2}{200}} \simeq 0.07$, or ici $|f - 0.5| = 0.05$ donc on ne rejette pas H_0, un habitant sur deux est bien favorable à l'implantation du centre commercial.

3.2.4 Test de l'aléatoire d'un échantillon

Étant donné une suite de variables aléatoires $X_1, \ldots, X_n$ nous cherchons à déterminer si cette suite est un échantillon indépendant et identiquement distribué. Nous testons pour cela
- $H_0 : X_1, \ldots, X_n$ indépendant et identiquement distribué,
- contre $H_1 : X_i = f(i) + \epsilon_i$ avec f une tendance monotone, ϵ_i i.i.d centrées.

3.2.4.1 Test de corrélation des rangs de Spearman

Une première façon de tester les hypothèses précédentes est de tester s'il existe une corrélation significative entre les rangs $R_1, \ldots, R_n$ associés à l'échantillon et la suite $1, \ldots, n$. La statistique de test est le coefficient de corrélation des rangs de Spearman

$$R_S = \frac{\sum_{i=1}^n (R_i - \bar{R})(i - \bar{i})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (i - \bar{i})^2}}$$

avec $\bar{R} = \bar{i} = (n+1)/2$ et $\sum_{i=1}^{n}(i-\bar{i})^2 = n(n^2-1)/12$.
En remarquant que

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

on voit que la statistique de test R_S sera égale à -1 dans le cas d'une tendance décroissante ($R_i = n+1-i$) et à 1 pour une tendance croissante ($R_i = i$).
On peut montrer que cette statistique admet les moments suivant :

$$E[R_S] = 0 \qquad V(R_S) = \frac{1}{n-1}.$$

Sous l'hypothèse H_0

- si $n \geq 30$, on utilise la statistique $R_S\sqrt{n-1}$ qui suit une $\mathcal{N}(0,1)$,
- si $10 < n < 30$, on utilise la statistique $R_S\sqrt{\frac{n-2}{1-R_S^2}}$ qui est approximativement distribuée selon une $\sim t_{n-2}$.

*Logiciel **R** : fonction `cor.test` avec option `spearman`.*

3.2.4.2 Test des changements de signes

Dans le cas où l'on veut tester plus qu'une dépendance monotone (par exemple croissance puis décroissance), on peut utiliser la statistique de test :

$$S = \#\{i : R_i > R_{i+1}, 1 \leq i < n\}$$

qui suit une loi normale d'espérance $\frac{n-1}{2}$ et de variance $\frac{n+1}{12}$.

3.2.5 Tests d'ajustement à une loi de probabilité spécifiée

Les tests d'ajustement ont pour but de vérifier si un échantillon provient ou non d'une certaine loi de probabilité spécifiée. Nous allons dans un premier temps présenter quelques méthodes empiriques qui permettent de s'orienter vers une distribution, puis nous présenterons deux tests : le test du χ^2 et le test de Kolmogorov-Smirnov.

3.2.5.1 Quelques méthodes empiriques

La forme de l'histogramme La forme de l'histogramme construit sur l'échantillon de données peut nous aider à avoir une idée de la distribution de la variable aléatoire dont il est issu. Par exemple, un histogramme symétrique nous orientera par exemple vers une loi normale, de Cauchy, de Student...

La nature du phénomène Suivant le phénomène étudié, il sera possible d'orienter son choix. Si on s'intéresse à une variable de comptage, on pourra penser à une loi de Poisson, pour une durée de vie on pensera à une loi exponentielle ou à une loi de Weibull... .

Utilisation des moments On sait que pour une loi de Poisson, la moyenne est égale à la variance. Pour une loi exponentielle la moyenne est égale à l'écart-type. Pour une loi normale le coefficient d'aplatissement (kurtosis) est égal à 3 et le coefficient d'asymétrie (skewness) est nul.

3.2.5.2 Ajustement graphiques

Pour un certain nombre de lois de probabilité, une transformation fonctionnelle permet de représenter la courbe de la fonction de répartition par une droite :

Loi exponentielle Pour $X \sim \mathcal{E}(\lambda)$, on a $p(X > x) = \exp(-\lambda x)$ d'où $\ln(1 - F(x)) = -\lambda x$. En rangeant dans l'ordre croissant les données x_i de l'échantillon, l'estimation de la fonction de répartition qu'est la fonction de répartition empirique s'écrit $F_e(x) = \frac{\text{effectif} \leq x_i}{n} = \frac{i-1}{n}$ pour $x_i \leq x \leq x_{i+1}$. Ainsi, les points de coordonnées $\left(x_i; \log(1 - \frac{i-1}{n})\right)$ sont approximativement alignés le long d'une droite dont la pente fournit une estimation graphique de λ.

Loi normale Si X est une variable gaussienne de moyenne μ et de variance σ^2 :

$$\mathbf{P}(X \leq x) = \Phi(\frac{x-\mu}{\sigma})$$

où Φ est la fonction de répartition de la loi normale centrée réduite.
Pour chaque valeur x_i de la variable X, on peut approcher $\mathbf{P}(X \leq x_i)$ empiriquement par $\frac{i}{n}$ (en ayant classé l'échantillon par ordre croissant), et en déduire le quantile u_i d'ordre $\mathbf{P}(X \leq x_i)$ tel que $\Phi(u_i) = \mathbf{P}(X \leq x_i)$.
Si la variable est gaussienne, les points de coordonnées (x_i, u_i) sont alignés sur la droite d'équation $u = \frac{x-\mu}{\sigma}$, appelée droite de Henry. On compare donc les valeurs des quantiles de la loi empirique x_i aux quantiles de la loi normale centrée réduite u_i.
*Logiciel **R** : la fonction $qqnorm$ permet de représenter la droite de Henry, et $qqplot$ généralise à d'autres lois que la loi normale.*

3.2.5.3 Test d'ajustement du χ^2

Soit une variable aléatoire X discrète ou discrétisée, c'est à dire divisée en K classes de probabilités $p_1, p_2, ..., p_K$ sous une certaine loi $\mathcal{L}(\theta)$.
Soit un échantillon de cette variable fournissant les **effectifs empiriques** aléatoires $N_1, N_2, ..., N_K$ dans chacune de ces classes. Ces effectifs empiriques N_i sont des variables aléatoires d'espérance np_i. Nous appellerons **effectifs théoriques** les quantités np_i.
Le test du χ^2 a pour but de tester :

$$H_0 : \text{X suit la loi de probabilité } \mathcal{L}(\theta),$$

et consiste à comparer les effectifs théoriques et empiriques.
Pour cela on introduit la variable D^2 définie par :

$$D^2 = \sum_{i=1}^{K} \frac{(N_i - np_i)^2}{np_i},$$

et qui est asymptotiquement distribué, lorsque $n \to \infty$, comme une loi du χ^2 à $K-1$ degrés de liberté.
La variable D^2 pouvant être interprétée comme une mesure de l'écart aléatoire entre les effectifs empirique et théorique, le test du χ^2 consiste à rejeter H_0 si la valeur d^2 de D^2 sur l'échantillon est trop grande :

$$\boxed{\text{on rejette } H_0 \text{ si } d^2 > \chi^2_{K-1,1-\alpha}.}$$

Si des estimations sont nécessaires
Pour faire le test du χ^2, il est nécessaire de savoir quelle est la loi à tester, c'est-à-dire quelle est sa nature (normale, Poisson...), mais aussi quels sont ses paramètres. Il est donc souvent nécessaire d'estimer ces paramètres.
Par exemple, pour tester une hypothèse de normalité, on teste la loi $\mathcal{N}(\bar{x}, s^2)$, où $\bar{x}$ et s^2 sont les estimations des paramètres de la loi. Soit l le nombre d'estimations indépendantes effectuées.
Le nombre de degrés de liberté du χ^2 utilisé dans le test devra alors être $K - l - 1$.

Effectif minimal d'une classe
La propriété qui assure que D^2 suit une loi du χ^2 suppose que chaque classe a un effectif théorique np_i supérieur à 5. Lors de la construction du test, cette propriété sera à vérifier. Souvent lorsque l'expérience conduit la création des classes, certaines classes "extrêmes" ne vérifient pas cette propriété. On regroupera alors les classes entre elles afin de créer des classes plus importantes qui vérifient cette propriété (en regroupant la classe extrême avec celle qui lui est contigüe, et ainsi de suite...).
Il ne faudra pas oublier alors d'affecter au nombre de classes K sa nouvelle valeur dans la détermination du nombre de degrés de liberté du χ^2.

*Logiciel **R** : le test du χ^2 peut être réalisé à l'aide de la fonction* `chisq.test`.

3.2.5.4 Test de Kolmogorov-Smirnov

Le test du χ^2 convient très bien aux variables discrètes, qui ne nécessitent aucune discrétisation. Par contre, lorsque les variables sont continues, on préfère généralement utiliser le test de Kolmogorov-Smirnov.
L'adéquation à une loi donnée porte cette fois sur les fonctions de répartition :
- $H_0 : F(x) = F_0(x)$ pour tout $x \in \mathbb{R}$
- contre $H_1 : \exists x \in \mathbb{R}, \quad F(x) \neq F_0(x)$

La statistique de test utilisée est

$$KS = \max_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

où $\hat{F}_n(x) = \#\{X_i : X_i \leq x\}/n$ est la fonction de répartition empirique estimée à partir de l'échantillon $X_1, \ldots, X_n$.
Il existe alors des tables de cette statistique KS sur lesquelles se baser pour conduire à rejeter ou non H_0.

*Logiciel **R** : le test de Kolmogorov-Smirnov peut être réalisé à l'aide de la fonction* `ks.test`.

3.2.5.5 Test de Shapiro-Wilk (normalité)

Le test de Shapiro-Wilk est le test le plus recommandé pour tester la normalité d'une série de données. Il est particulièrement puissant pour les petits effectifs.
Supposons les X_i rangés par ordre croissant.La statistique du test s'écrit :

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_i\right)^2}{\sum_{i=1}^{n}(X_i - \bar{(X)})^2} = \frac{\left(\sum_{i=1}^{\left[\frac{n}{2}\right]} a_{n+1-i}(X_{n+1-i} - X_i)\right)^2}{\sum_{i=1}^{n}(X_i - \bar{(X)})^2}$$

où
- $\left[\frac{n}{2}\right]$ est la partie entière de $\frac{n}{2}$,
- a_i sont des constantes fournies dans des tables spécifiques (Annexe 4.2.4),

$$(a_1, \ldots, a_n) = \frac{m^t V^{-1}}{(m^t V^{-1} V^{-1} n)^2}$$

où $m = (m_1, \ldots, m_n)^t$ sont les espérances des statistiques d'ordre d'un échantillon de variables indépendantes et identiquement distribuée suivant une loi normale, et V est la matrice de variance-covariance de ces statistiques d'ordre.

La statistique W peut donc être interprétée comme le coefficient de détermination entre la série des quantiles générés à partir de la loi normale et les quantiles empiriques obtenus à partir des données. Plus W est élevé, plus la compatibilité avec la loi normale est crédible. La région critique, rejet de la normalité, s'écrit :

on rejette la normalité si $W < w_{\alpha,n}$,

la valeur critique $w_{\alpha,n}$ étant lue dans les tables de Shapiro-Wilk (Annexe 4.2.4) en fonction du risque de première espèce α et de la taille d'échantillon n.
*Logiciel **R** : le test de Shapiro-Wilk peut être réalisé à l'aide de la fonction* `shapiro.test`.

3.2.6 Test d'indépendance entre deux variables aléatoires

3.2.6.1 Cas de deux variables aléatoires quantitatives

Test de corrélation linéaire Le coefficient de corrélation linéaire ρ_{XY} entre deux variables continues X et Y, introduit au chapitre 1, est défini par :

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Son estimateur est

$$R_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

La statistique suivante

$$T = \sqrt{n-2}\frac{R_{XY}}{\sqrt{1-R_{XY}^2}}$$

qui suit une loi de Student t_{n-2} permet de tester la nullité du coefficient de corrélation linéaire, en rejetant l'hypothèse nulle $\rho_{XY} = 0$ si la valeur t de cette statistique est trop grande ou trop petite, autrement dit si elle vérifie :

$$t > t_{n-2,1-\frac{\alpha}{2}} \qquad \text{ou} \qquad t < t_{n-2,\frac{\alpha}{2}}.$$

Il conviendra donc de tester la nullité de ce coefficient de corrélation linéaire avant de tenter de modéliser Y en fonction de X par une relation linéaire.

*Logiciel **R** : fonction* `cor.test`.

Test de corrélation des rangs de Spearman Un indicateur de corrélation entre deux variables quantitatives plus robuste aux valeurs extrêmes, est le coefficient de corrélation des rangs de Spearman, défini comme le coefficient de corrélation linéaire entre les rangs associés aux variables testées. Ce test, déjà présenté dans la section 3.2.4.1, permet également de tester la corrélation entre des variables ordinales.

3.2.6.2 Cas de deux variables aléatoires qualitatives : Test du χ^2

Ce test découle du test d'ajustement du χ^2. Soient X et Y deux variables aléatoires qualitatives pouvant prendre respectivement k et r modalités. Les données sont présentées dans un tableau de contingence :

X Y	modalité 1	modalité 2	...	modalité r	total
modalité 1	n_{11}	n_{12}		n_{1r}	$n_{1.}$
modalité 2	n_{21}	n_{22}		n_{2r}	$n_{2.}$
$\vdots$					
modalité k	n_{11}	n_{12}		n_{1r}	$n_{1.}$
total	$n_{.1}$	$n_{.2}$		$n_{.r}$	n

où

- n_{ij} est le nombre d'individus ayant la modalité i de X et la modalité j de Y,
- $n_{i.} = \sum_{j=1}^{r} n_{ij}$ est le nombre total d'individus ayant la modalité i de X ,
- $n_{.j} = \sum_{i=1}^{k} n_{ij}$ est le nombre total d'individus ayant la modalité j de Y,
- $n = \sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}$ est le nombre d'individus total.

Le test consiste à tester H_0 : « les deux variables sont indépendantes ».
Si H_0 est vrai, cela a un sens de considérer les probabilités $p_1^X, \ldots, p_k^X$ d'avoir les modalités $1, \ldots, k$ de la variable X et les probabilités $p_1^Y, \ldots, p_r^Y$ d'avoir les modalités $1, \ldots, r$ de la variable Y.
Le test consiste, comme pour le test d'ajustement, à comparer les effectifs empiriques n_{ij} aux effectifs théoriques $p_i^X p_j^Y$ que l'on devrait observer si X et Y étaient indépendantes. Les p_i^X et p_j^Y étant inconnues, on les estime par $\hat{p}_i^X = \frac{n_{i.}}{n}$ et $\hat{p}_j^Y = \frac{n_{.j}}{n}$.
On construit alors la mesure d'écart suivante :

$$d^2 = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} = n\left(\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right)$$

qui est la réalisation d'une statistique dont la loi peut être approximée par une loi de χ^2 à $(k-1)(r-1)$ degrés de liberté, lorsque les effectifs sont de tailles suffisantes ($\frac{n_{i.}n_{.j}}{n} > 5$ pour tout i, j).
Le test consiste donc à rejeter H_0 si d^2 est trop grand, comme pour un test d'ajustement du χ^2.

3.2.6.3 Cas de deux variables aléatoires binaires et de petits échantillons : Test exact de Fisher

Dans le cas d'échantillons de petites tailles (effectifs théoriques inférieurs à 5 par croisement de variables), une alternative consiste à utiliser le test exact de Fisher.

Lorsque les variables sont binaires, sous l'hypothèse H_0 d'indépendance de X et Y, la probabilité d'observer l'effectif n_{11} est donnée :

$$\mathbf{P}(N_{11} = n_{11}|n_{1.}, n_{2.}, n_{.1}, n_{.2}) = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{21}!n_{21}!n_{22}!} = \frac{C_{n_{1.}}^{n_{11}} C_{n_{2.}}^{n_{21}}}{C_n^{n_{.1}}}.$$

On reconnait une variable aléatoire de loi Hypergéométrique (tirage de n individus parmi n dont). Le test peut donc être construit de façon exacte en utilisant cette loi.
Ce test est généralisable à plus de deux modalités par variable.
*Logiciel **R** : fonction* `fisher.test`.

3.2.6.4 Cas d'une variable qualitative et d'une variable quantitative : ANOVA à 1 facteur

Soient X une variable quantitative que l'on observe pour différentes modalités (*niveaux*) d'une variable qualitative A (*facteur*). On dispose de K échantillons indépendants de X de tailles n_1 à n_K correspondant chacun à un niveau différent du facteur A :

- $X_1^1, X_1^2, \ldots, X_1^{n_1}$ correspondant au niveau A_1 du facteur A,
- $X_2^1, X_2^2, \ldots, X_2^{n_2}$ correspondant au niveau A_2 du facteur A,
- ...
- $X_K^1, X_K^2, \ldots, X_K^{n_K}$ correspondant au niveau A_K du facteur A.

On suppose que le facteur A influe uniquement sur la moyenne des échantillons et non sur leur dispersion. Ainsi, chaque échantillon est supposé suivre une **loi normale** $\mathcal{N}(\mu_k, \sigma^2)$.
Le problème est donc de tester

$$H_0 : \mu_1 = \ldots = \mu_K = \mu \qquad \text{contre } H_1 : \exists 1 \leq i, j \leq K \text{ t.q. } \mu_i \neq \mu_j.$$

Pour cela on appelle $\bar{X}_k$ la moyenne empirique de l'échantillon k et $\bar{X}$ la moyenne empirique globale :

$$\bar{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_k^i \quad \text{et} \quad \bar{X} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} X_k^i,$$

où $n = \sum_{k=1}^K n_k$.
En remarquant que $X_k^i - \bar{X} = X_k^i - \bar{X}_k + \bar{X}_k - \bar{X}$, on montre facilement la **formule d'analyse de variance** :

$$\underbrace{\frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_k^i - \bar{X})^2}_{V_T^2} = \underbrace{\frac{1}{n} \sum_{k=1}^{K} n_k (\bar{X}_k - \bar{X})^2}_{V_A^2} + \underbrace{\frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_k^i - \bar{X}_k)^2}_{V_R^2}$$

qui représente la décomposition de la variance totale V_T^2 en la **variance** V_A^2 **due au facteur** A (variance **inter**-groupe) plus la **variance résiduelle** V_R^2 (ou variance **intra**-groupe).

Remarque 3.2.3. *Cette formule est l'équivalente empirique de la formule vue en cours de probabilité :*

$$V(X) = E[V(X|A)] + V(E[X|A]).$$

En remarquant que $V_R^2 = \frac{1}{n} \sum_{k=1}^K n_k V_k^2$ où $V_k^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} (X_k^i - \bar{X}_k)^2$, on montre que $\frac{n}{\sigma^2} V_R^2 = \sum_{k=1}^K \frac{n_k V_k^2}{\sigma^2}$ suit une loi du χ^2 à $n - K$ degrés de liberté, car chaque $\frac{n_k V_k^2}{\sigma^2}$ suit une loi du χ^2 à $n_k - 1$ degrés de liberté.
De même, sous H_0 cette fois, $\frac{nV_T^2}{\sigma^2}$ suit une loi du χ^2 à $n - 1$ degrés de liberté (car V_T^2 est la variance d'un n-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$) et $\frac{nV_A^2}{\sigma^2}$ suit une loi du χ^2 à $K - 1$ degrés de liberté (car V_A^2 peut être vue comme la variance du K-échantillon $(\bar{X}_1, \ldots, \bar{X}_K)$).
L'équation de l'analyse de variance revient alors à $\chi_{n-1}^2 = \chi_{K-1}^2 + \chi_{n-K}^2$, ce qui permet en outre de conclure via le théorème de Cochran que V_A^2 et V_R^2 sont indépendantes.
La statistique du test est donc

$$F = \frac{\frac{V_A^2}{K-1}}{\frac{V_R^2}{n-K}}$$

qui suit sous H_0 une loi de Fisher-Snedecor $F_{K-1,n-K}$, et on rejette l'hypothèse H_0 si la statistique F est supérieure au quantile de la loi $F_{K-1,n-K}$ d'ordre $1-\alpha$.

*Logiciel **R** : fonction* `aov`.

Test de l'homogénéité des variances : test de Levene. En plus de la normalité des échantillons, dont on peut se passer si les échantillons sont de tailles suffisantes, nous avons supposé que les variances étaient homogènes ($\sigma_1 = \ldots = \sigma_K$).
Le test de Levene permet de tester cette hypothèse. La statistique de ce test est la suivante :

$$L = \frac{n-K}{K-1} \frac{\sum_{k=1}^{K} (\bar{Z}_k - \bar{Z})^2}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} (Z_k^i - \bar{Z}_k)^2},$$

où

$$Z_k^i = |X_k^i - \bar{X}_k|, \quad \bar{Z}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} Z_k^i \quad \text{et} \quad \bar{Z} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Z_k^i.$$

Sous l'hypothèse $H_0 : \sigma_1 = \ldots = \sigma_K$, cette statistique suit une loi de Fisher-Snedecor $F_{K-1,n-K}$. Nous rejetons donc l'hypothèse H_0 si la statistique L est supérieure au quantile de la loi $F_{K-1,n-K}$ d'ordre $1-\alpha$.

*Logiciel **R** : fonction* `levene.test` *du package* `lawstat`.

Comparaison des moyennes deux à deux
Rejeter H_0 permet de dire que toutes les moyennes ne sont pas égales. Il peut cependant être intéressant de tester l'égalité des moyennes deux à deux.
Pour cela, on effectue un test de comparaison multiple des moyennes (pour $1 \leq k, k' \leq K$) :

$$H_0 : \mu_k = \mu_{k'}.$$

Un résultat dû à Scheffé montre que

$$p\left(|\bar{X}_k - \bar{X}_{k'} - (\mu_k - \mu_{k'})| \leq S_R \sqrt{(K-1) f_{K-1,n-K,1-\alpha}} \sqrt{\frac{1}{n_k} + \frac{1}{n_{k'}}} \right) = 1 - \alpha$$

où $f_{K-1,n-K,1-\alpha}$ est le quantile de la loi de Fisher de paramètres $K-1$ et $n-K$ d'ordre $1-\alpha$.
On rejette donc l'hypothèse d'égalité des moyennes μ_k et $\mu_{k'}$ si

$$|\bar{X}_k - \bar{X}_{k'}| > S_R \sqrt{(K-1) f_{K-1,n-K,1-\alpha}} \sqrt{\frac{1}{n_k} + \frac{1}{n_{k'}}}.$$

Remarque. *Attention, l'égalité des moyennes n'est pas transitive.*

3.3 Tests de comparaison de deux populations indépendantes

L'objectif de cette section est de dire si deux échantillons indépendants sont issus d'une même population ou non. Voici quelques exemples d'application :

- les rendements journaliers de deux usines d'un même groupe sont-ils semblables ?
- les ventes par semaine de deux actions sont-elles similaires ?

On formule le problème de la façon suivante : on observe deux échantillons $(X_{1,1}, ..., X_{1,n_1})$ et $(X_{2,1}, ..., X_{2,n_2})$, indépendants et de fonctions de répartition $F_1(x)$ et $F_2(x)$. Le test exact revient à tester l'égalité de ces fonctions de répartitions :

$$H_0 : F_1(x) = F_2(x) \text{ contre } H_1 : F_1(x) \neq F_2(x).$$

Nous verrons dans un premier temps des tests paramétriques qui, sous l'hypothèse de normalité des échantillons (ou de grandes tailles), consistent à tester l'égalité des variances et des espérances des deux populations. Dans un second temps, lorsque les échantillons sont de petites tailles nous présenterons des alternatives non paramétriques.

3.3.1 Cas de deux échantillons gaussiens ou de grandes tailles

Supposons dans un premier temps que les deux échantillons sont gaussiens.
Si les **variances sont connues**, ce qui n'arrive que rarement en pratique, la statistique de test utilisée pour tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$ repose sur la différence entre les estimateurs des moyennes des deux échantillons :

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

qui suit, sous H_0, une loi normale centrée réduite.
Ainsi, on rejettera H_0 si

$$|\bar{x}_1 - \bar{x}_2| > -u_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Dans le cas le plus courant, les **variances sont inconnues**. On doit alors tester dans un premier temps si elles sont égales ou non (**test de Fisher**) avant de pouvoir effectuer le test de comparaison des moyennes (**test de Student**).

3.3.1.1 Test de comparaison des variances de Fisher

Nous testons

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ contre } H_1 : \sigma_1^2 \neq \sigma_2^2.$$

D'après les résultats de la théorie de l'échantillonnage :

$$\frac{n_1 V_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1} \quad \text{et} \quad \frac{n_2 V_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}.$$

Ainsi, sous l'hypothèse H_0 que $\sigma_1^2 = \sigma_2^2$, la statistique du test F suivante suit une loi de Fisher F_{n_1-1,n_2-1} :

$$F = \frac{\frac{n_1 V_1^2}{n_1-1}}{\frac{n_2 V_2^2}{n_2-1}} = \frac{S_1^2}{S_2^2} \tag{3.1}$$

Cette variable de décision s'interprète comme le rapport des estimateurs de σ_1^2 et σ_2^2. Elle doit donc ne pas être trop différentes de 1 si H_0 est vérifiée. En pratique on met toujours au numérateur la plus grande des deux quantités, ou autrement dit on suppose que $S_1^2 > S_2^2$ (sinon on permute les indices).
La région de rejet sera donc de la forme $F > k$ avec k plus grand que 1 :

$$\boxed{\text{on rejette } H_0 \text{ si } \frac{\frac{n_1 V_1^2}{n_1-1}}{\frac{n_2 V_2^2}{n_2-1}} > f_{n_1-1,n_2-1,1-\alpha},}$$

où $f_{n_1-1,n_2-1,1-\alpha}$ est le quantile de la loi de Fisher-Snedecor F_{n_1-1,n_2-1} d'ordre $1-\alpha$.

3.3.1.2 Test de comparaison des moyennes de Student avec variances égales

Nous testons

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

en supposant les variances égales $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
On a pour $i = 1, 2$:

$$\frac{n_i V_i^2}{\sigma^2} \sim \chi^2_{n_i-1} \quad \text{et} \quad \bar{X}_i \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{n_i}).$$

Ainsi, la statistique

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 V_1^2 + n_2 V_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté. D'où la conclusion :

on rejette H_0 si $|\bar{x}_1 - \bar{x}_2| > -t_{n_1+n_2-2,\frac{\alpha}{2}} \sqrt{\frac{n_1 v_1^2 + n_2 v_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$.

Remarque 3.3.1 (Tests unilatéraux de comparaison de moyennes)**.** *Le test unilatéral* $H_0 : \mu_1 = \mu_2$ *contre* $H_1 : \mu_1 < \mu_2$, *conduit au rejet de* H_0 *si* $\bar{x}_1 - \bar{x}_2 < t_{n_1+n_2-2,\alpha} \sqrt{\frac{n_1 v_1^2 + n_2 v_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

3.3.1.3 Test de comparaison des moyennes avec variances différentes

Lorsque les échantillons sont de grandes tailles (> 30), le test de Student reste encore approximativement valable.
Pour de petits échantillons gaussiens, l'approximation d'Aspin-Welch consiste à utiliser le test de Student avec un degré de liberté non plus égal à $n_1 + n_2 - 2$ mais égal à l'entier le plus proche de :

$$n = \frac{1}{\frac{c^2}{n_1-1} + \frac{(1-c)^2}{n_2-1}} \quad \text{où } c = \frac{\frac{v_1^2}{n_1-1}}{\frac{v_1^2}{n_1-1} + \frac{v_2^2}{n_2-1}}$$

3.3.1.4 Échantillons non gaussiens

Théoriquement, le test de la variance de Fisher n'est plus valable car la statistique $\frac{nV^2}{\sigma^2}$ ne suit plus une loi du χ^2. Néanmoins, le test de comparaison de moyennes de Student étant relativement robuste à un changement dans la loi des échantillons, il est possible de l'utiliser pour comparer les moyennes des deux échantillons, que les variances soit égales ou non, si les tailles d'échantillons sont suffisamment grandes (au minimum 30 observations par échantillon).

3.3.2 Échantillons de petites tailles

Lorsque les échantillons ne sont pas suffisamment grands pour permettre une utilisation du test de Student, on utilise des alternatives non paramétriques, qui ont pour but de tester :

$$H_0 : F_1(x) = F_2(x) \text{ contre } H_1 : F_1(x) \neq F_2(x)$$

où $F_1(x)$ et $F_2(x)$ sont les fonctions de répartition de deux échantillons $(X_{1,1}, \ldots, X_{1,n_1})$ et $(X_{2,1}, \ldots, X_{2,n_2})$.
Dans cette section nous concaténons les deux échantillons en un seul $(X_1, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_{n_1+n_2})$, et nous allons travailler avec les rangs $(R_1, \ldots, R_{n_1+n_2})$ associés à cet échantillon global.
Les statistiques de test utilisées seront de la forme

$$S = \sum_{i=1}^{n_1} a(R_i)$$

où a est une fonction de $\{1, \ldots, n_1 + n_2\}$ dans $\mathbb{R}$. A noter que seuls les rangs du premier échantillon sont utilisés dans la statistique S puisque la somme s'arrête à n_1.
Lorsque les tailles d'échantillons n_1 et n_2 sont petites (< 30), il existe des tables suivant la fonction a choisie (Wilcoxon, médiane, scores normaux). Lorsque les tailles sont plus grandes (cas dans lequel les tests paramétriques sont également utilisables), la statistique S est approximativement distribuée suivant une loi normale.
Les moments de S sont :

$$E[S] = \frac{n_1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} a(i) \qquad V(S) = \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)} \sum_{i=1}^{n_1+n_2} (a(i) - \bar{a})^2$$

où $\bar{a} = \frac{1}{n_1+n_2} \sum_{i=1}^{n_1+n_2} a(i)$

3.3.2.1 Test de Wilcoxon

On supposera ici que $n_1 \leq n_2$. En choisissant $a(i) = i$ la statistique de test devient

$$W = \sum_{i=1}^{n_1} R_i$$

et correspond à la somme des rangs du premier échantillon (le plus petit en nombre d'observations).

$$E_{H_0}[W] = \frac{n_1(n_1+n_2+1)}{2}$$
$$V_{H_0}(W) = \frac{n_1 n_2(n_1+n_2+1)}{12}$$

La loi de cette statistique a été tabulée pour de petites tailles d'échantillons (moins de 10), et la table en Annexe 4.2.3 donne les bornes critiques de W pour des risques de première espèce de 5% et 1%.
Pour de plus grandes tailles d'échantillons, la loi de W peut être approchée par une loi normale.

Cas des ex-æquo Nous avons vu section 3.2.1.2 qu'en présence d'ex-æquo nous remplacions les rangs des ex-æquo par le rang moyen des rangs qu'ils devraient occuper. Si les tailles d'échantillons sont inférieures à 10, les tables sont toujours utilisables. Pour de plus grandes tailles, l'approximation gaussienne est toujours valable mais la variance de W n'est plus identique à celle donnée précédemment.
Soit e le nombre de valeurs distinctes dans l'échantillon $(X_1, \ldots, X_{n_1+n_2})$, et soit $V_1, \ldots, V_e$ ces valeurs distinctes. Soit D_j le nombre d'apparitions de la valeur V_j dans l'échantillon ($1 \leq j \leq e$). La statistique W a alors pour variance :

$$V_{H_0}(W^*) = V(W) - \frac{n_1 n_2 \sum_{j=1}^{e}(D_j^3 - D_j)}{12(n_1+n_2)(n_1+n_2+1)}.$$

*Logiciel **R** : fonction* `wilcox.test`.

3.3.2.2 Test U de Mann-Whitney

Le test U de Mann-Whitney est basé sur la statistique U égale au nombre de paires (X_i, X_j) avec X_i dans le premier échantillon ($1 \leq i \leq n_1$) et X_j dans le second ($n_1 + 1 \leq j \leq n_2$) telle que $X_i > X_j$.
Ce test est identique au test de Wilcoxon puisque $U = W - \frac{n_1(n_1+1)}{2}$.

3.3.2.3 Test de la médiane

En choisissant $a(i) = \mathbb{1}_{](n_1+n_2+1)/2, +\infty]}(i)$, où $(n_1+n_2+1)/2$ est le rang moyen des observations, la statistique de test est

$$M = \sum_{i=1}^{n_1} \mathbb{1}_{](n_1+n_2+1)/2, +\infty]}(R_i)$$

et correspond au nombre d'éléments du premier échantillon supérieur à la médiane de l'échantillon total. La loi de M correspond à une loi hypergéométrique (on tire n_1 individus parmi n_1+n_2 avec sous H_0 probabilité $1/2$ d'être supérieur à la médiane de l'échantillon total).
Ce test est performant uniquement lorsque les distributions des deux échantillons sont très diffuses.

*Logiciel **R** : test à implémenter*

3.3.2.4 Test des scores normaux

En choisissant $a(i) = \Phi^{-1}\left(\frac{i}{n_1+n_2+1}\right)$ la statistique de test devient

$$SN = \sum_{i=1}^{n_1} \Phi^{-1}\left(R_i/(n_1+n_2+1)\right).$$

*Logiciel **R** : test à implémenter*

3.3.2.5 Test de Kolmogorov-Smirnov

Le test est le même que dans le cas de l'adéquation d'une distribution empirique à une distribution théorique, en remplaçant la fonction de répartition théorique par la version empirique du second échantillon :

$$KS = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \max_{x \in \mathbb{R}} |\hat{F}^1_{n_1}(x) - \hat{F}^2_{n_2}(x)|$$

où $\hat{F}^1_n$ et $\hat{F}^2_n$ sont les fonctions de répartitions empiriques des deux échantillons.

3.3.3 Cas de deux échantillons dépendants

Lorsque les deux échantillons ne sont pas indépendants, et qu'il s'agit par exemple d'une mesure sur les même individus statistiques dans deux conditions différentes (avant et après la prise un médicament par exemple), la solution est alors de travailler sur la différence des deux échantillons, que l'on comparera à la valeur centrale 0.

3.3.4 Tests de comparaison de deux proportions, pour de grands échantillons

Deux populations possèdent des individus ayant un certain caractère, en proportion p_1 et p_2. L'objet du présent test est de tester :

$$H_0 : p_1 = p_2 = p \text{ contre } H_1 : p_1 \neq p_2$$

On relève dans deux échantillons de tailles n_1 et n_2 les proportions f_1 et f_2 d'individus ayant ce caractère. Les tailles sont supposées suffisamment grandes ($n_i p_i > 5$ et $n_i(1-p_i) > 5$ pour $i = 1, 2$).
Ainsi les lois des fréquences empiriques F_1 et F_2 peuvent être approximées par des lois normales, d'où la statistique du test

$$U = \frac{F_1 - F_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}},$$

qui suit une loi normale centrée réduite sous H_0.
Si p est inconnue on la remplace par son estimation

$$\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2},$$

où f_1 et f_2 sont les estimations de p_1 et p_2.
La région critique sera alors déterminée par $|U| > u_{1-\frac{\alpha}{2}} = -u_{\frac{\alpha}{2}}$, d'où

$$\boxed{\text{on rejette } H_0 \text{ si } |f_1 - f_2| > u_{1-\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}.}$$

3.4 Tests de comparaison de K populations

Soit X une variable aléatoire quantitative, que l'on a observée pour K populations (ou de façon équivalente dans K conditions différentes). On dispose des K échantillons suivants :

- population $\mathcal{P}_1$: $X_{11}, \ldots, X_{n_1 1}$,
- population $\mathcal{P}_2$: $X_{12}, \ldots, X_{n_2 2}$,
- ...
- population $\mathcal{P}_K$: $X_{1K}, \ldots, X_{n_K K}$.

On note $n = \sum_{k=1}^{K} n_k$ est le nombre total d'observations.
Le test que l'on cherche à définir est le suivant :

- H_0 : les K populations $\mathcal{P}_k$ sont identiquement distribuées,
- H_1 : $\exists i, j$ telle que les populations $\mathcal{P}_i$ et $\mathcal{P}_j$ soient différentes.

L'hypothèse primordiale définissant le type de tests à effectuer est l'indépendance des populations entre elles. Nous présentons ci-après des tests paramétriques et non paramétriques dans le cas de populations indépendantes, puis nous examinerons le cas d'une dépendance particulière, celle des *mesures répétées*.

3.4.1 Tests de comparaison de K populations indépendantes

Exemple. On cherche à tester l'effet de K traitements médicamenteux, et pour cela on donne ces traitements à K groupes différents d'individus. Les K populations correspondent aux K groupes d'individus ayant reçu respectivement un des K traitements possibles. $X_{1k}, \ldots, X_{n_k k}$ sont les mesures de la réponse au traitement pour les n_k individus ayant reçus le traitement k.

3.4.1.1 Échantillons gaussiens ou de grandes tailles : ANOVA 1 facteur

Sous l'hypothèse que les populations sont de variances identiques (homoscedasticité), nous sommes en présence d'un problème d'analyse de variance (ANOVA) à un facteur (ici le facteur population), qui a déjà été présenté dans la section 3.2.6.4.

3.4.1.2 Échantillons de petites tailles : test de Kruskal-Wallis

La version non-paramétrique de l'ANOVA à un facteur est le test de Kruskal-Wallis, basés sur les rangs.
Soit R_{jk} le rang de la variable X_{jk} dans le classement dans l'ordre croissant de toutes les observations des K échantillons (supposé sans ex-æquo).
Soit $R_{.k} = \frac{1}{n_k} \sum_{j=1}^{n_k} R_{jk}$ le rang moyen dans l'échantillon de la population $\mathcal{P}_k$.
Sous l'hypothèse H_0 d'égalité des fonctions de répartitions F_k de chaque population

$$H_0 : F_1 = \ldots = F_K,$$

le rang moyen $R_{.k}$ de chaque population doit être proche de $E[R_{jk}] = \frac{n+1}{2}$.
La statistique du test de Kruskal-Wallis est

$$KW = \frac{12}{n(n+1)} \sum_{k=1}^{K} \left(R_{.k} - \frac{n+1}{2} \right)^2$$

qui suit sous H_0, lorsque les tailles n_k des échantillons tendent vers l'infini, approximativement une loi du χ^2 à $K-1$ degrés de liberté. Cette approximation est valable lorsque $K > 3$ et $min(n_1, \ldots, n_K) > 5$, et des tables existent lorsque ce n'est pas le cas.

Remarque. *On retrouve le test de Wilcoxon lorsque $K = 2$.*

En présence d'**ex-aequo**, les rangs seront remplacés par les rangs moyens et les lois de la statistique KW données ci-dessus restent approximativement valable.
*Logiciel **R** : fonction* `kruskal.test`

3.4.2 Tests de comparaison de K populations dépendantes (cas des mesures répétées)

Supposons maintenant que les K populations consistent en les mesures des mêmes individus statistiques dans K conditions différentes. On est alors dans une problématique de **mesures répétées** puisque les mesures sont répétées sur les même individus. De fait, on perd l'indépendance entre les populations puisqu'en particulier $X_{j1}, \ldots, X_{jK}$ sont liées en tant que mesures d'un même individu. A noter que comme on suppose que ce sont les mêmes individus qui sont mesurés, le nombre n_k est constant ($n_k = n$).

Exemple. On mesure le taux de diabète de n patients à K différents instants après l'ingestion d'un médicament.

3.4.2.1 Échantillons gaussiens ou de grandes tailles : ANOVA 2 facteurs

Dans le cas d'échantillons gaussiens ou de grandes tailles, une solution classique est de réaliser un analyse de variance à 2 facteurs : 1 facteur pour la population/condition/traitement, comme précédemment, et un facteur individu.

Nous présentons ci-après l'ANOVA à 2 facteurs génériques A et B, dans le cas légèrement plus général d'un plan *équilibré* ou *équirépété*, c'est-à-dire où le nombre de mesures pour chaque croisement des facteurs des deux niveaux est constant égal à r (et non plus égal à 1 comme précédemment).

L'objectif de l'analyse de variance à deux facteurs consiste à étudier les liens éventuels entre une variable continue X et deux facteurs A et B à J et K niveaux.
On note :

- X_{jk} la variable X observée pour les j-ème et k-ème valeurs respectives des facteurs A et B,
- X_{ijk} la variable aléatoire correspondant à la i-ème observation de X_{jk},
- n_{jk} le nombre d'observations X_{ijk},
- $n_{j.} = \sum_{k=1}^{K} n_{jk}$, $n_{.k} = \sum_{j=1}^{J} n_{jk}$ et $n = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}$.

On suppose que pour chaque croisement des modalités de A et B, les observations X_{ijk} sont indépendantes et identiquement distribuées selon une $\mathcal{N}(\mu_{jk}, \sigma^2)$. Les échantillons correspondant à des modalités différentes sont eux aussi supposés indépendants les uns des autres. Nous supposons enfin que les n_{jk} sont constants ($n_{jk} = r$ plan équilibré ou équirépété).
Dans le modèle le plus général pour la moyenne μ_{jk}, on suppose qu'elle peut s'écrire comme une somme d'un terme constant et de termes dépendants du facteur A, du facteur B et de l'interaction entre les facteurs A et B :

$$\mu_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}, \tag{3.2}$$

avec les contraintes d'unicité $\sum_j \alpha_j = \sum_k \beta_k = \sum_k \gamma_{jk} = \sum_j \gamma_{jk} = 0$.
On considère les moyennes suivantes :

$$\bar{X}_{.jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} X_{ijk}, \qquad \bar{X}_{..k} = \frac{1}{n_{.k}} \sum_{j=1}^{J} \bar{X}_{.jk}, \qquad \bar{X}_{.j.} = \frac{1}{n_{j.}} \sum_{k=1}^{K} \bar{X}_{.jk} \quad \text{et} \quad \bar{X}_{...} = \frac{1}{n} \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} X_{ijk}.$$

ainsi que les sommes des carrés suivantes :

$$SST = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} (X_{ijk} - \bar{X}_{...})^2, \qquad SSA = \sum_{j=1}^{J} n_{j.}(\bar{X}_{.j.} - \bar{X}_{...})^2, \qquad SSB = \sum_{k=1}^{K} n_{.k}(\bar{X}_{..k} - \bar{X}_{...})^2,$$

$$SSAB = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}(\bar{X}_{.jk} - \bar{X}_{.j.} - \bar{X}_{..k} + \bar{X}_{...})^2, \quad \text{et} \quad SSR = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} (X_{ijk} - \bar{X}_{.jk})^2,$$

où SST est la somme des carrés totale, SSA est la somme des carrés relatifs au facteur A, SSB est la somme des carrés relatifs au facteur B, $SSAB$ est la somme des carrés relatifs à l'interaction entre les facteurs A et B et SSR est la somme des carrés résiduels.
En remarquant que que l'on peut écrire $SST = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}} X_{ijk}^2 - n\bar{X}_{...}^2$, on obtient l'équation d'analyse de la variance à deux facteurs :

$$SST = SSA + SSB + SSAB + SSR$$

Comme en analyse de variance à un facteur, sous l'hypothèse $H_0 : \alpha_j = 0$, les quantités SSA et SSR suivent à σ^2 près des lois du χ^2 indépendantes à $J-1$ et $n-JK$ degrés de liberté. La statistique suivante est donc de loi de Fisher de paramètres $J-1$ et $n-JK$:

$$F_A = \frac{SSA/(J-1)}{SSR/(n-JK)}.$$

De même, sous les hypothèses respectives $H_0 : \beta_k = 0$ et $H_0 : \gamma_{jk} = 0$, les statistiques

$$F_B = \frac{SSB/(K-1)}{SSR/(n-JK)} \quad \text{et} \quad F_{AB} = \frac{SSAB/(K-1)(J-1)}{SSR/(n-JK)}$$

suivent des lois de Fisher de paramètres $K-1$ et $n-JK$ pour F_B, $(K-1)(J-1)$ et $n-JK$ pour F_{AB}.
Ainsi, on peut donc tester l'existence des effets principaux des deux facteurs et de leur interaction en comparant ces statistiques aux quantiles de la loi de Fisher : si les valeurs observées de ces statistiques sont supérieures au quantile de la loi de Fisher d'ordre $1-\alpha$ on conclura à un effet significatif.
On présente usuellement l'analyse de variance sous la forme du tableau suivant

Facteur	Somme des carrés	degrés de liberté	carré moyen	F
A	SSA	$J-1$	$SSA/(J-1)$	$F_A = \frac{SSA/(J-1)}{SSR/(n-JK)}$
B	SSB	$K-1$	$SSB/(K-1)$	$F_B = \frac{SSB/(K-1)}{SSR/(n-JK)}$
Interaction AB	$SSAB$	$(J-1)(K-1)$	$SSAB/(K-1)(J-1)$	$F_{AB} = \frac{SSAB/(K-1)(J-1)}{SSR/(n-JK)}$
Résidu	SSR	$n-JK$	$SSR/(n-JK)$	
Total	SST	$n-1$		

Estimation des effets Sous les hypothèses de contraintes $\sum_k \alpha_k = \sum_j \beta_j = \sum_k \gamma_{jk} = \sum_j \gamma_{jk} = 0$, les paramètres α_j, β_k et γ_{jk} de la décomposition (3.2) de μ_{jk} peuvent être estimés par les relations suivantes :

$$\alpha_j = \bar{x}_{.j.} - \bar{x}_{...}, \qquad \beta_k = \bar{x}_{..k} - \bar{x}_{...} \quad \text{et} \quad \gamma_{jk} = \bar{x}_{.jk} - \bar{x}_{.j.} - \bar{x}_{..k} + \bar{x}_{...}$$

3.4.2.2 Échantillons de petites tailles

Nous revenons au cas dans lequel on dispose des K échantillons :
- $X_{11}, \ldots, X_{n1}$: mesure des n individus dans la conditions 1,
- $X_{12}, \ldots, X_{n2}$: mesure des n individus dans la conditions 2,
- ...
- $X_{1K}, \ldots, X_{nK}$: mesure des n individus dans la conditions K,

Puisque les observations $X_{j1}, \ldots, X_{jK}$ sont les mesures d'un même individu, elles sont dépendantes entre elles. On ne peut donc comparer ces valeurs avec les valeurs des mesures des autres individus.
Nous nous intéressons donc aux **rangs intra-individu** R_{jk} des variables X_{jk} dans le classement dans l'ordre croissant de $X_{j1}, \ldots, X_{jK}$, qui correspond aux mesures de l'individu j pour chaque condition (supposé sans ex-æquo).

Exemple. Revenons à l'exemple dans lequel X_{jk} est la mesure du diabète de l'individu j au temps k. Comme $X_{j1}, \ldots, X_{jK}$ sont les mesures du diabète d'une même personne à différents instants, ces mesures peuvent par exemple être toute extrêmement élevées en comparaison des autres valeurs, uniquement parce que la personne est la seule diabétique de l'étude. Afin de prendre en compte cet effet individu, nous nous intéressons aux rangs intra-individu des mesures $X_{j1}, \ldots, X_{jK}$.

Test de Friedman On teste l'hypothèse H_0 d'égalité des fonctions de répartitions F_k de chaque population

$$H_0 : F_1 = \ldots = F_K.$$

Soit $R_{.k} = \frac{1}{n}\sum_{j=1}^{n} R_{jk}$ le rang moyen de la condition/population k. Sous l'hypothèse H_0, on doit avoir $E[R_{.k}] = (K+1)/2$.
La statistique de Friedman est alors

$$F = \frac{12n}{K(K+1)} \sum_{k=1}^{K} \left(R_{.k} - \frac{K+1}{2} \right)^2 = \frac{12}{nK(K+1)} \sum_{k=1}^{K} R_{.k}^2 - 3n(K+1)$$

qui suit asymptotiquement sous H_0 une loi du χ^2 à $K-1$ degrés de liberté. Puisqu'on s'intéresse généralement à des échantillons de petites tailles, la distribution asymptotique de F n'est rarement utilisable et on se référera généralement à la table statistique tabulant ses valeurs (Annexe 4.2.5).

En présence d'ex-æquo, il faut corriger la statistique F en la divisant par

$$C = 1 - \frac{\sum_{i=1}^{s} (t_i^3 - t_i)}{n(K^3 - K)}$$

où s est le nombre de séries de valeurs ex-aequo et t_i le nombre d'éléments de la $i$ème série d'ex-aequo.
*Logiciel **R** : fonction* `friedman.test`

Test de Quade Le test de Friedman peut être amélioré en prenant en compte les différences de valeurs X_{jk} pour un même individu. Pour cela, on introduit l'étendue $E_j = max_k(X_{jk}) - min_k(X_{jk})$ qui est la différence entre la valeur maximale et la valeur minimale pour un individu.
Soit S_j le rang de l'étendue E_j dans le classement des étendues intra-individu $E_1, \ldots, E_n$ (rang moyen en présence d'ex-æquo).
On remplace chaque observation X_{jk} par

$$Q_{jk} = S_j(R_{jk} - \frac{K+1}{2})$$

et soit $Q_k = \sum_{j=1}^{n} Q_{jk}$.
Les statistiques $T = \sum_{j=1}^{n} \sum_{k=1}^{K} Q_{jk}^2$ et $B = \sum_{k=1}^{K} Q_k^2$ peuvent être interprétées comme représentant respectivement les variations intra-individu et inter-individus.
La statistique du test de Quade est

$$Q = \frac{(n-1)B}{T-B}$$

qui suit approximativement sous H_0 une loi de Fisher à $K-1$ et $(n-1)(K-1)$ degrés de libertés.

*Logiciel **R** : fonction* `quade.test`

Remarque. *Le test de Quade est plus puissant que le test de Friedman lorsque les distributions des données sont très hétérogènes et lorsque le nombre K d'échantillons est pas trop grand ($K < 5$).*

Test de Page Le test de Page est une variante du test de Friedman dans le cas où un ordre est imposé dans l'hypothèse alternative :

$$H_0 : F_1 = \ldots = F_K,$$

contre

$$H_1 : F_1 > \ldots > F_K.$$

Ce type de test peut être intéressant pour tester une évolution monotone de la variable X au sein des populations/conditions $\mathcal{P}_1, \ldots, \mathcal{P}_K$ (évolution temporelle dans le cas où les populations/conditions sont indexées par le temps).
La statistique du test de Page est

$$L = \sum_{k=1}^{K} k \sum_{j=1}^{n} R_{jk}$$

où R_{jk} est toujours le rang intra-individu de l'observation X_{jk} parmi $(X_{j1}, \ldots, X_{jK})$. Pour des petites valeurs de n et de K, la loi de L a été tabulée.
Lorsque $n \geq 12$ et $K \geq 4$, la statistique suivante suis sous H_0 une loi du χ^2 à 1 degré de liberté :

$$\frac{a^2}{nK^2(K^2-1)(K+1)} \qquad \text{avec} \quad a = 12L - 3nK(K+1)^2$$

L'hypothèse H_0 sera rejetée au profit de $H_1 : F_1 > \ldots > F_K$ si $a > 0$ et si la valeur de la statistique est supérieure au quantile d'ordre $1-\alpha$ de la loi χ_1^2. Si $a < 0$ et si la valeur de la statistique est supérieure au quantile d'ordre $1-\alpha$ de la loi χ_1^2, l'hypothèse H_0 sera alors rejetée au profit de $H_1 : F_1 < \ldots < F_K$.
*Logiciel **R** : test à implémenter.*

Chapitre 4

Annexes

4.1 Rappel sur les convergences des suites de variables aléatoires

Soit (X_n) une suite de variables aléatoires réelles.

Définition 1. *La suite (X_n) **converge en probabilité** vers une variables aléatoire X si $\forall \epsilon, \eta$ positifs, il existe n_0 tel que*

$$\forall n > n_0, \qquad P(|X_n - X| > \epsilon) < \eta$$

Définition 2. *La suite (X_n) **converge presque sûrement** vers la variable aléatoire X si*

$$P(\{\omega | \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}) = 0$$

Définition 3. *La suite (X_n) **converge en moyenne d'ordre** p vers la variable aléatoire X si*

$$E[|X_n - X|^p] \to 0$$

Définition 4. *La suite (X_n) **converge en loi** vers la variable aléatoire X de fonction de répartition F si en tout point de continuité de F, la suite F_n des fonctions de répartition de X_n converge vers F*

Propriété 1.

$$\begin{array}{ccccc} (X_n) \xrightarrow{p.s.} X & \searrow & & & \\ & & (X_n) \xrightarrow{P} X & \to & (X_n) \xrightarrow{\mathcal{L}} X \\ (X_n) \xrightarrow{\text{moyenne ordre p}} X & \nearrow & & & \end{array}$$

4.1.0.3 Loi faible des grands nombres

Soit $(X_1, \ldots, X_n)$ un échantillon indépendant et identiquement distribué, avec $E[X_i] = \mu$ et $V(X_i) = \sigma^2 < \infty$. On a alors

$$\bar{X} \xrightarrow{P} \mu$$

4.1.0.4 Loi forte des grands nombres

Soit $(X_1, \ldots, X_n)$ un échantillon indépendant et identiquement distribué, avec $E[X_i] = \mu < \infty$ et $V(X_i) = \sigma^2$

$$\bar{X} \xrightarrow{p.s.} \mu$$

4.1.0.5 Théorème centrale limite

Soit $(X_1, \ldots, X_n)$ un échantillon indépendant et identiquement distribué, avec $E[X_i] = \mu$ et $V(X_i) = \sigma^2 < \infty$. On a alors

$$\bar{X} \xrightarrow{\mathcal{L}} \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

4.2 Tables statistiques pour test

4.2.1 Test des rangs signés

Attention : les bornes délimitent la zone de rejet (bilatérale), et il faut donc rejeter H_0 lorsque la statistique de test tombe sur ces bornes.

ST2238 Introductory Biostatistics

Critical values for the Wilcoxon signed rank test

This table gives critical values for the Wilcoxon signed rank test for sample sizes n of 20 or less for the hypothesis that the median of the population is zero. The tabulated values are the values of the test statistic W beyond which the p-value is less than the column heading. For $n > 20$, W is approximately normally distributed with mean $n(n+1)/4$ and variance $n(n+1)(2n+1)/24$, and so the p-value can be determined by comparing

$$Z = \frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

to the standard normal tables. The tabulated p-values are for the two-sided alternative.

n	5%		1%	
6	0	21		
7	2	26		
8	3	33	0	36
9	5	40	1	44
10	8	47	3	52
11	10	56	5	61
12	13	65	7	71
13	17	74	9	82
14	21	84	12	93
15	25	95	15	105
16	29	107	19	117
17	34	119	23	130
18	40	131	27	144
19	46	144	32	158
20	52	158	37	173

4.2.2 Test du signe

ST2238 Introductory Biostatistics

p-values for the sign test

This table gives p-values for the sign test for sample sizes n of 20 or less for the hypothesis that the median of the population is zero. The tabulated values are the p-values corresponding to the values of the test statistic, C, where C is the number of positives in the sample. If any data are equal to 0, they should be omitted and the sample size reduced accordingly. For $n > 20$, C is approximately normally distributed with mean $\mu = n/2$ and variance $s^2 = n/4$, and so the p-value can be approximately computed by comparing $2C/n - 1$ to standard normal tables. For $n \leq 4$, p-values are always more than 6%. The tabulated p-values are for the two-sided alternative.

$n = 5$

C	C	p
0	5	0.0625
1	4	0.3750
2	3	1.0000

$n = 6$

C	C	p
0	6	0.0313
1	5	0.2188
2	4	0.6875
3		1.0000

$n = 7$

C	C	p
0	7	0.0156
1	6	0.1250
2	5	0.4531
3	4	1.0000

$n = 8$

C	C	p
0	8	0.0078
1	7	0.0703
2	6	0.2891
3	5	0.7266
4		1.0000

$n = 9$

C	C	p
0	9	0.0039
1	8	0.0391
2	7	0.1797
3	6	0.5078
4	5	1.0000

$n = 10$

C	C	p
0	10	0.0020
1	9	0.0215
2	8	0.1094
3	7	0.3438
4	6	0.7539
5		1.0000

$n = 11$

C	C	p
0	11	0.0010
1	10	0.0117
2	9	0.0654
3	8	0.2266
4	7	0.5488
5	6	1.0000

$n = 12$

C	C	p
0	12	0.0005
1	11	0.0063
2	10	0.0386
3	9	0.1460
4	8	0.3877
5	7	0.7744
6		1.0000

$n = 13$

C	C	p
0	13	0.0002
1	12	0.0034
2	11	0.0225
3	10	0.0923
4	9	0.2668
5	8	0.5811
6	7	1.0000

$n = 14$

C	C	p
0	14	0.0001
1	13	0.0018
2	12	0.0129
3	11	0.0574
4	10	0.1796
5	9	0.4240
6	8	0.7905
7		1.0000

$n = 15$

C	C	p
0	15	0.0001
1	14	0.0010
2	13	0.0074
3	12	0.0352
4	11	0.1185
5	10	0.3018
6	9	0.6072
7	8	1.0000

$n = 16$

C	C	p
0	16	0.0000
1	15	0.0005
2	14	0.0042
3	13	0.0213
4	12	0.0768
5	11	0.2101
6	10	0.4545
7	9	0.8036
8		1.0000

$n = 17$

C	C	p
0	17	0.0000
1	16	0.0003
2	15	0.0023
3	14	0.0127
4	13	0.0490
5	12	0.1435
6	11	0.3323
7	10	0.6291
8	9	1.0000

$n = 18$

C	C	p
0	18	0.0000
1	17	0.0001
2	16	0.0013
3	15	0.0075
4	14	0.0309
5	13	0.0963
6	12	0.2379
7	11	0.4807
8	10	0.8145
9		1.0000

$n = 19$

C	C	p
0	19	0.0000
1	18	0.0001
2	17	0.0007
3	16	0.0044
4	15	0.0192
5	14	0.0636
6	13	0.1671
7	12	0.3593
8	11	0.6476
9	10	1.0000

$n = 20$

C	C	p
0	20	0.0000
1	19	0.0000
2	18	0.0004
3	17	0.0026
4	16	0.0118
5	15	0.0414
6	14	0.1153
7	13	0.2632
8	12	0.5034
9	11	0.8238
10		1.0000

4.2.3 Test de Wilcoxon (2 populations)

Attention : les bornes délimitent la zone de rejet (bilatérale), et il faut donc rejeter H_0 lorsque la statistique de test tombe sur ces bornes.

ST2238 Introductory Biostatistics

Critical values for the Wilcoxon rank sum test

This table gives critical values for the Wilcoxon rank sum test for two samples both of size 10 or less for the hypothesis that the two populations have the same underlying distributions. The tabulated values are the values of the test statistic R equal to the sum of the ranks in the smaller sample (with sample size n_S) beyond which the p-value is less than the column heading (the larger sample is of size n_B). The tabulated p-values are for the two-sided alternative.

For larger samples, R is approximately normally distributed with mean $\mu_R = n_S(n_S+n_B+1)/2$ and variance $\sigma_R^2 = (n_S n_B)(n_S+n_B+1)/12$, and so the p-value can be determined by comparing

$$Z = \frac{R-\mu_R}{\sigma_R}$$

to the standard normal tables.

n_B	n_S	5%		1%	
4	4	10	26	—	—
5	4	11	29	—	—
5	5	17	38	15	40
6	4	12	32	10	34
6	5	18	42	16	44
6	6	26	52	23	55
7	4	13	35	10	38
7	5	20	45	16	49
7	6	27	57	24	60
7	7	36	69	32	73
8	4	14	38	11	41
8	5	21	49	17	53
8	6	29	61	25	65
8	7	38	74	34	78
8	8	49	87	43	93
9	4	14	42	11	45
9	5	22	53	18	57
9	6	31	65	26	70
9	7	40	79	35	84
9	8	51	93	45	99
9	9	62	109	56	115
10	4	15	45	12	48
10	5	23	57	19	61
10	6	32	70	27	75
10	7	42	84	37	89
10	8	53	99	47	105
10	9	65	115	58	122
10	10	78	132	71	139

4.2.4 Test de Shapiro-Wilk (normalité)

Ces tables sont dues à Christophe Chesneau `http://www.math.unicaen.fr/~chesneau/`.

(Table 9) Coefficients de Shapiro-Wilk

Les colonnes des tableaux ci-dessous donnent les coefficients de Shapiro-Wilk $(a_1, \ldots, a_\rho)$ où ρ est l'entier tel que $n = 2\rho$ ou $n = 2\rho + 1$ selon la parité de n.

i \ n	2	3	4	5	6	7	8	9	10
1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739
2			0,1677	0,2413	0,2806	0,3031	0,3164	0,3244	0,3291
3					0,0875	0,1401	0,1743	0,1976	0,2141
4							0,0561	0,0947	0,1224
5									0,0399

i \ n	11	12	13	14	15	16	17	18	19	20
1	0,5601	0,5475	0,5359	0,5251	0,5150	0,5056	0,4963	0,4886	0,4808	0,4734
2	0,3315	0,3325	0,3325	0,3318	0,3306	0,3290	0,3273	0,3253	0,3232	0,3211
3	0,2260	0,2347	0,2412	0,2460	0,2495	0,2521	0,2540	0,2553	0,2561	0,2565
4	0,1429	0,1586	0,1707	0,1802	0,1878	0,1939	0,1988	0,2027	0,2059	0,2085
5	0,0695	0,0922	0,1099	0,1240	0,1353	0,1447	0,1524	0,1587	0,1641	0,1686
6		0,0303	0,0539	0,0727	0,0880	0,1005	0,1109	0,1197	0,1271	0,1334
7				0,0240	0,0433	0,0593	0,0725	0,0837	0,0932	0,1013
8						0,0196	0,0359	0,0496	0,0612	0,0711
9								0,0163	0,0303	0,0422
10										0,0140

i \ n	21	22	23	24	25	26	27	28	29	30
1	0,4643	0,4590	0,4542	0,4493	0,4450	0,4407	0,4366	0,4328	0,4291	0,4254
2	0,3185	0,3156	0,3126	0,3098	0,3069	0,3043	0,3018	0,2992	0,2968	0,2944
3	0,2578	0,2571	0,2563	0,2554	0,2543	0,2533	0,2522	0,2510	0,2499	0,2487
4	0,2119	0,2131	0,2139	0,2145	0,2148	0,2151	0,2152	0,2151	0,2150	0,2148
5	0,1736	0,1764	0,1787	0,1807	0,1822	0,1836	0,1848	0,1857	0,1064	0,1870
6	0,1399	0,1443	0,1480	0,1512	0,1539	0,1563	0,1584	0,1601	0,1616	0,1630
7	0,1092	0,1150	0,1201	0,1245	0,1283	0,1316	0,1346	0,1372	0,1395	0,1415
8	0,0804	0,0878	0,0941	0,0997	0,1046	0,1089	0,1128	0,1162	0,1192	0,1219
9	0,0530	0,0618	0,0696	0,0764	0,0823	0,0876	0,0923	0,0965	0,1002	0,1036
10	0,0263	0,0368	0,0459	0,0539	0,0610	0,0672	0,0728	0,0778	0,0822	0,0862
11		0,0122	0,0228	0,0321	0,0403	0,0476	0,0540	0,0598	0,0650	0,0697
12				0,0107	0,0200	0,0284	0,0358	0,0424	0,0483	0,0537
13						0,0094	0,0178	0,0253	0,0320	0,0381
14								0,0084	0,0159	0,0227
15										0,0076

(Table 10) Valeurs de Shapiro-Wilk

Les valeurs intérieures du tableau ci-dessous donnent les coefficient $w_{\alpha,n}$ utilisé dans le test de Shapiro-Wilk. Ici, n est la taille de l'échantillon et α est la valeur du risque.

n \ α	0,05	0,01
3	0,767	0,753
4	0,748	0,687
5	0,762	0,686
6	0,788	0,713
7	0,803	0,730
8	0,818	0,749
9	0,829	0,764
10	0,842	0,781
11	0,850	0,792
12	0,859	0,805
13	0,856	0,814
14	0,874	0,825
15	0,881	0,835
16	0,837	0,844
17	0,892	0,851
18	0,897	0,858
19	0,901	0,863
20	0,905	0,868
21	0,908	0,873
22	0,911	0,878
23	0,914	0,881
24	0,916	0,884
25	0,918	0,888
26	0,920	0,891

n \ α	0,05	0,01
27	0,923	0,894
28	0,924	0,896
29	0,926	0,898
30	0,927	0,900
31	0,929	0,902
32	0,930	0,904
33	0,931	0,906
34	0,933	0,908
35	0,934	0,910
36	0,935	0,912
37	0,936	0,914
38	0,938	0,916
39	0,939	0,917
40	0,940	0,919
41	0,941	0,920
42	0,942	0,922
43	0,943	0,923
44	0,944	0,924
45	0,945	0,926
46	0,945	0,927
47	0,946	0,928
48	0,947	0,929
49	0,947	0,929
50	0,947	0,930

4.2.5 Test de Friedman

Critical values for the Friedman Test

$$M = \frac{12}{nk(k+1)} \sum R_j^2 - 3n(k+1)$$

	k=3		k=4		k=5		k=6	
n	**α=5%**	**α=1%**	**α=5%**	**α=1%**	**α=5%**	**α=1%**	**α=5%**	**α=1%**
2	—	—	6.000	—	7.600	8.000	9.143	9.714
3	6.000	—	7.400	9.000	8.533	10.130	9.857	11.760
4	6.500	8.000	7.800	9.600	8.800	11.200	10.290	12.710
5	6.400	8.400	7.800	9.960	8.960	11.680	10.490	13.230
6	7.000	9.000	7.600	10.200	9.067	11.870	10.570	13.620
7	7.143	8.857	7.800	10.540	9.143	12.110	10.670	13.860
8	6.250	9.000	7.650	10.500	9.200	13.200	10.710	14.000
9	6.222	9.556	7.667	10.730	9.244	12.440	10.780	14.140
10	6.200	9.600	7.680	10.680	9.280	12.480	10.800	14.230
11	6.545	9.455	7.691	10.750	9.309	12.580	10.840	14.320
12	6.500	9.500	7.700	10.800	9.333	12.600	10.860	14.380
13	6.615	9.385	7.800	10.850	9.354	12.680	10.890	14.450
14	6.143	9.143	7.714	10.890	9.371	12.740	10.900	14.490
15	6.400	8.933	7.720	10.920	9.387	12.800	10.920	14.540
16	6.500	9.375	7.800	10.950	9.400	12.800	10.960	14.570
17	6.118	9.294	7.800	10.050	9.412	12.850	10.950	14.610
18	6.333	9.000	7.733	10.930	9.422	12.890	10.950	14.630
19	6.421	9.579	7.863	11.020	9.432	12.880	11.000	14.670
20	6.300	9.300	7.800	11.100	9.400	12.920	11.000	14.660
∞	5.991	9.210	7.815	11.340	9.488	13.280	11.070	15.090

For values of ***n*** greater than 20 and/or values of ***k*** greater than 6, use χ^2 tables with ***k*-1** degrees of freedom

4.2.6 Test de Kolmogorov-Smirnov

Critical values, $d_{alpha};(n)^a$, of the maximum absolute difference between sample $F_n(x)$ and population $F(x)$ cumulative distribution.

Number of trials, n	Level of significance, α			
	0.10	0.05	0.02	0.01
1	0.95000	0.97500	0.99000	0.99500
2	0.77639	0.84189	0.90000	0.92929
3	0.63604	0.70760	0.78456	0.82900
4	0.56522	0.62394	0.68887	0.73424
5	0.50945	0.56328	0.62718	0.66853
6	0.46799	0.51926	0.57741	0.61661
7	0.43607	0.48342	0.53844	0.57581
8	0.40962	0.45427	0.50654	0.54179
9	0.38746	0.43001	0.47960	0.51332
10	0.36866	0.40925	0.45662	0.48893
11	0.35242	0.39122	0.43670	0.46770
12	0.33815	0.37543	0.41918	0.44905
13	0.32549	0.36143	0.40362	0.43247
14	0.31417	0.34890	0.38970	0.41762
15	0.30397	0.33760	0.37713	0.40420
16	0.29472	0.32733	0.36571	0.39201
17	0.28627	0.31796	0.35528	0.38086
18	0.27851	0.30936	0.34569	0.37062
19	0.27136	0.30143	0.33685	0.36117
20	0.26473	0.29408	0.32866	0.35241
21	0.25858	0.28724	0.32104	0.34427
22	0.25283	0.28087	0.31394	0.33666
23	0.24746	0.27490	0.30728	0.32954
24	0.24242	0.26931	0.30104	0.32286

Critical values, $d_{alpha};(n)^a$, of the maximum absolute difference between sample $F_n(x)$ and population $F(x)$ cumulative distribution.

Number of trials, n	Level of significance, α			
	0.10	0.05	0.02	0.01
25	0.23768	0.26404	0.29516	0.31657
26	0.23320	0.25907	0.28962	0.31064
27	0.22898	0.25438	0.28438	0.30502
28	0.22497	0.24993	0.27942	0.29971
29	0.22117	0.24571	0.27471	0.29466
30	0.21756	0.24170	0.27023	0.28987
31	0.21412	0.23788	0.26596	0.28530
32	0.21085	0.23424	0.26189	0.28094
33	0.20771	0.23076	0.25801	0.27677
34	0.20472	0.22743	0.25429	0.27279
35	0.20185	0.22425	0.26073	0.26897
36	0.19910	0.22119	0.24732	0.26532
37	0.19646	0.21826	0.24404	0.26180
38	0.19392	0.21544	0.24089	0.25843
39	0.19148	0.21273	0.23786	0.25518
40^b	0.18913	0.21012	0.23494	0.25205

[a] Values of $d_\alpha(n)$ such that $p(\max)|F^n(x) - F(x)|d^\alpha(n) = \alpha$.

[b] $N > 40 \approx \frac{1.22}{N^{1/2}}, \frac{1.36}{N^{1/2}}, \frac{1.51}{N^{1/2}}$ and $\frac{1.63}{N^{1/2}}$ for the four levels of significance.

Bibliographie

[1] M. Carbon, C. Franck. *Estimation non paramétrique de la densité et de la régression - Prévision non paramétrique*. La revue MODULAD, numéro 15, juin 1995.

[2] G. Saporta. *Probabilités, analyse de données et statistique*. 2ème édition, Editions Technip, 2006.

[3] D.J. Sheskin. *Handbook of parametric and nonparametric statistical procedures.* Fifth edition. Chapman & Hall/CRC, 2011.