Notes de Cours UE STA401

Statistique et Calcul des Probabilités

Frédérique Leblanc

January 14, 2025

Plan de cours :

- Chap. 1 : Introduction
- Chap. 2 : Variables aléatoires discrètes et données discrètes
- Chap. 3 : Variables aléatoires continues et données continues
- Chap. 4 : Estimation et intervalles de confiance
- Chap. 5 : Tests paramétriques
- Chap. 6 : Tests de comparaisons d'échantillons
- Chap. 7 : Tests du Chi2

Chapter 1

Introduction

On souhaite étudier un caractère X sur une population $\mathcal{P}$.

1.1 Exemple :

Dans l'exemple qui illustrera les méthodes proposées dans ce document on observe plusieurs caractères : le sexe, le nombre de jours par semaine où sont consommés au moins quatre fruits ou légumes (appelé alimentation), le poids (à 20 ans et à 15 ans) ou encore la taille d'un étudiant d'une ancienne promotion de L2 à Valence. Le fichier de données est renvoyé à la fin de l'introduction.

1.2 Problématique :

En général, on ne peut pas observer le caractère X sur tous les individus d'une grande population, mais seulement sur une sous-population de $\mathcal{P}$ de taille n. Quelques définitions et notations s'imposent avant d'aller plus avant :

sous-population : $\{i_1, ..., i_j, ..., i_n\}$ un ensemble de n individus choisis au hasard dans $\mathcal{P}$.
échantillon de données : $x_1, ..., x_j, ..., x_n$ les n valeurs observées du caractère X sur les individus de la sous-population.

Plusieurs problèmes se posent alors naturellement :

1. Quelles informations sur le caractère X, peut-on tirer de l'échantillon ?
2. Quelle prévision pourrait on faire sur un individu non observé de $\mathcal{P}$, à partir des données observées $x_1, ..., x_j, ..., x_n$?
3. Quelle décision prendra-t-on pour choisir entre deux hypothèses (formulées sur les paramètres du modèle), en utilisant l'échantillon observé $x_1, ..., x_n$, avec un risque d'erreur spécifié ?
4. ...

Pour répondre à ces questions on utilisera la démarche suivante :

1. *Description du jeu de données* : résumés graphiques ou numériques des échantillons qui constituent le jeu de données (dans R les données sont rangées dans un objet nommé `data.frame`). Cette étape de visualisation et fouille des données est essentielle, car elle permettra, entre autres, de proposer un modèle pour X, d'apprécier l'effet ou non de certaines variables sur d'autres,...

2. *Modélisation* : le caractère X d'un individu choisi au hasard dans $\mathcal{P}$ est aléatoire (puisqu'il dépend de l'individu choisi). Il sera décrit par l'ensemble de ses valeurs possibles que l'on notera $\mathcal{X}$ et par la probabilité d'obtenir l'une ou l'autre des valeurs de $\mathcal{X}$. Modéliser le caractère d'intérêt revient à se donner l'ensemble $\mathcal{X}$ et la loi de probabilité associée P. Ce qui en d'autres termes revient à supposer que X est une variable aléatoire de distribution (ou loi de probabilité) P sur $\mathcal{X}$. Si cette distribution est connue, on peut prévoir (avec un certain risque d'erreur) la valeur de X pour n'importe quel individu tiré au hasard dans $\mathcal{P}$, à l'aide du calcul des probabilités.

3. *Inférence statistique* : Si on ne connaît pas P, on supposera que P dépend d'un paramètre inconnu θ. On estimera alors θ à l'aide de $x_1, ..., x_j, ..., x_n$, afin de pouvoir ensuite prévoir X pour tout individu de $\mathcal{P}$. On pourra aussi décider, au vu de l'échantillon observé, que le paramètre inconnu θ dépasse (ou non) un certain seuil, en contrôlant le risque de donner une conclusion erronée. Dans tous les cas, les décisions seront prises au vu de l'échantillon observé, c'est-à-dire à partir d'une information partielle sur $\mathcal{P}$. On risque donc de prendre une mauvaise décision, si par exemple l'échantillon représente mal la population totale. Les méthodes de décision seront donc développées de sorte que l'on puisse contrôler le risque de donner une mauvaise conclusion concernant $\mathcal{P}$, en utilisant l'échantillon observé.

Selon la forme de l'ensemble $\mathcal{X}$, la variable X aura trois "types" différents. On dira que X est une

- **variable qualitative** lorsque $\mathcal{X}$ est un ensemble fini de mots ou codes (par ex. $\mathcal{X} = \{\text{Femme, Homme}\}$ ou $\mathcal{X} = \{1, 0\}$). Dans ce cas les éléments de $\mathcal{X}$ ne peuvent pas être ordonnés.

- **variable quantitative discrète** lorsque $\mathcal{X}$ est une suite finie ou infinie d'éléments de $\mathbb{N}$ (par ex $\mathcal{X} = \{0, 1, ..., 5\}$ ou $\mathcal{X} = \mathbb{N}$).

- **variable quantitative continue** lorsque $\mathcal{X}$ est un intervalle de $\mathbb{R}$ (par ex $\mathcal{X} = \mathbb{R}$ ou $\mathcal{X} = [120, 210]$).

La description graphique de données qualitatives ou quantitatives discrètes et leurs modélisations diffèrent de celles de données quantitatives continues. Les premières seront abordées dans le chapitre 2 et les suivantes dans le chapitre 3. L'inférence statistique fera l'objet des chapitres 4 à 7. Mais avant de rentrer dans le vif du sujet il convient de faire de brefs rappels de calculs élémentaires des probabilités.

Exemple de données : On a observé sur un groupe d'étudiants les variables : sexe (1 pour F, 0 pour M) noté S, poids à 20 ans noté P, poids à 15 ans noté P', taille notée T, nombre de jours par semaine où l'étudiant a consommé au moins quatre fruits ou légumes dans la journée noté A.

individu i	sexe s_i	poids 20 ans p_i	poids 15 ans p'_i	taille t_i	alimentation a_i
1	1	49	45	160	5
2	1	53	45	164	4
3	1	50	45	161	4
4	1	49	45	175	6
5	1	70	74	166	4
6	1	43	40	165	5
7	1	52	45	164	5
8	1	50	47	164	7
9	1	49	45	166	6
10	1	49	43	161	5
11	1	59	55	167	4
12	1	70	60	179	3
13	1	50	45	166	6
14	1	50	45	168	5
15	1	61	50	175	5
16	1	47	44	152	4
17	1	54	49	157	3
18	1	53	49	162	4
19	1	55	50	164	4
20	1	59	54	162	1
21	1	45	44	162	6
22	1	60	50	175	5
23	1	55	53	158	3
24	1	68	55	170	2
25	0	75	70	179	3
26	0	78	72	180	2
27	0	64	58	170	3
28	0	69	63	180	5
29	0	70	58	183	4
30	0	62	56	179	5
31	0	75	65	186	4
32	0	78	65	182	3

Résumés numériques :

$$\sum s_i = \quad , \quad \sum p_i = \quad , \quad \sum p'_i = \quad , \quad \sum t_i = \quad , \quad \sum a_i =$$

$$\sum p_i t_i = \quad , \quad \sum p_i^2 = \quad , \quad \sum p_i'^2 = \quad , \quad \sum t_i^2 = \quad , \quad \sum a_i^2 =$$

$$\bar{p} = \quad , \quad \bar{p'} = \quad , \quad \bar{t} = \quad , \quad \bar{a} =$$

$$\mathrm{cov}(p,t) = \quad , \quad s_p^2 = \quad , \quad s_{p'}^2 = \quad , \quad s_t^2 = \quad , \quad s_a^2 =$$

Le calcul de ces résumés numériques est laissé au lecteur en exercice (à faire avec la calculatrice ou en utilisant R). Les définitions de $\bar{x}$, s_x^2 et $cov(x,y)$ sont rappelées dans les chapitres 2 et 3.

1.3 Probabilités : rappels

Soit Ω l'ensemble de toutes les éventualités d'une expérience. Soit $E = \mathcal{P}(\Omega)$ l'ensemble des évènements possibles relatifs à cette expérience. Soient A et B deux évènements (quelconque de E) et $\bar{A}$, et $\bar{B}$ leurs évènements complémentaires respectifs. Rappelons que $A \cap B$ est l'évènement A est réalisé **et** B est réalisé tandis que $A \cup B$ désigne A est réalisé **ou** B est réalisé. Ω est l'évènement certain et $\emptyset$ l'évènement impossible. $A|B$ l'évènement qui décrit A est réalisé sachant que B l'est (on dit A sachant B).

Définition : Une probabilité P sur Ω est une application de E dans $[0, 1]$ qui satisfait les trois axiomes suivants :

1. $\forall A \in E : 0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. pour une famille dénombrable d'ensembles $(A_i)_i$ deux à deux disjoints (t.q. $A_i \cap A_j = \emptyset$ pour tous $i \neq j$)
   $P(A_1 \cup A_2 ... \cup A_i ...) = \sum_{i=1}^{\infty} P(A_i)$

En utilisant les relations et propriétés des ensembles on établit :

Propriétés :

1. $P(\emptyset) = 0$
2. $P(\bar{A}) = 1 - P(A)$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   en particulier si $A \cap B = \emptyset$ alors $P(A \cup B) = P(A) + P(B)$
   (cas part. du troisième axiome définissant une probabilité sur Ω)
4. $P(A) = P(A \cap B) + P(A \cap \bar{B})$

On peut également vouloir évaluer la probabilité d'un évênement conditionnellement à la réalisation d'un autre évènement décrit par B tel que $P(B) \neq 0$.

Définition : probabilité conditionnelle
Soit $B \in E$ tel que $P(B) \neq 0$ alors on appelle probabilité de A conditionnelle à B la quantité

$$P_B(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \forall A \in E.$$

P_B satisfaisant les trois axiomes précédents c'est donc une probabilité sur Ω.

Dans de nombreux problèmes, étant donnée une partition $B_1, ..., B_n$ de Ω on a facilement accès aux probabilités conditionnelles à chaque B_i d'un évènement quelconque A et dans ce cas on utilisera la propriété suivante appelée formule des probabilités totales.

Formule des probabilités totales :
Soit $B_1, ..., B_n$ une partition de Ω alors

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

On peut en déduire facilement la propriété suivante très utile en pratique :

Formule de Bayes :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

Terminons avec deux définitions importantes :

- A et B sont dits incompatibles (ne peuvent se réaliser simultanément) s'ils sont disjoints (c. à d. $A \cap B = \emptyset$) ou s'ils vérifient $P(A \cap B) = 0$.
- A et B sont dits indépendants si : $P(A \cap B) = P(A)P(B)$ ou de façon équivalente si $P(A|B) = P(A)$ (que B soit réalisé ou non n'impacte pas qu'A le soit)

Chapter 2

Variables aléatoire discrètes et données discrètes

Nous nous intéressons au cas où $\mathcal{X}$ est un ensemble fini de q éléments : $m_1, ..., m_q$ appelés modalités de X. S'il s'agit d'éléments quantitatifs, on conviendra d'ordonner les modalités dans l'ordre croissant $(m_1 < m_2 < ... < m_q)$.

2.1 Analyse descriptive de données discrètes

2.1.1 Définitions et notations

Tableau de données : Tableau croisant une population (ensemble d'individus de cardinal n) et un ensemble de p caractères (observés sur les individus de la population). Un tableau de données peut se présenter sous deux formes : tableau de données brutes ou tableau d'effectifs. On utilisera les notations suivantes :

- p caractères notés : $X^1, ..., X^p$.
- n individus notés : $i_1, ..., i_n$.
- n échantillon de données : réalisations d'un caractère X notées : $x_1, ..., x_n$, avec $x_i \in \mathcal{X}$. (x_j étant la réalisation de X pour l'individu i_j et $\mathcal{X}$ l'ensemble des modalités de X).
- q modalités de la variable X notées : $m_1, ..., m_q$ où $q \leq n$ (c.-à-d. $\mathcal{X} = \{m_1, ..., m_q\}$). Les modalités désignent les valeurs distinctes possibles des réalisations de X dans l'échantillon de données. Si toutes les réalisations $x_1, ..., x_n$ sont distinctes alors X sera modélisé de préférence par une variable continue.
- la taille du sous-échantillon pour lequel X prend la modalité m_k sera notée n_k. On a évidemment la relation $\sum_{k=1}^{q} n_k = n$.

Les données sont en général présentées sous la forme suivante lorsque plusieurs caractères sont observés sur une même population. Il s'agit d'un tableau dit *tableau de données brutes* ou encore `data.frame` :

indiv	X	X^1	...	X^p
1	x_1	x_1^1	...	x_1^p
...				
i	x_i	x_i^1	...	x_i^p
...				
n	x_n	x_n^1	...	x_n^p

Pour chaque caractère (ou variable) X observé sur une population, on peut aussi présenter données sous la forme d'un *tableau en effectifs* :

X	effectifs
m_1	n_1
...	
m_k	n_k
...	
m_q	n_q

2.1.2 Tableau de distribution

C'est le tableau en effectifs auquel on ajoute une colonne donnant les proportions notées $f_k = n_k/n$ associées à chaque modalité m_k. Cette colonne est souvent intitulée *pourcentages* ou *fréquences*. Comme par convention $\mathcal{X} = \{m_1, ..., m_q\}$ est un ensemble numérique ordonné, on ajoute à ce tableau une colonne contenant les fréquences cumulées : $F_k = f_1 + ... + f_k$. Il s'agit de la proportion d'individus de l'échantillon pour lesquels on observe une valeur $\leq m_k$ qui sera notée $freq]-\infty, m_k]$.

X	effectifs	fréquences	fréq. cumul.
m_1	n_1	$f_1 = n_1/n$	$F_1 = f_1$
...			
m_k	n_k	$f_k = n_k/n$	$F_k = f_1 + ... + f_k$
...			
m_q	n_q	$f_q = n_q/n$	$F_q = f_1 + ... + f_q = 1$

On peut plus généralement définir la proportion de la population pour laquelle la variable X prend une valeur $\leq x$, et on définit ainsi la fonction de répartition empirique (empirical cumulative distribution function).

Fonction de répartition empirique de X :

$$\begin{aligned} \hat{F} : I\!R &\rightarrow [0,1] \\ x &\rightarrow freq]-\infty, x] = \frac{card\{x_i | x_i \leq x\}}{n} \end{aligned}$$

Propriétés :
$\forall x < m_1; \hat{F}(x) = 0$ et $\forall x \geq m_q; \hat{F}(x) = 1$.
$\forall x \in [m_k, m_{k+1}[; \hat{F}(x) = F_k, 1 \leq k \leq q-1$.
$\forall a \leq b; \hat{F}(b) - \hat{F}(a) = freq]a, b]$.
Avec le logiciel R, on en calculera les valeurs avec la fonction de R, `ecdf()`.

2.1.3 Représentations graphiques

Répartition observée : représentation des fréquences par un diagramme en barres, où la hauteur de chaque barre est la fréquence de la modalité qu'elle représente. On peut également représenter

les fréquences observées à l'aide d'un diagramme en secteurs (c.-à-d. "camembert"). Par exemple s'agissant de la variable *alimentation* de l'exemple 1, les deux représentations possibles de la répartition sont présentées dans la figure 2.1.

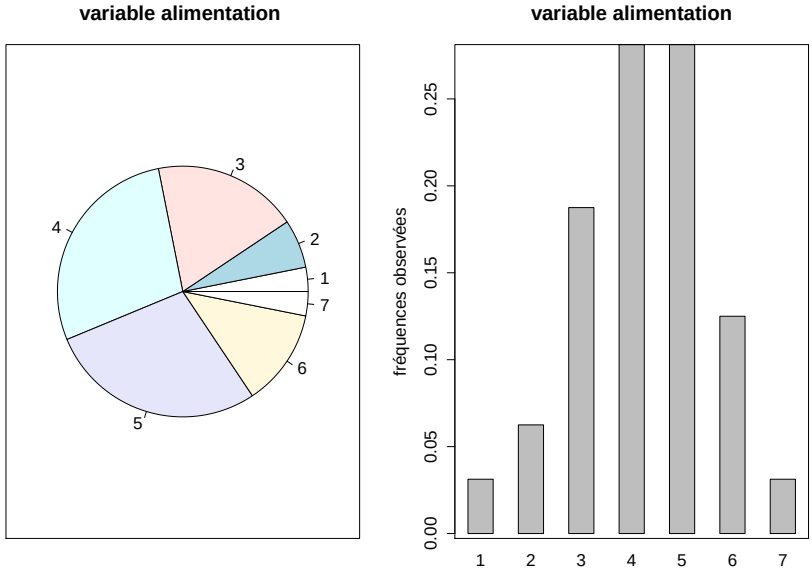

Figure 2.1: Répartition de la variable alimentation (nombre de jours de la semaine où on été consommés au moins 4 fruits et légumes), représentée par un diagramme en secteurs à gauche et un diagramme en barres à droite.

Un résumé de la répartition observées : La boite de distribution, appelée aussi "boite à moustaches" ou boxplot. Elle permet de visualiser les quartiles et valeurs extrêmes de l'échantillon et elle n'utilise qu'un seul axe muni d'une unité (l'axe des abscisses si le boxplot est horizontal et des ordonnées s'il est vertical). Ce graphique permet d'apprécier en un coup d'oeil, la symétrie de la répartition observée, les positions des quartiles, et la possible présence d'individus extrêmes très différents des autres. De plus lorsque l'on souhaite comparer plusieurs échantillons, c'est un outil très utile pour représenter sur une même figure toutes les boites de distributions à des fins de comparaisons (voir TP2).

Fonction de répartition empirique : Le graphe de $\hat{F}$ est une fonction en escaliers croissante valant 0 en tout point strictement inférieur à m_1 et valant 1 en tout point supérieur ou égal à m_q.

Par exemple pour la variable alimentation de l'exemple 1 on obtient la figure 2.3

2.1.4 Résumés numériques

On peut résumer un jeu de données par différentes caractéristiques, telles que celles utilisées dans le boxplot ainsi que celles permettant dévaluer le centrage et la dispersion des données.

– **Caractéristiques centrales** –
Les plus utilisées sont le mode, la médiane et la moyenne arithmétique.

Mode : Valeur (modalité) en laquelle l'histogramme des fréquences présente un maximum relatif.

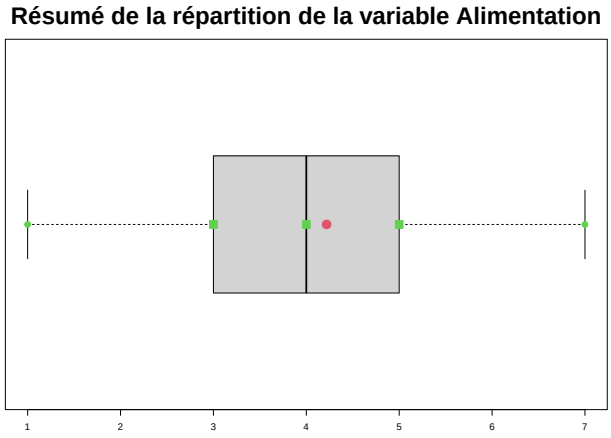

Figure 2.2: Boxplot de la variable alimentation, les deux points verts sont le minimum et le maximum, les carrés verts les trois quartiles et le point rouge est la moyenne.

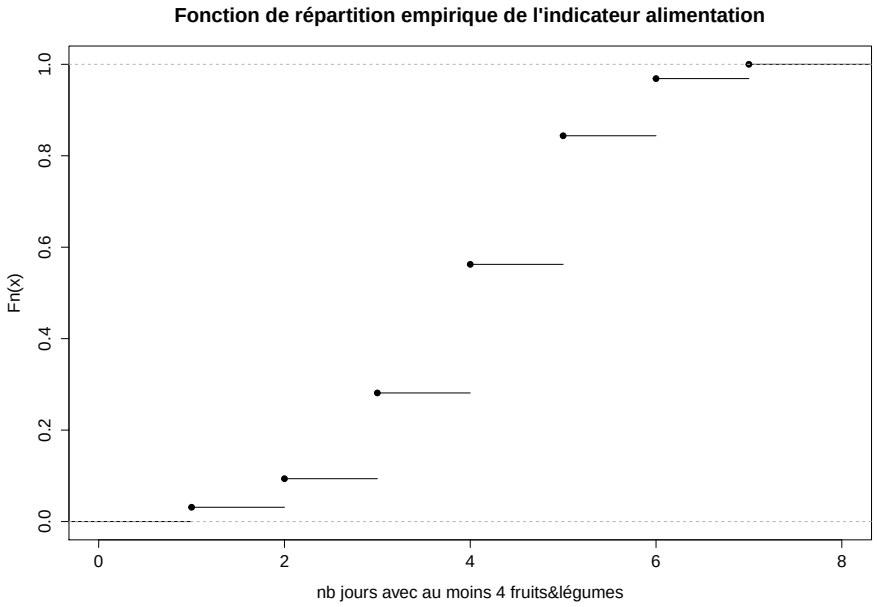

Figure 2.3: Fonction de répartition empirique de la variable alimentation (nombre de jours d'une semaine où au moins 4 fruits et légumes sont consommés par jour), représentée par une fonction en escaliers.

Interprétation : modalité la plus représentée dans l'échantillon.

Médiane : Valeur qui partage la population en deux effectifs (presque) égaux, notée $\hat{q}_{0.5}$. On définit donc la médiane $\hat{q}_{0.5}$ comme la plus petite modalité à partir de laquelle la fonction de répartition empirique $\hat{F}$ est supérieure ou égale à 0.5 (c.-à-d. $\hat{F}(\hat{q}_{0.5}^{-}) < 0.5$ et $\hat{F}(\hat{q}_{0.5}) \geq 0.5$).

Interprétation : modalité en dessous de laquelle (au sens strict) on trouve au plus la moitié des individus et au dessus de laquelle (au sens large) on trouve au moins la moitié des individus.

Graphiquement, c'est la plus petite valeur en laquelle le graphe de $\hat{F}$ franchit le palier 0.5.
La médiane est le quantile empirique d'ordre 0.5 et on étend naturellement cette notion à un ordre α quelconque ($\alpha \in [0, 1]$)).

Quantile empirique d'ordre α :

$$\hat{q}_\alpha = \inf\{x \in \mathcal{X}, \hat{F}(x) \geq \alpha\}$$

C'est en fait la plus petite modalité en laquelle la fonction de répartition empirique dépasse α. La médiane est le quantile d'ordre 0.5 et les quartiles sont les trois quantiles d'ordre $0.25; 0.5; 0.75$.

Moyenne : notée $\bar{x}_n$, moyenne arithmétique de l'ensemble des n réalisations de X c.-à-d. :

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{q} n_k m_k = \sum_{k=1}^{q} f_k m_k.$$

Interprétation : valeur qu'auraient tous les individus s'ils prennaient tous la même valeur.

– Caractéristiques de dispersion –

Afin de compléter les caractéristiques centrales on peut définir des mesures de dispersion telles que :

Ecart quadratique moyen (de l'échantillon à une cible x quelconque):

$$Q(x) = \frac{1}{n}\sum_{j=1}^{n}(x_j - x)^2 = \frac{1}{n}\sum_{k=1}^{q} n_k(m_k - x)^2 = \left[\frac{1}{n}\sum_{j=1}^{n} x_j^2\right] - 2x\bar{x} + x^2,$$

Ecart absolu moyen (de l'échantillon à une cible x quelconque):

$$e_m(x) = \frac{1}{n}\sum_{j=1}^{n}|x_j - x|.$$

L'écart quadratique moyen est une parabole qui atteint son minimum au point $\bar{x}_n$ et la valeur de ce minimum s'appelle la variance de l'échantillon. Par contre le minimum de l'écart absolu moyen est atteint au point $\hat{q}_{0.5}$.

Variance empirique de l'échantillon :

$$s_n^2 = Q(\bar{x}_n) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2 = \frac{1}{n}\sum_{k=1}^{q} n_k(m_k - \bar{x}_n)^2.$$

Ecart-type empirique de l'échantillon : s_n.

Propriété : $s_n^2 = \bar{x^2}_n - \bar{x}_n^2$ où $\bar{x^2}_n = \sum x_j^2/n$.

Remarque : sous R on obtient la variance empirique avec la fonction `var(x)*(length(x)-1)/length(x)` où `x` est l'objet qui contient l'échantillon des données brutes $(x_1, ..., x_n)$ (voir TP2), tandis que `var(x)` permet de calculer la version corrigée de la variance empirique définie par $s'^2 = n/(n-1)s^2$.

2.2 Variable aléatoire discrète

2.2.1 Loi de probabilité et FdR d'une variable aléatoire

La quantité x observée de X pour un individu choisi au hasard dans $\mathcal{P}$ est la réalisation d'une variable aléatoire X (v.a. X). On la caractérise par sa **loi de probabilité** définie dans le cas discret par :

$$\mathcal{X} = \{m_1, ..., m_q\} \qquad \text{et} \qquad \{p_k = P(X = m_k), 1 \leq k \leq q\}$$

Les p_k sont les probabilités théoriques qu'a la variable aléatoire X de prendre les modalités m_k. On a évidemment

$$p_k \in [0, 1] \qquad \text{et} \qquad \sum_k p_k = 1$$

Si l'échantillon de données étudié est constitué de n tirages indépendants de la même variable aléatoire X alors les fréquences f_k d'occurence de la modalité m_k, observées sur $(x_1, ..., x_n)$ approchent de mieux en mieux les p_k lorsque n augmente. Ce qui revient à dire que la répartition observée (c'est à dire la suite des f_k) approxime la loi de probabilité (c'est à dire la suite des p_k). De même la fonction de répartition empirique (basée sur les données observées $(x_1, ..., x_n)$) approxime la fonction de répartition théorique lorsque $n \to \infty$.

Fonction de répartition (FdR) : La fonction de répartition de la variable aléatoire X est définie sur $\mathbb{R}$ par :

$$\forall x \in \mathbb{R} \qquad F_X(x) = P(X \leq x) = \sum_{m_k \leq x} P(X = m_k).$$

C'est aussi une fonction en escaliers (avec $q + 1$ paliers) croissante où le saut entre le k-ième et le $(k + 1)$-ième palier vaut p_k.

On peut représenter la loi de probabilité de la v.a. X par un diagramme en barres (avec hauteurs p_k) et sa fonction de répartition par le graphe de F_X (en escaliers avec des sauts de hauteur p_k).

En utilisant la définition de F_X et les propriétés des probabilités sur les ensembles (qui décrivent des évènements relatifs à une expérience aléatoire) on établit :

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) = \sum_{a < m_k \leq b} P(X = m_k).$$

2.2.2 Quelques caractéristiques théoriques d'une variable aléatoire discrète

De même que l'on caractérise les tendances centrales et de dispersion d'un jeu de données $(x_1, ..., x_n)$, on définit la moyenne, l'écart type et la variance d'une variable aléatoire discrète X par :

Espérance mathématique de X le nombre :

$$E(X) = \sum_{m_k \in \mathcal{X}} m_k \, P(X = m_k).$$

Remarques :

- Lorsque toutes les probabilités sont égales (ie. $p_1 = p_k = p_q = 1/q$) l'espérance mathématique est égale à la moyenne arithmétique des éléments de $\mathcal{X}$.
- L'espérance de la variable aléatoire X sera aussi appelée dans ce cours *moyenne théorique* de X, et notée μ_X ou μ (la moyenne empirique est la moyenne arithmétique des données x_i). En général dans le cas discret, $\mu_X \notin \mathcal{X}$.

Variance de X le nombre :

$$\sigma^2 = V(X) = \sum_{m_k \in \mathcal{X}} (m_k - \mu)^2 P(X = m_k).$$

Ecart-type de X le nombre :

$$\sigma = \sqrt{V(X)}.$$

Quantile d'ordre α de X le nombre :

$$q_\alpha = \inf\{x \in \mathcal{X}, F_X(x) \geq \alpha\}.$$

Une variable de moyenne nulle ($E(X) = 0$) sera dite centrée et une variable de variance 1 ($V(X) = 1$) sera dite réduite.

On a les propriétés suivantes, très utiles, de l'espérance et de la variance qui sont satisfaites pour toutes les variables aléatoires (continues ou discrètes) :

Propriétés de l'espérance et de la variance
Soient deux variables aléatoires X et Y et deux nombres réels quelconques a et b :

1. $E(aX + b) = aE(X) + b$ et en particulier $E(b) = b$
2. $E(X + Y) = E(X) + E(Y)$
3. $V(aX + b) = a^2V(X)$ et en particulier $V(b) = 0$
4. si de plus X et Y sont telles que $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ elles sont dites indépendantes et $V(X + Y) = V(X) + V(Y)$.

2.2.3 Centrer et réduire

Pour centrer une variable on lui retire son espérance et pour la réduire on la divise par la racine de sa variance. On peut ainsi montrer que :

Définitions de la variable centrée et de la variable centrée réduite
Soit X d'espérance μ et variance σ^2 alors :

1. $E(X - \mu) = 0$ et $X - \mu$ est dite centrée
2. $V(X/\sigma) = 1$ et X/σ est dite réduite
3. $(X - \mu)/\sigma$ est la centrée réduite de X en effet :

$$E\left(\frac{X - \mu}{\sigma}\right) = 0 \qquad \text{et} \qquad V\left(\frac{X - \mu}{\sigma}\right) = 1$$

2.2.4 Modèles discrets usuels

Loi uniforme : on choisit au hasard un objet parmi n objets distincts avec la même probabilité d'obtenir chacun d'entre eux (par exemple lancé d'un dé non pipé). Ainsi la probablité de sortie de chaque objet numéroté de $k = 1, ..., n$ est $P(X = k) = 1/n$ si X représente le numéro de l'objet tiré.

Loi hypergéométrique : on choisit n individus au hasard dans une population de taille N et dont M remplissent une condition A (c. à d. une proportion $p = M/N$ qui satisfont la condition A). Soit X le nombre de personnes choisies parmi les n qui remplissent la condition A. X sera dite variable de loi hypergéométrique notée $\mathcal{H}(N, n, p)$ et pour tout $k \in \{\max(0, n-(N-Np)), ..., \min(Np, n)\}$:

$$P(X = k) = C_{Np}^k C_{N-Np}^{n-k} / C_N^n \quad \text{avec} \quad C_n^k = n!/((n-k)!k!).$$

Cette loi est peu utilisée car dès que $n << M$ on utilisera à la place la loi binômiale ci dessous :

Loi de Bernoulli et Loi binomiale : on considère n expériences telles que le lancer répétitif de pièces ou de dés, ou le tirage répété d'un individu dans un ensemble; chaque lancer est dit *essai*. Au cours de chacun des essais, à un événement particulier (c.-à-d. remplir une condition A) est associé une probabilité de réussite. Si les tirages ou les essais sont indépendants et réalisés dans les mêmes conditions on aura la même probabilité de réussite à chaque tirage.

Soit X_i la variable décrivant la réussite au i-ème tirage et $Y = \sum X_i$ le nombre cumulé de réussites parmi n tirages.

- **Loi de Bernoulli :** elle décrit la réalisation d'une expérience n'ayant que deux issues possibles, 1="succès" et 0="échec". La distribution d'une v.a. X de Bernoulli, notée $\mathcal{B}(p)$, est donnée par :

  $$P(X = 1) = p \quad \text{et} \quad P(X = 0) = 1 - p.$$

  On note que $q = 1 - p$ est la probabilité d'échec. Chaque variable X_i suit une loi de Bernoulli de paramètre p : "probabilité d'obtenir A".

- **Loi binomiale :** la probabilité pour que l'événement A (le succès) se réalise k fois exactement au cours de n essais est donnée par la probabilité :

  $$P(Y = k) = P\left(\sum_{i=1}^{n} X_i = k\right) = C_n^k p^k (1-p)^{n-k}, \quad \text{pour tout} \quad k \in \{0, ..., n\}.$$

  On dira que Y suit une loi binomiale $\mathcal{B}(n, p)$.

Remarques :

- La répartition de Y est appelée distribution binomiale, dans la mesure où, pour $k = 0, 1, 2, ..., n$, elle correspond aux termes successifs du développement du binôme :

  $$(q+p)^n = q^n + C_n^1 p^1 (1-p)^{n-1} + C_n^2 p^2 (1-p)^{n-2} + ... + p^n = \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} = \sum_{k=0}^{n} p_k.$$

- La loi hypergéométrique $\mathcal{H}(N, n, M)$ est approchée par la loi binomiale $\mathcal{B}(n, M/N)$, lorsque n est petit devant M.

Exemple : On admet qu'un étudiant a la même probabilité p chaque jour de consommer au moins quatre fruits ou légumes. On suppose de plus qu'il y a indépendance entre ses choix quotidiens. La variable A décrivant le nombre de jours de "bonne alimentation" de l'étudiant en une semaine est donc modélisée par une variable aléatoire de loi $\mathcal{B}(7, p)$.

Loi géométrique : c'est la loi du temps Z d'attente du premier succès dans les réalisations de tirages indépendants de variables de Bernoulli, $\mathcal{B}(p)$. Elle est notée $\mathcal{G}(p)$ et

$$P(Z = k) = (1-p)^{k-1} p, \quad \text{pour tout} \quad k \geq 1.$$

Loi de Poisson : cette distribution approche la loi binomiale $\mathcal{B}(n, \lambda/n)$ lorsque n est grand. C'est-à-dire que dans ce cas $P(Y = k) = C_n^k(\lambda/n)^k(1 - \lambda/n)^{n-k} \approx P(W = k)$ où W est une variable qui suit la loi de Poisson, notée $\mathcal{P}(\lambda)$, définie par

$$P(W = k) = \lambda^k e^{-\lambda}/k!, \quad \text{pour tout} \quad k \in I\!N.$$

Loi et Notation	$\mathcal{X}$	$P(X = k)$	$E(X)$	$V(X)$
uniforme $\mathcal{U}(n)$	$\{1, ..., n\}$	$1/n$	$(n+1)/2$	$(n+1)(n-1)/12$
hypergéométrique $\mathcal{H}(N, n, p), Np \in I\!N$	$k \geq \max(0, n-(N-Np))$ $k \leq \min(Np, n)$	$C_{Np}^k C_{N-Np}^{n-k}/C_N^n$	np	$np\,(1-p)\,\frac{N-n}{N-1}$
Bernoulli $\mathcal{B}(1, p)$	$\{0, 1\}$	$P(X = 1) = p$ $P(X = 0) = 1-p$	p	$p(1-p)$
binomiale $\mathcal{B}(n, p)$	$[0, n]$	$C_n^k p^k (1-p)^{n-k}$	np	$np(1-p)$
Poisson $\mathcal{P}(\lambda)$	$I\!N$	$e^{-\lambda}\lambda^k/k!$	λ	λ
géométrique $\mathcal{G}(p)$	$I\!N^*$	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

2.3 Adéquation entre distribution observée et loi

Considérons les données "alimentation" $(a_1, ..., a_n)$ proposées dans l'introduction. On se demande si la répartition observée sur $\mathcal{X} = \{m_1, ..., m_q\}$ ressemble à une distribution théorique sur $\mathcal{X}$ spécifiée par $(p_1^*, ..., p_q^*)$. Autrement dit, les valeurs a_i auraient-elles pu être obtenues comme n réalisations indépendantes d'une variable aléatoire X à valeurs dans $\mathcal{X}$ et ayant pour loi de probabilité $(p_1^*, ..., p_q^*)$? Ou encore, les fréquences observées $(f_1, ..., f_q)$ sont-elles "proches" des probabilités théoriques $(p_1^*, ..., p_q^*)$?

On visualise côte à côte dans la figure 2.4 la répartition observée, la répartition uniforme et la répartition binômiale sur un même graphique afin d'apprécier la proximité ou non de la répartition observée à l'une des lois de probabilité théoriques proposées.

D'après la figure précédente il semble visuellement clair que la loi de probabilité binômiale est plus proche de la répartition observée que la loi de probabilité uniforme. On peut quantifier l'écart entre la répartition observée et une loi de probabilité à l'aide de la distance du Chi-deux, définie par :

$$d^2 = \sum_{k=1}^{q} \frac{(n_k - np_k^*)^2}{np_k^*} = n \sum_{k=1}^{q} \frac{(f_k - p_k^*)^2}{p_k^*}.$$

Plus d^2 est proche de zéro meilleure est l'adéquation entre la loi théorique et la distribution observée. Les n_k sont les effectifs observés de la modalité m_k tandis que np_k^* sont les effectifs théoriques que l'on s'attend à obtenir sur n tirages indépendants sous la loi $(p_1^*, ..., p_q^*)$.

Par exemple, on souhaite comparer la distribution observée de la variable alimentation A d'une part à une loi uniforme

$$(p_1^*, ..., p_8^*) = (1/8, 1/8, ..., 1/8)$$

et d'autre part à une loi binomiale $\mathcal{B}(7, 0.5)$ donnée par les probabilités théoriques

$$(p_1^*, ..., p_8^*) = (0.008, 0.055, 0.164, 0.273, 0.273, 0.164, 0.055, 0.008).$$

On dresse le tableau suivant :

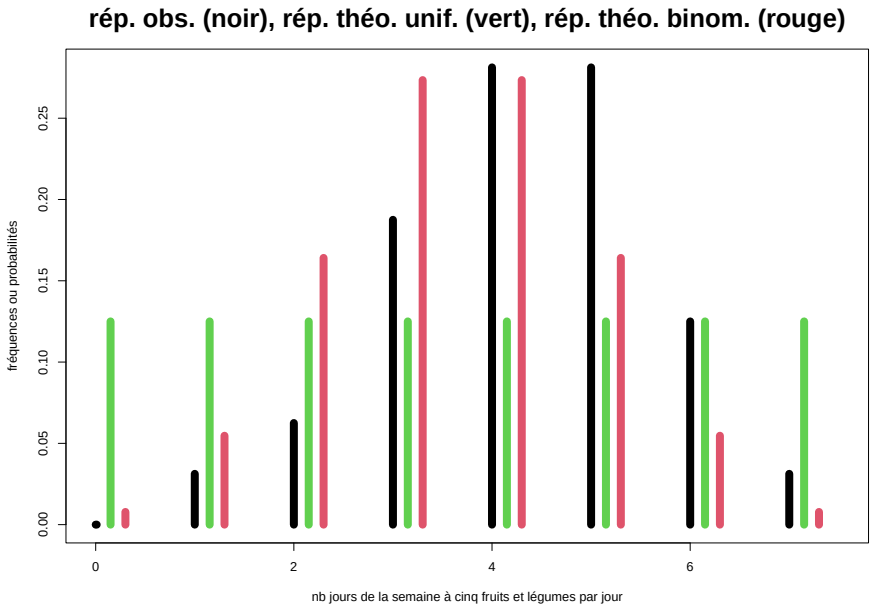

Figure 2.4: Répartition observée de la variable Alimentation et deux lois de probabilités envisagées comme modèles.

$\mathcal{X}$	0	1	2	3	4	5	6	7	d^2
Eff. obsv.	0	1	2	6	9	9	4	1	
Eff. theo. avec unif.	4	4	4	4	4	4	4	4	23
Eff. theo. avec binom.	0.25	1.75	5.25	8.75	8.75	5.25	1.75	0.25	11.276

Pour les données proposées, avec le modèle uniforme on obtient un d^2 plus grand qu'avec le modèle binomial. La $\mathcal{B}(7, 0.5)$ modélise donc bien mieux les données observées que l'uniforme sur $\{0, 1, 2, .., 7\}$.

Chapter 3

Variables aléatoires continues et données continues

Nous considérons dans cette partie des données observées dans un intervalle de $\mathbb{R}$, noté $]m, M]$. Par exemple les variables poids à 15 ans, poids à 20 ans ou taille seront considérées comme continues (les poids et la taille étant donnés au kg ou cm près on pourrait penser qu'il s'agit de variables discrètes sur l'ensemble des entiers mais comme on observe peu ou pas de répétitions dans les données il est préférable de les décrire comme des variables continues).

3.1 Analyse descriptive de données continues

Toutes les notions vues pour les données discrètes se "déclinent" dans le cas continu, moyennant quelques modifications naturelles. Pour cela on fait une partition de $]m, M]$ en q morceaux, appelés classes et notés $C_1, ..., C_q$. Et, $\forall k = 0, ..., q$, on note :

$e_k \in \mathbb{R}$: extrémité droite de la classe C_k,

$a_k = (e_k - e_{k-1})$: amplitude de la classe C_k et

$m_k = (e_k + e_{k-1})/2$: milieu de C_k.

3.1.1 Tableau de distribution et représentations graphiques

Ce tableau est défini comme pour les données discrètes, à ceci près qu'ici m_k désigne le centre de la k-ième classe au lieu de la k-ième modalité du cas discret.

Toutes les classes ne sont pas nécessairement de même amplitude, ce qui nous conduit à ajouter une colonne dans le tableau de distribution contenant les *fréquences relatives* à l'unité d'amplitude.

X	milieux	effectifs	fréquences	fréq. cumul.	fréq. rel.
$]e_0, e_1]$	m_1	n_1	$f_1 = n_1/n$	$F_1 = f_1$	f_1/a_1
...					
$]e_{k-1}, e_k]$	m_k	n_k	$f_k = n_k/n$	$F_k = f_1 + ... + f_k$	f_k/a_k
...					
$]e_{q-1}, e_q]$	m_q	n_q	$f_q = n_q/n$	$F_q = f_1 + ... + f_q = 1$	f_q/a_q

De même que dans le cas discret, on représente les fréquences observées et les fréquences cumulées avec les graphiques suivants:

- **Répartition observée :** avec un histogramme des fréquences (et non des effectifs) : graphe où sont portées en abscisses les extrémités de classes et où l'on trace un rectangle de surface f_k (de largeur a_k et hauteur f_k/a_k) au dessus de la classe k. La hauteur du rectangle k est la fréquence relative à la largeur de classe k.

- **Fonction de répartition empirique :** elle est définie comme dans le cas discret par $\hat{F}(x) = freq]-\infty, x]$. Si on ne dispose pas des données brutes (c. à d. de la suite $x_1, ..., x_n$) mais seulement des données en effectifs par classe alors on l'approche avec la fonction $\tilde{F}$ telle que $\tilde{F}(e_k) = F_k$ pour tout $k = 1, ..., q$, $\tilde{F}(e_0) = 0$ continue, linéaire par morceaux. Comme la fonction $\hat{F}$ elle est croissante, vaut 0 pour $x \leq e_0$ et 1 pour $x \geq e_q$. Son graphe est celui d'une fonction qui passe par les points $(-\infty, 0)$, $(e_0, 0)$ et (e_k, F_k) pour $k = 1, ..., q$, $(1, +\infty)$ relié par des segments.

Par exemple pour la variable poids on obtient les représentations suivantes de la répartition observée avec un histogramme ou avec un boxplot (plus grossier que l'histogramme) :

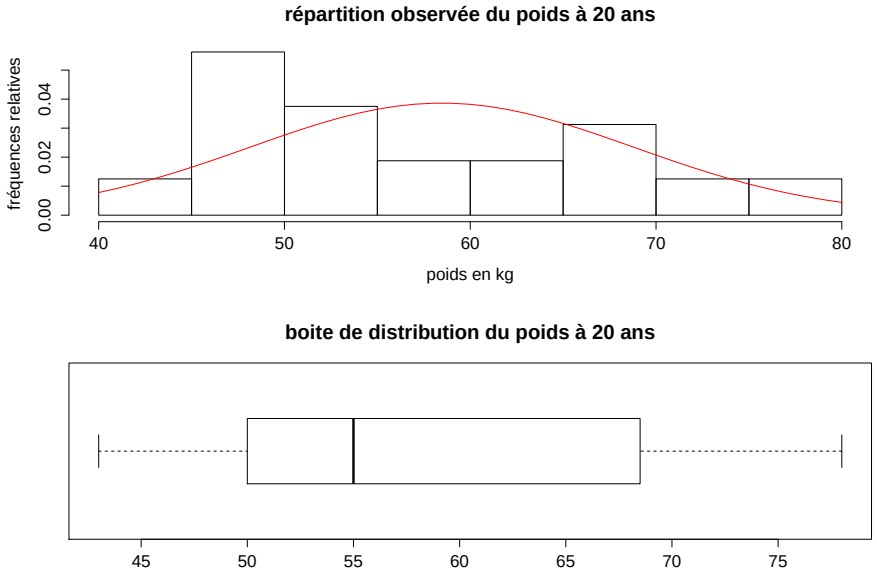

Figure 3.1: Répartition de la variable poids à 20 ans (histogramme et boxplot)

Un des intérêts du boxplot est de pouvoir représenter simultanément plusieurs échantillons de façon à les comparer. Ainsi dans la figure 3.2 il apparaît nettement que le poids est plus bas à 15 ans qu'à 20 ans mais également moins variable.

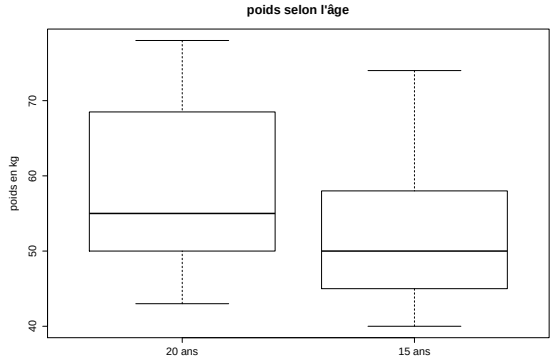

Figure 3.2: Boites de distribution du poids à selon l'âge

Pour représenter la répartition cumulée on utilise la fonction de répartition empirique définie comme dans le cas discret si on dispose des données brutes et la fonction $\tilde{F}$ si on dispose seulement des données en effectifs par classes (six classes dans l'exemple).

Dans le cas de la variable poids à 20 ans, avec des données en effectifs sur six classes on obtient la figure 3.3.

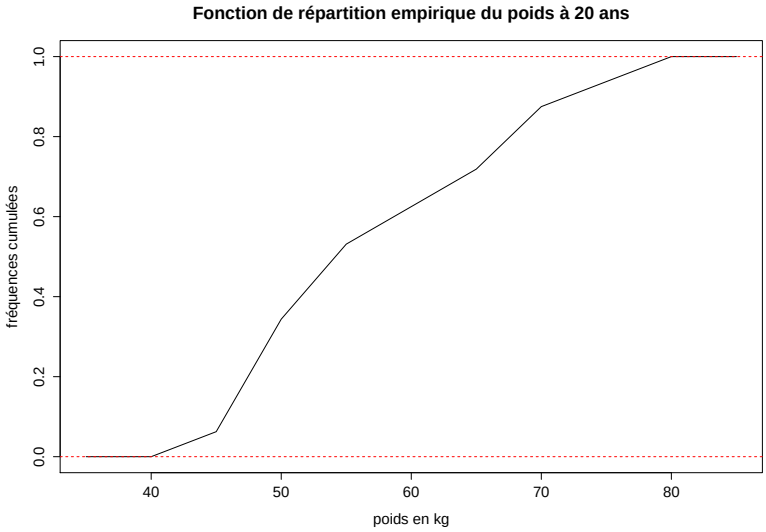

Figure 3.3: Répartition cumulée de la variable poids à 20 ans (approchant la FdR empirique)

3.1.2 Caractéristiques centrales et de dispersion

Classe Modale : classe pour laquelle l'histogramme des fréquences présente un maximum relatif.

Médiane : $\hat{q}_{0.5} \in \mathcal{X}$ telle que $\tilde{F}(\hat{q}_{0.5}) = 0.5$ si on a seulement les données en effectifs et si on a les données brutes elle est définie comme dans le cas discret.

Fractile d'ordre α : $\hat{q}_\alpha$ tel que $\tilde{F}(\hat{q}_\alpha) = \alpha$ avec les données en effectifs et comme dans le cas discret avec les données brutes.

Remarque : on appelle quartiles les quantiles qui partagent l'échantillon en quatre et percentiles ceux qui la partagent en cent.

Moyenne :

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \simeq \frac{1}{n}\sum_{k=1}^{q} n_k m_k = \sum_{k=1}^{q} f_k m_k.$$

Pour les caractéristiques de dispersion, nous considèrerons seulement l'écart type et l'amplitude de l'intervalle interquartile. La variance et l'écart type se définissent comme en discret :

Variance :

$$s_n^2 = Q(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \simeq \frac{1}{n}\sum_{k=1}^{q} n_k (m_k - \bar{x})^2.$$

Ecart type : s_x.

Intervalle inter-quartile : $[\hat{q}_{0.25}, \hat{q}_{0.75}]$.

3.2 Variables aléatoires continues

La courbe rouge superposée sur l'histogramme de la figure 3.1 est la densité d'une variable normale convenablement ajustée. Lorsque l'histogramme ressemble à une densité normale (voir ci-dessous des représentation de densités gaussiennes) on a tendance a poser un modèle gaussien sur les données, même lorsque la ressemblance est lointaine. D'autres critères graphiques comme la droite de Henry permettent de juger de la pertinence d'une telle hypothèse et nous verrons ultérieurement dans le cours qu'il est possible de faire un test statistique sur cette question (et d'en quantifier la pertinence). Quoiqu'il en soit lorsque l'on modélise une variable aléatoire continue on essaie toujours en premier la loi normale car elle a des propriétés aisément manipulables (entre autres la symétrie de la densité...) et qu'elle apparait naturellement dans les théorèmes limites fournis pas le calcul des probabilités (voir fin de ce chapitre). Nous rappelons dans la suite ce qui peut caractériser le comportement aléatoire d'une variable continue.

3.2.1 Loi de probabilité

Dans le cas où la variable X étudiée est à valeur dans un intervalle de $[m, M]$, la probabilité d'obtenir x, c'est-à-dire $P(X = x)$ sera nulle car autrement on aurait $P(m \leq X \leq M) = \sum_{x\in[m,M]} P(X = x) = \infty$! Par contre si la probabilité d'avoir $X = x$ est nulle, celle d'être autour de x ne l'est pas et on définit la loi de probabilité de la variable X à l'aide d'une fonction appelée densité et définie par :

$$f_X(x) \approx P(X \in [x, x + \delta])/\delta \quad \text{si} \quad \delta \quad \text{petit.}$$

On peut aussi définir une variable aléatoire à l'aide de sa fonction de répartition (qui est continue) donnée par :

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt \quad \text{pour tout} \quad x \in I\!R \quad \text{et} \quad f_X(x) = F'_X(x).$$

Notons que comme $P(X = x) = 0$, on a toujours $P(X \leq x) = P(X < x)$.

Définir une variable aléatoire continue sur $I\!R$ c'est donc se donner une fonction positive f_X telle que $\int_{-\infty}^{+\infty} f_X(t)dt = 1$ ou se donner F_X une fonction continue croissante à valeurs dans $[0, 1]$ telle que $F_X(-\infty) = 0$ et $F_X(\infty) = 1$. On a alors pour tout couple de nombres (a, b) :

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) \\ &= P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) = \int_a^b f_X(t)dt. \end{aligned}$$

3.2.2 Espérance Variance et α-quantiles

Ces quantités sont définies par analogie avec le cas discret, en remplaçant m_k par x, $P(X = m_k)$ par $f_X(x)dx$ et la somme $\sum$ par l'intégrale $\int$.
Espérance de la variable X :

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x)dx = \mu.$$

Variance de la variable X, le nombre :

$$V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f_X(x) dx = \sigma^2.$$

La variance étant évidemment positive on peut en définir la racine.

Ecart-type de la variable X :

$$\sigma = \sqrt{V(X)}.$$

Quantile d'ordre α de la variable X :

$$q_\alpha \text{ tel que } F_X(q_\alpha) = \alpha.$$

L'indépendance entre X **et** Y est définie mathématiquement par

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \text{ pour tous } (x, y).$$

Elle signifie que le résultat obtenu sur X n'a aucune incidence sur celui obtenu sur Y et vice versa.

Les propriétés de l'espérance et de la variance d'une variable continue, énoncées dans 2.2.2, sont les mêmes que celles d'une variable discrète.

3.2.3 Modèles continus usuels

Le modèle le plus couramment utilisé est le modèle gaussien défini par la loi normale (aussi appelée loi gaussienne). Un autre modèle très utilisé, notamment pour modéliser des durées de vie est donné par la loi exponentielle de paramètre $\lambda > 0$.

Loi exponentielle $\mathcal{E}(\lambda)$: Une variable X suit la loi exponentielle si sa densité est donnée par :

$$f_\lambda(x) = \lambda e^{-\lambda x} \quad \forall x \geq 0 \quad \text{et} \quad f_\lambda(x) = 0 \quad \forall x < 0.$$

L'espérance et la variance d'une variable de loi exponentielle sont $E(X) = 1/\lambda$ et $V(X) = 1/\lambda^2$.

Loi normale (ou loi gaussienne) : la densité d'une variable aléatoire normale d'espérance μ et de variance σ^2 a une forme de cloche symétrique autour de l'axe $x = \mu$ et de largeur proportionnelle à σ (afin de ne pas alourdir les notations et n'en ayant pas besoin dans ce cours sa densité n'est pas donnée ici). La loi est notée $\mathcal{N}(\mu, \sigma^2)$. On notera sa fonction de répartition $\Phi_{\mu,\sigma}$ et lorsque $\mu = 0$ et $\sigma = 1$ on parlera de la variable normale centrée et réduite et on conviendra pour simplifier les notations d'écrire Φ à la place de $\Phi_{0,1}$.

Dans la figure 3.4 suivante sont représentées plusieurs densités gaussiennes (normales) pour différents choix de paramètres μ et σ^2. Le paramètre μ de centrage permet de définir l'axe de symétrie de la densité (axe vertical passant par le point d'abscisse μ) et le paramètre σ détermine lui la "largeur" du pic. Si on garde en tête que la surface sous la courbe doit valoir 1, une densité correspondant à une petite valeur de σ et qui est très resserrée autour de sa moyenne sera une courbe avec un pic "haut" et pointu. A l'inverse plus σ sera grand plus la courbe sera applatie.

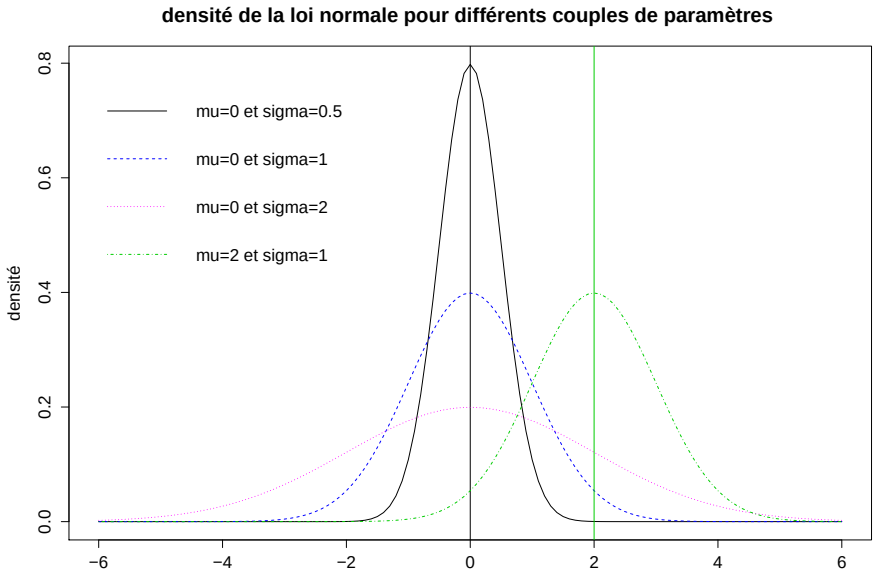

Figure 3.4: densités gaussiennes pour différents paramètres

La loi normale a plusieurs bonne propriétés : la première est que si on fait une transformation linéaire (αX) ou affine ($\alpha X + \beta$) d'une variable normale alors la variable résultante suit encore une loi normale. Ainsi on a la par la propriété suivante :

Propriété 1 :

- si X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ alors $U = (X - \mu)/\sigma$ suit une loi normale $\mathcal{N}(0, 1)$. Par conséquent :

$$\Phi_{\mu,\sigma}(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(U \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

- Réciproquement si U suit une loi $\mathcal{N}(0, 1)$ alors $X = \sigma U + \mu$ suit une loi $\mathcal{N}(\mu, \sigma^2)$.

La seconde propriété importante est que si l'on fait une combinaison linéaire de deux variables normales et indépendantes alors la variable résultante est elle aussi de loi normale :

Propriété 2 : si X_1 suit une loi normale $\mathcal{N}(\mu_1, \sigma_1^2)$ et si X_2 suit une loi normale $\mathcal{N}(\mu_2, \sigma_2^2)$ et sont indépendantes alors $aX_1 + bX_2$ suit une loi normale $\mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.

Une autre raison qui rend cette loi si attractive est qu'elle est dans des conditions raisonnables la distribution limite d'une moyenne sur n tirages indépendants d'une même variable X de loi non gaussienne. Cette question sera traitée dans le paragraphe suivant.

Par ailleurs, grâce aux méthodes numériques on sait calculer très précisément la fonction de répartition de la loi normale (elle est programmée sur n'importe quelle calculette de lycée ou logiciel statistique) et celle de la loi centrée réduite est tabulée sur papier. On conviendra dans la suite de réserver la notation Φ à la fonction de répartition de la $\mathcal{N}(0, 1)$.

Lecture de tables : la première table donne les couples $(x, \Phi(x))$ pour un certain nombre de valeurs de $x \geq 0$. La densité de la loi normale centrée réduite étant symétrique, lorsque $x \leq 0$ on utilisera

que $\Phi(x) = 1 - \Phi(-x)$. En effet par symétrie de la densité de la loi normale centrée réduite autour de l'axe des ordonnées on a pour tout x, $\Phi(x) = 1 - \Phi(-x)$ (un simple croquis à main levée de la densité, où sont placés x et $-x$ sur l'axe des abscisses et où les probabilités $\Phi(x)$ et $\Phi(-x)$ sont représentées par des surfaces sous la courbe de densité permet de s'en convaincre - à faire en exercice). On a aussi grâce à la symétrie la relation entre le quantile d'ordre p et celui d'autre $1-p$: $u_p = \Phi^{-1}(p) = -u_{(1-p)} = -\Phi^{-1}(1-p)$ (comme précédemment placer sur un croquis représentant la densité p, $1-p$, u_p et $u_{(1-p)}$).

Le plus souvent s'il n'y a aucune contre-indication flagrante on utilisera la loi Normale pour modéliser une variable continue.

Différentes méthodes graphiques (histogramme, droite de Henry,...) et quantitatives (tests statistiques de normalité) permettent d'apprécier si l'hypothèse de normalité sur la variable ayant produit un jeu de données est raisonnable.

La droite de Henry (qui sera utilisée en TP) est un nuage de points croisant n quantiles théoriques d'ordres $1/(n+1), 2/(n+1), ..., n/(n+1)$ de la loi normale centrée réduite (en abscisse) et les x_i (qui classés dans l'ordre croissants sont les quantiles empiriques d'ordres $1/(n+1), 2/(n+1), ..., n/(n+1)$, attention ils sont placés en ordonnée).

Utilisation de la Droite de Henry : lorsque les points sont alignés autour de la droite de Henry on pourra considérer que l'échantillon est issu d'une variable gaussienne. Par exemple la figure 3.5 représente les points et la droite obtenue dans le cas du poids à 20 ans. Cet indicateur graphique n'est pas très convainquant pour dire que le poids suit une loi gaussienne lui non plus (comme la superposition de la densité d'une gaussienne sur l'histogramme des poids de la figure 3.1).

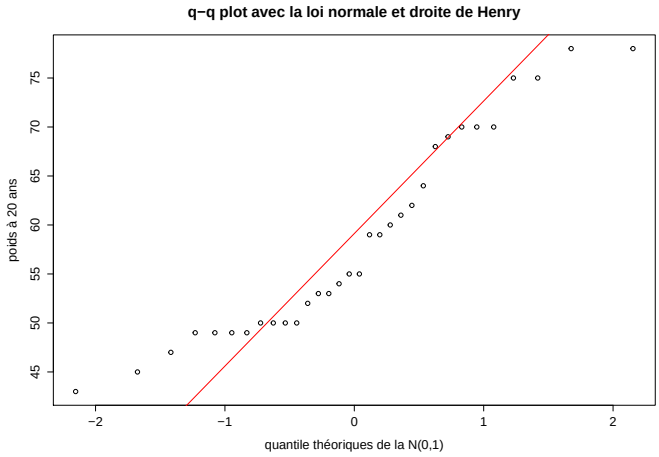

Figure 3.5: droite de Henry et qqplot pour le poids à 20 ans

A partir d'une ou plusieurs variables indépendantes de loi $\mathcal{N}(0,1)$ on peut définir d'autre lois telles que celles du Chi2, de Student ou de Fisher-Snedecor. Elle seront utilisées dans la suite pour décrire les lois de certaines statistiques permettant de calculer des intervalles de confiances ou faire des tests.

Loi du Chi2 : Une variable qui suit une loi du Chi2 à ν degrés de liberté s'obtient comme la somme de ν carrés de variables normales centrées réduites indépendantes. On la note $\mathcal{X}^2_\nu$. Sa densité est nulle

si $x \leq 0$ et sa fonction de répartition est tabulée.

Lecture de la table : on lit le couple $(z_{\nu,p}, p)$ qui satisfait $P(Z_\nu < z_{\nu,p}) = p$ lorsque Z_ν suit une loi du Chi2 à ν degrés de liberté, $\mathcal{X}^2_\nu$.

Dans la figure 3.6 sont représentées des densités de $\mathcal{X}^2_\nu$ pour différentes valeurs de ν. Plus ν est petit moins la densité est symétrique.

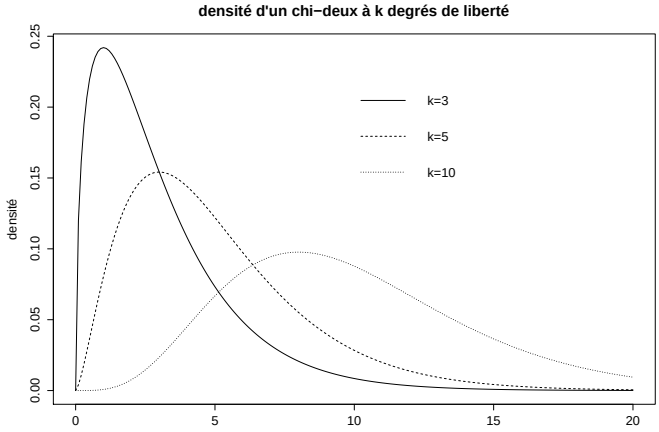

Figure 3.6: densité d'une variable $\mathcal{X}^2_k$

On construit ensuite la variable de Student avec la normale centrée réduite et une variable du Chi2.

Loi de Student : $T_\nu = U/\sqrt{V/\nu}$ avec U normale centrée réduite $\mathcal{N}(0,1)$ et V indépendante de U et de loi du Chi2 à ν degrés de libertés $\mathcal{X}^2_\nu$. Elle est notée $\mathcal{T}_\nu$ Cette densité ressemble beaucoup à celle de la $\mathcal{N}(0,1)$ et est aussi symétrique. On aura aussi dans ce cas l'expression du quantile d'ordre $1-p$ en fonction du quantile d'ordre p.

Lecture de la table : pour T_ν de loi de Student $\mathcal{T}_\nu$, on lit les couples $(t_{\nu,p}, p)$ tels que $P(T_\nu < t_{\nu,p}) = p$. pour tout $p > 1/2$ et pour $p < 1/2$ on utilisera la relation $t_{\nu,p} = -t_{\nu,1-p}$

Loi de Fisher-Snedecor : $F = (U/\nu_1)/(V/\nu_2)$ avec U et V deux variables indépendantes de lois respectives $\mathcal{X}^2_{\nu_1}$ et $\mathcal{X}^2_{\nu_2}$, suit une loi de Fisher-Snedecor à (ν_1, ν_2) degrés de libertés. Elle est notée $\mathcal{F}_{\nu_1,\nu_2}$.

Lecture de la table : pour F_{ν_1,ν_2} de loi $\mathcal{F}_{\nu_1,\nu_2}$ on lit sur les abaques f_{ν_1,ν_2} le quantile d'ordre p pour les valeurs de p les plus couramment utilisées $0.975, 0.995$. Lorsque $p = 0.005$ ou $p = 0.025$ on utilisera que $f_{\nu_1,\nu_2}(p) = 1/f_{\nu_2,\nu_1}(1-p)$ (en inversant F on obtient une $\mathcal{F}_{\nu_2,\nu_1}$ et comme on a que $P(F \leq f) = 1 - P(1/F \leq 1/f) = p$ qui implique que $f_{\nu_1,\nu_2}(p) = 1/f_{\nu_2,\nu_1}(1-p)$).

3.3 Suite de variables aléatoires et théorèmes limites

En statistique on travaille avec une suite de données $x_1, ..., x_n$ que l'on modélise comme une réalisation d'un échantillon aléatoire $X_1, ..., X_n$ défini comme :

Echantillon aléatoire de X: séquence de n variables $X_1, ..., X_n$ indépendantes et de même loi que X.

On utilise principalement les quantités suivantes fonctions de l'échantillon (qui sont les versions aléatoires des deux quantités $\bar{x}$ et s^2_x) :

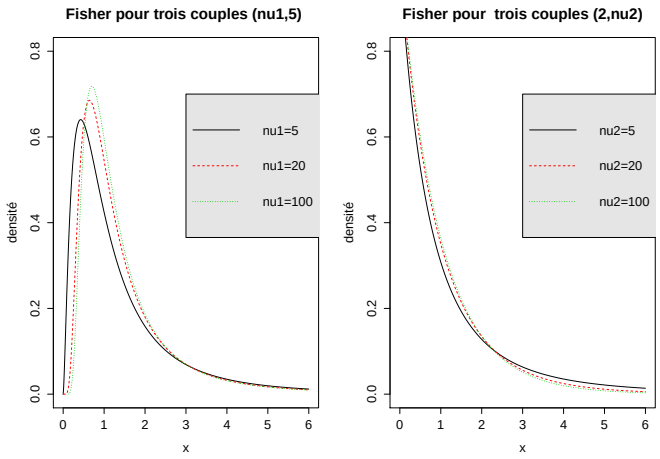

Figure 3.7: densité d'une variable $\mathcal{X}^2_k$

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

Propriétés :

1. $E(\bar{X}_n) = E(X) = \mu$ et $V(\bar{X}_n) = V(X)/n = \sigma^2/n$
2. $E(S_n^2) = (n-1)V(X)/n = (n-1)\sigma^2/n$
3. Si l'échantillon est gaussien $\bar{X}_n$ suit une loi normale pour tout n et nS_n^2/σ^2 suit une loi du $\mathcal{X}^2_{n-1}$

Lorsque les échantillons sont gaussiens les propriétés des statistiques étudiées découleront du résultat précédent (dont on pourra facilement démontrer le point 1 en exercice). En revanche lorsque l'échantillon n'est pas gaussien on pourra toujours établir les propriétés requises mais dans une version asymptotique, c'est à dire pour n grand à l'aide des deux théorèmes fondamentaux : loi (forte) des grands nombres et théorème central limite (TCL).

Loi des grands nombres : Soit $X_1, ..., X_n$ un échantillon de la variable X qui admet une valeur moyenne finie, soit $-\infty < E(X) < +\infty$, alors

$$\bar{X}_n \longrightarrow E(X) \quad \text{lorsque} \quad n \to +\infty$$

Le Théorème central limite : Soit $X_1, ..., X_n$ un échantillon de la variable X qui admet une variance finie, soit $V(X) < +\infty$, alors

$$\frac{\sqrt{n}(\bar{X}_n - E(X))}{\sqrt{V(X)}} \longrightarrow \mathcal{N}(0,1) \quad \text{lorsque} \quad n \to +\infty$$

Ce dernier résultat indique que la moyenne empirique d'un échantillon suit approximativement une loi normale $\mathcal{N}(E(X), V(X)/n) = \mathcal{N}(\mu, \sigma^2/n)$ à condition que n soit suffisamment grand. Dans le

cas du modèle binômial on demandera que $np > 10$ et $n(1-p) > 10$. En revanche si le modèle est normal ce résultat est vrai et exact pour tout $n \geq 1$.

Grâce au résultat fondamental énoncé dans le Théorème central limite, on montre que de nombreuses lois peuvent être approchées par la loi normale. C'est le cas, par exemple des lois Binômiale ou de Poisson. On retiendra

Propriété 3 : Soient X de loi $\mathcal{B}(n,p)$ et Y de loi $\mathcal{P}(\lambda)$ avec $\lambda \in \mathbb{N}^*$, alors on a les approximations suivantes :

- Une loi $\mathcal{B}(n,p)$ est approximativement une loi $\mathcal{N}(np, np(1-p))$ si n est grand :

$$\text{si} \quad np > 10 \quad \text{et} \quad n(1-p) > 10, \quad P(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right) \quad \forall x \in \mathbb{R};$$

- Une loi $\mathcal{P}(\lambda)$ est approximativement une $\mathcal{N}(\lambda, \lambda)$ si λ est grand.

$$\text{si} \quad \lambda > 10, \quad P(Y \leq y) \approx \Phi\left(\frac{y - \lambda}{\sqrt{\lambda}}\right) \quad \forall y \in \mathbb{R}.$$

La figure 3.8 suivante illustre la qualité d'approximation d'une binômiale, tracée en noir par une loi normale (convenablement choisie, c'est à dire de paramètre $\mu = np$ et $\sigma^2 = np(1-p)$) tracée en rouge.

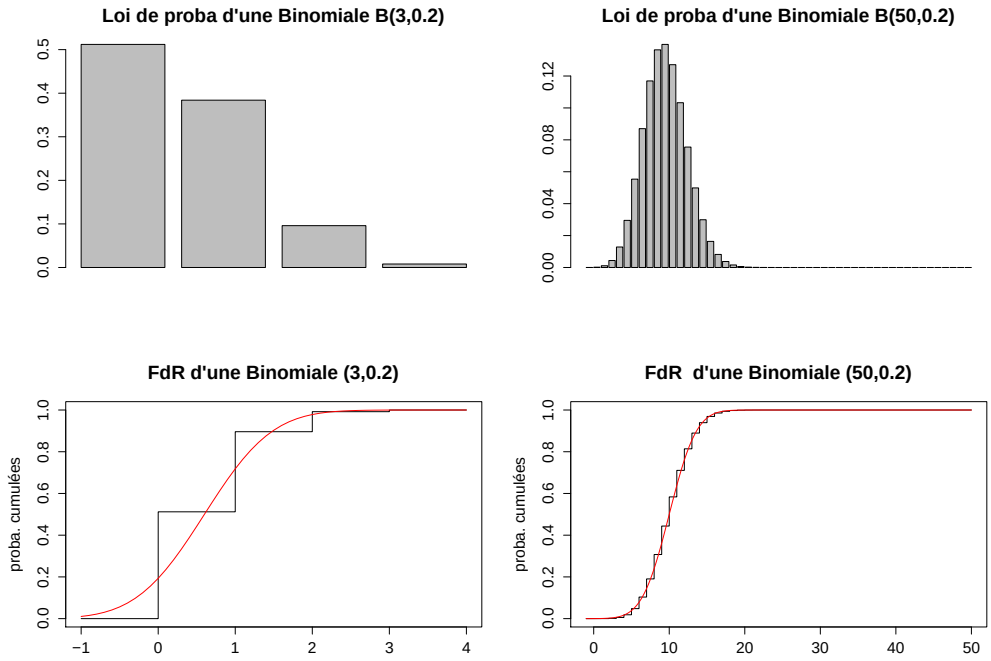

Figure 3.8: loi de proba et FdR d'une variable $\mathcal{B}(n,p)$ et son approximation normale $\mathcal{N}(np, np(1-p))$

On peut observer que pour n assez grand (colonne de droite) la qualité de l'approximation est assez bonne, ce qui n'est pas du tout le cas pour $n = 3$ et $p = 0.2$.

On retiendra *que pour un problème posé sur le paramètre p d'une loi de Bernoulli $\mathcal{B}(p)$ on ne pourra utiliser l'approximation de la loi de $\bar{X}_n$ par une $\mathcal{N}(p, p(1-p)/n)$ que lorsque n est tel que $np > 10$ et $n(1-p) > 10$.*

3.4 Intervalles de fluctuation

L'intervalle de fluctuation de niveau $1-p$, d'une variable aléatoire X de densité symétrique autour de son espérance est l'intervalle symétrique $I = [E(X) - a, E(X) + a]$ (centré en $E(X)$) tel que $P(X \in I) = 1 - p$.

Par exemple on établira en exercice que l'intervalle $[\mu - \sigma, \mu + \sigma]$ est un intervalle de fluctuation symétrique pour la variable X de loi normale de probabilité $68,5\%$ ... et que $[\mu - 3\sigma, \mu + 3\sigma]$ est un intervalle de niveau $99,7\%$.

On pourra retenir que :

Si X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ alors l'intervalle suivant est un intervalle de fluctuation de niveau $1-p$ pour X:

$$[\mu - u_{1-\frac{p}{2}}\sigma; \mu + u_{1-\frac{p}{2}}\sigma]$$

et en particulier on peut en déduire que pour un échantillon d'une variable X quelconque comme la loi de $\bar{X}_n$ est une $\mathcal{N}(\mu, \sigma^2/n)$ l'intervalle de fluctuation de niveau approximatif $1-p$ de $\bar{X}_n$ est donné par :

$$[\mu - u_{1-\frac{p}{2}}\frac{\sigma}{\sqrt{n}}; \mu + u_{1-\frac{p}{2}}\frac{\sigma}{\sqrt{n}}]$$

à condition que n soit assez grand (sauf dans le cas gaussien où il est de niveau exactement $1-p$ pour tout n).

Chapter 4

Estimation et Intervalles de confiance

Considérons la variable A : "alimentation d'un étudiant choisi au hasard dans le groupe". Nous avons vu, dans le premier chapitre, que la distribution de l'échantillon de données : $(a_1, ..., a_{32}) = (5, 4, 4, ..., 5, 4, 3)$ ressemblait plus à la répartition d'une $\mathcal{B}(7, 0.5)$ qu'à celle d'une répartition uniforme sur $\{0, 1, ..., 7\}$. On pourrait plus généralement comparer cette distribution observée à une $\mathcal{B}(7, p)$, où p serait ajusté à l'aide de $a_1, ..., a_{32}$. Supposons donc que A suit une $\mathcal{B}(7, p)$ avec p inconnu et essayons d'estimer p. L'espérance de A valant $7p$, on peut penser que cette moyenne théorique est assez bien approchée par la moyenne arithmétique des observations $\bar{a} = 4.22$. Il semble, alors raisonnable de proposer comme estimation de p, la quantité $\hat{p} = 4.22/7 = 0.603$. On peut ensuite, à l'aide de la distance d^2 voir si la $\mathcal{B}(7, 0.603)$ est meilleure que la $\mathcal{B}(7, 0.5)$. Le calcul du d^2 associé à la répartition de la $\mathcal{B}(7, 0.603)$ donnée par $(0.002, 0.017, 0.076, 0.191, 0.290, 0.264, 0.133, 0.029)$, donne $d^2 = 0.606$. Nous avions trouvé $d^2 = 11.276$, pour l'adéquation à une $\mathcal{B}(7, 0.5)$. La loi $\mathcal{B}(7, 0.603)$ modélise donc mieux les données que la $\mathcal{B}(7, 0.5)$. C'est donc $\mathcal{B}(7, 0.603)$ que l'on utilisera pour effectuer des prévisions sur n'importe quel individu de $\mathcal{P}$. Par exemple, on prévoira que la probabilité qu'un étudiant ne consomme jamais plus de quatre fruits ou légumes par jour sera environ de $P(A = 0) = C_7^0 0.603^0 (1 - 0.603)^7 = 0.16\%$.

4.1 Modèle et Echantillon aléatoire

Données : $x_1, ..., x_n$. On regardera x_i comme le i-ème tirage d'une variable aléatoire X ; ou de façon équivalente comme une réalisation (ou un tirage) d'une variable X_i de même loi que X. On supposera de plus que les variables X_i sont indépendantes.

Définition : On appelle *échantillon aléatoire* de taille n d'une variable X de loi P, l'ensemble $X_1, ..., X_n$ de variables indépendantes et de même loi que X. Un échantillon de données noté $x_1, ..., x_n$ est une réalisation (ou un tirage) de l'échantillon aléatoire $X_1, ..., X_n$.

Nous supposerons désormais que les données collectées $(x_1, ..., x_n)$ sont n tirages indépendants (ou un tirage de $(X_1, ..., X_n)$) d'une variable aléatoire X de loi P_θ, où le paramètre θ est un nombre réel inconnu. $\theta = p$ et $P_\theta = \mathcal{B}(7, p)$ dans l'exemple ci-dessus.

4.2 Estimation ponctuelle

Nous nous intéresserons dans la suite à l'estimation du paramètre θ décrivant la loi de X, dans les cas où $\theta = \mu_X$, $\theta = \sigma_X^2$ ou $\theta = p$ dans un modèle $\mathcal{B}(p)$ avec p inconnu.

Dans la suite, on considère $X_1, ..., X_n$, un échantillon aléatoire de la variable X de moyenne inconnue $\mu_X = \mu$ et de variance inconnue $\sigma_X^2 = \sigma^2$.

Estimateur : un estimateur de θ, est une variable aléatoire construite à l'aide des X_i. Une estimation de θ, notée $\hat{\theta}$, est l'application d'un estimateur aux données $x_1, ..., x_n$.

Un même estimateur appliqué à différents jeux de données produira des *estimations différentes.*

ex : X_1, $\sum X_i$ ou $\sum X_i/n$ sont des estimateurs de $E(X) = \mu_X$.

On peut construire autant d'estimateurs que l'on veut du paramètre inconnu θ (toute fonction connue de $X_1, ..., X_n$) mais ils ne sont pas tous équivalents. Une des premières propriétés que l'on souhaite vérifier est que l'estimateur ajuste bien le paramètre d'intérêt θ, c'est à dire qu'il est sans biais.

Estimateur sans biais : Un estimateur $T_n = T(X_1, ..., X_n)$ sera dit sans biais pour estimer θ si $E(T_n) = \theta$.

Dire que T_n est sans biais revient à vérifier qu'en moyenne il permet de retrouver assez correctement θ. On peut également souhaiter qu'il soit de plus en plus précis lorsque n augmente, c'est-à-dire que sa variance diminue avec n.

Estimateur convergent : T_n un estimateur sans biais de θ sera dit convergent (en moyenne quadratique) si $V(T_n)$ tend vers 0 lorsque $n \to \infty$.

4.2.1 Estimation d'une moyenne

Soit un échantillon d'une variable aléatoire X où on note $E(X) = \mu$ et $V(X) = \sigma^2$. On rappelle :

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

Estimateur : $\bar{X}_n$ est un estimateur sans biais et convergent de $\mu = E(X)$.

L'estimation $\hat{\mu}$ de μ obtenue par l'application de l'estimateur $\bar{X}$ aux données est $\bar{x}$. Par abus de langage, on appellera le résultat d'un estimateur sans biais estimation non biaisée du paramètre.

Loi de l'estimateur : Si la variable X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ alors $\bar{X}$ suit une loi normale $\mathcal{N}(\mu, \sigma^2/n)$.

4.2.2 Estimation d'une variance

On définit :

$$V_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2, \quad S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \left[\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right] - \bar{X}^2 \quad \text{et} \quad S_n'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Estimateurs :

- Si μ est connue V^2 est un estimateur de σ^2 et il est sans biais.

- Si μ est inconnue S^2 est un estimateur biaisé de σ^2 et S'^2 est un estimateur sans biais. En effet :

$$E(S^2) = \frac{n-1}{n}\sigma^2 \quad \text{et} \quad E(S'^2) = \sigma^2$$

Dans le cas de l'échantillon gaussien, on dispose également de la loi de ces estimateurs.

Loi des estimateurs :
Pour X de loi $\mathcal{N}(\mu, \sigma^2)$:

- nV_n^2/σ^2 suit une loi du Chi2 à n degrés de liberté : $\mathcal{X}_n^2$
- $nS_n^2/\sigma^2 = (n-1)S_n'^2/\sigma^2$ suit une loi du Chi2 à $n-1$ degrés de liberté : $\mathcal{X}_{n-1}^2$

4.2.3 Estimation d'une proportion

L'estimation d'une proportion peut être vu comme un problème d'estimation de moyenne. En effet, soit p la proportion dans une population $\mathcal{P}$ d'individus remplissant une condition A. Dans un échantillon de taille n, on observe une proportion, f_n, d'individus qui vérifient A. On souhaite estimer p à l'aide de f_n.

On pose le modèle suivant : soit X_i la variable indicatrice du succès de A pour l'individu i de l'échantillon. L'échantillon aléatoire $X_1, ..., X_n$ est celui d'une variable X de loi de Bernoulli $\mathcal{B}(p)$. On a alors $f_n = \sum x_i/n$, puisque $\sum x_i$ est le nombre d'individus dans l'échantillon observé qui remplissent A. Notons $F_n = \sum X_i/n$ la variable aléatoire associée.

Le problème d'estimation de p dans ce cas, est celui de l'estimation de la moyenne inconnue $\mu = p$ d'une variable X de Bernoulli $\mathcal{B}(p)$.

Estimateur : $F_n = \bar{X}$ est un estimateur sans biais et convergent de p. En effet $E(F_n) = p$ et $V(F_n) = p(1-p)/n$.

Lois de l'estimateur :

- nF_n suit une loi binomiale $\mathcal{B}(n, p)$
- si $np > 10$ et $n(1-p) > 10$, $\frac{\sqrt{n}(F_n - p)}{\sqrt{p(1-p)}}$ suit approximativement une loi normale $\mathcal{N}(0, 1)$. Ce qui se traduit par

$$P\left(\frac{\sqrt{n}(F_n - p)}{\sqrt{p(1-p)}} \leq t\right) \approx \Phi(t).$$

- si $np > 10$ et $n(1-p) > 10$, $\frac{\sqrt{n}(F_n - p)}{\sqrt{F_n(1-F_n)}}$ suit approximativement une loi normale $\mathcal{N}(0, 1)$. C'est-à-dire

$$P\left(\frac{\sqrt{n}(F_n - p)}{\sqrt{F_n(1-F_n)}} \leq t\right) \approx \Phi(t).$$

4.3 Intervalles de confiance (IC)

Dans le paragraphe précédent ont été proposées des évaluations ponctuelles du paramètre θ. Plutôt qu'une seule valeur du paramètre, on souhaite à présent donner une fourchette de deux valeurs entre lesquelles "on s'attend" à trouver θ. Au lieu de donner un estimateur de θ, cela revient à s'en donner deux qui encadrent θ avec une probabilité $1 - \alpha$ fixée.

Définition : Un intervalle de confiance de niveau de confiance $1-\alpha$ pour le paramètre θ est défini par :

$$I(\theta,\alpha) = [T_1, T_2] \quad \text{tel que} \quad P(T_1 \leq \theta \leq T_2) = 1-\alpha$$

avec $T_1 = f_1(X_1, ..., X_n)$ et $T_2 = f_2(X_1, ..., X_2)$ deux fonctions connues de l'échantillon aléatoire.

Remarques :

- α étant la probabilité que l'intervalle aléatoire $I(\theta,\alpha)$ ne contienne pas le paramètre inconnu θ, on le choisit en général petit.
- T_1 et T_2 sont en fait des fonctions d'estimateurs sans biais de θ.
- un intervalle de confiance est aléatoire et par abus de langage on dira aussi intervalle de confiance pour désigner l'application de l'intervalle $I(\theta,\alpha)$ au jeu de données qui produira un intervalle calculé. La réalisation de $I(\theta,\alpha)$ sera notée $i(\theta,\alpha)$. *Un même intervalle* $I(\theta,\alpha)$ appliqué à différents jeux de données produira des *"fourchettes"* : $i(\theta,\alpha)$ *différentes.*

Construction d'un IC :

1. construire un estimateur (de préférence sans biais et convergent) de θ : T ;
2. trouver une fonction simple de T et θ, $g(T,\theta)$, dont la loi est connue (c.-à-d. elle ne dépend pas des paramètres inconnus du modèle) ;
3. en partant de $T_1 \leq \theta \leq T_2$ trouver une inégalité équivalente de la forme $A \leq g(T,\theta) \leq B$; en utilisant la table de la loi de $g(T,\theta)$ ajuster A et B pour que $P(g(T,\theta) \leq A) = \alpha/2$ et $P(g(T,\theta) \geq B) = \alpha/2$;
4. retrouver à partir de l'inégalité $A \leq g(T,\theta) \leq B$ l'inégalité équivalente $T_1 \leq \theta \leq T_2$.

4.3.1 Intervalles de confiance pour la moyenne d'une variable normale

Soit $X_1, ..., X_n$, un échantillon de la loi $\mathcal{N}(\mu,\sigma^2)$. Soient U une variable normale centrée réduite et T_n une variable de loi de Student à n degrés de liberté. On rappelle qu'est noté u_α (respc. $t_{n,\alpha}$) le quantile d'ordre α de U (respc. de T_n). Ils vérifient :

$$P(U \leq u_\alpha) = \alpha \quad \text{et} \quad P(T_n \leq t_{n,\alpha}) = \alpha.$$

Les intervalles suivants sont des intervalles de confiance de niveau de confiance $1-\alpha$ et symétriques pour $E(X) = \mu$ selon que σ^2 est connu ou pas.

σ^2 connue :

$$I(\mu,\alpha,\sigma^2) = \left[\bar{X} - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \bar{X} + \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right]$$

σ^2 inconnue :

$$I(\mu,\alpha) = \left[\bar{X} - \tfrac{S'}{\sqrt{n}} t_{n-1,1-\alpha/2}, \bar{X} + \tfrac{S'}{\sqrt{n}} t_{n-1,1-\alpha/2}\right]$$

4.3.2 Intervalles de confiance pour la variance d'une variable normale

Soit $X_1, ..., X_n$, un échantillon de la loi $\mathcal{N}(\mu,\sigma^2)$. Soit Z_n une variable qui suit une loi du Chi2 à n degrés de liberté. On note $z_{n,\alpha}$ le nombre :

$$P(Z_n \leq z_{n,\alpha}) = \alpha.$$

Selon que l'on connait ou non μ on a les intervalles suivants :

μ **connue :**	μ **inconnue :**
$J(\sigma^2, \alpha, \mu) = \left[\frac{nV^2}{z_{n,1-\alpha/2}}, \frac{nV^2}{z_{n,\alpha/2}}\right]$	$J(\sigma^2, \alpha) = \left[\frac{nS^2}{z_{n-1,1-\alpha/2}}, \frac{nS^2}{z_{n-1,\alpha/2}}\right]$

Si la variable aléatoire qui modélise le caractère étudié n'est pas une variable normale, les résultats précédents restent valables à condition que n soit assez grand.

4.3.3 Intervalles de confiance pour une proportion

Soit $X_1, ..., X_n$, un échantillon de la loi $\mathcal{B}(p)$ alors l'intervalle de confiance suivant est approximativement de niveau $1-\alpha$, si $np > 10$ et $n(1-p) > 10$. C'est-à-dire

$$P(p \in I(p, \alpha)) \simeq 1 - \alpha.$$

$$I(p, \alpha) = \left[F_n - \frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}} u_{1-\alpha/2}, F_n + \frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}} u_{1-\alpha/2}\right]$$

Chapter 5

Tests paramétriques

A propos de la variable alimentation nous avons *estimé* avec l'échantillon observé que le nombre moyen de jours de "bonne alimentation" (nombre quotidien de fruits et légumes supérieur à 4) était de 4.22. On sait de plus, que si ce nombre moyen (c.à d. l'espérance ou moyenne théorique de la variable A) est inférieur à cinq une campagne de prévention et d'information sera mise en place par les organismes de santé publique auprès de la population étudiante. Pour arrêter une décision on dispose de l'échantillon observé qui indique $\hat{p} = 4.22/7 = 0.603$. Cela nous permet-il de déduire $p < 5/7$? Accepter (ou refuser) $p < 5/7$, au vu des données ne se fait pas sans risque de se tromper. Le but de cette partie sera donc de se donner des règles de décision ou de choix entre deux situations pour lesquelles on sait contrôler le risque de donner une mauvaise conclusion.

5.1 Généralités

Dans l'exemple de l'alimentation : A la variable d'intérêt caractérisée par le paramètre $\theta = p$ est une variable $\mathcal{B}(7, p)$. Pour répondre à la question : "est-il utile de lancer une campagne?", il faut décider entre $p \geq 5/7$ et $p < 5/7$. En pratique cela revient au même que de décider entre $p = 5/7$ et $p < 5/7$. De façon générale, le test $\mathcal{H}_0 : \theta \leq \theta_0$ contre $\mathcal{H}_1 : \theta > \theta_0$ sera remplacé par le test plus simple $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta > \theta_0$. De même $\mathcal{H}_0 : \theta \geq \theta_0$ contre $\mathcal{H}_1 : \theta < \theta_0$ sera remplacé par le test plus simple $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta < \theta_0$. On pourra également être amené à traiter le cas $\mathcal{H}_0 : \theta = \theta_0$ contre $\mathcal{H}_1 : \theta \neq \theta_0$.

Les deux premiers tests sont dits unilatéraux et le dernier bilatéral. On se fixe θ_0. Dans la suite on désignera les tests par leur type défini par :

- **type 1** : $\mathcal{H}_0 : \theta = \theta_0 \qquad \mathcal{H}_1 : \theta > \theta_0$
- **type 2** : $\mathcal{H}_0 : \theta = \theta_0 \qquad \mathcal{H}_1 : \theta < \theta_0$
- **type 3** : $\mathcal{H}_0 : \theta = \theta_0 \qquad \mathcal{H}_1 : \theta \neq \theta_0$

Faire un test, c'est construire *une règle de décision* qui, à l'échantillon observé, associe l'une ou l'autre des conclusions : on rejette $\mathcal{H}_0$ (c.-à-d. on accepte $\mathcal{H}_1$) ou on ne rejette pas $\mathcal{H}_0$ (c.-à-d. on refuse $\mathcal{H}_1$). La règle de décision sera définie de la façon suivante :

- si $\hat{\theta} \in W$, on refuse $\mathcal{H}_0$ (c.-à-d. on accepte $\mathcal{H}_1$)
- si $\hat{\theta} \notin W$, on refuse $\mathcal{H}_1$ (c.-à-d. on accepte $\mathcal{H}_0$)

On notera dans la suite T l'estimateur de θ qui appliqué aux données $(x_1, ..., x_n)$ fournit l'estimation $\hat{\theta}$.

5.1.1 Les risques

Au vu des données, on rejettera ou non $\mathcal{H}_0$. On ne pourra prendre de décision sans risque de se tromper. Il y a deux risques possibles : celui de rejeter $\mathcal{H}_0$ à tort ou celui de rejeter $\mathcal{H}_1$ à tort. On définit les deux risques d'erreurs comme :

Risque de première espèce	**Risque de seconde espèce**
$\alpha = P(\text{refus de } \mathcal{H}_0 \mid \mathcal{H}_0 \text{ vraie})$ $= P_{\mathcal{H}_0}(T \in W)$	$\beta = P(\text{refus de } \mathcal{H}_1 \mid \mathcal{H}_1 \text{ vraie})$ $= P_{\mathcal{H}_1}(T \notin W)$

Définition : On dira que *faire un test de niveau* α (ou *tester au seuil* α), $0 \leq \alpha \leq 1$, c'est construire une région de rejet de $\mathcal{H}_0$, notée W_α telle que $\alpha = P_{\mathcal{H}_0}(T \in W_\alpha)$.

On retiendra donc que :

si dans un test de niveau α, on est conduit à rejeter $\mathcal{H}_0$, alors on conclura $\mathcal{H}_1$ avec un risque de se tromper α.
si au contraire, dans un test de niveau α, on est conduit à ne pas rejeter $\mathcal{H}_0$, alors on conservera $\mathcal{H}_0$ avec un risque de se tromper β.

Propriété : Pour les trois tests (1, 2 et 3) qui nous intéressent, la somme des risques de première et seconde espèce vaut 1.

5.1.2 Le choix de $\mathcal{H}_0$ et $\mathcal{H}_1$

Si on veut limiter le risque de refuser $\mathcal{H}_0$ quand elle est vraie, on se donne α petit. Ceci aura pour conséquence de prendre plus de risque de refuser $\mathcal{H}_1$ à tort, puisqu' il vaut $\beta = 1 - \alpha$. Cela aura également pour conséquence qu'en cas d'acceptation de $\mathcal{H}_1$ on prend un risque faible de se tromper puisque qu'il vaut α.

A l'inverse, si c'est $\mathcal{H}_0$ que l'on souhaite valider, en prenant peu de risque de l'accepter à tort (c.-à-d. on limite le risque $\beta = 1 - \alpha$ de refuser $\mathcal{H}_1$ à tort) alors on choisit α grand. En pratique, on prendra comme première règle de choix des hypothèses :

règle : mettre sous $\mathcal{H}_1$ l'hypothèse que l'on souhaite valider, sous $\mathcal{H}_0$ celle que l'on ne veut pas refuser à tort trop souvent et prendre α petit. De plus dans les tests étudiés ici on met toujours l'égalité à un paramètre cible θ_0 sous l'hypothèse nulle.

Par exemple, imaginons qu'au vu de données radar on souhaite choisir entre "un missile se dirige vers nous" et "aucun missile ne se dirige vers nous". Il semble évident dans ce cas, qu'il est beaucoup plus grave de conclure à tort "aucun missile ne se dirige vers nous" que de conclure à tort "un missile se dirige vers nous". L'hypothèse que l'on veut voir refusée le moins souvent à tort est dans ce cas "un missile se dirige vers nous" et celle que l'on voudrait voir validée est "aucun missile ne se dirige vers nous". L'application de la règle ci-dessus conduit à poser : $\mathcal{H}_0$: "un missile se dirige vers nous", $\mathcal{H}_1$: "aucun missile ne se dirige vers nous" **et** α **petit**.

Autre exemple : on étudie un nouveau vaccin contre la grippe et on souhaite vérifier qu'il est plus efficace que le vaccin habituel dont les effets secondaires sont bien connus et qui de plus est économique. On souhaite montrer, statistiquement, que ce nouveau traitement est meilleur que le vaccin habituel mais en limitant le risque de le juger meilleur à tort, car ce nouveau vaccin est cher et l'on en connaît pas encore les effets secondaires. L'application de la règle nous conduit alors à poser : $\mathcal{H}_0$: "le nouveau vaccin n'est pas meilleur que le vaccin habituel", $\mathcal{H}_1$: "le nouveau vaccin est meilleur que le vaccin habituel" **et** α **petit**.

La règle doit cependant être appliquée sous la contrainte que l'hypothèse à valider, placée sous $\mathcal{H}_1$, soit de la forme $\mathcal{H}_1 : \theta > \theta_0$; $\mathcal{H}_1 : \theta < \theta_0$ ou $\mathcal{H}_1 : \theta \neq \theta_0$.

Par construction il y a une disymétrie entre les deux hypothèses et lorsque la conclusion d'un test au niveau α est de ne pas rejeter $\mathcal{H}_0$ soit ne pas valider $\mathcal{H}_1$ alors on dira que l'on ne peut pas rejeter $\mathcal{H}_0$ en prenant un risque de le rejeter à tort α. Autrement dit on ne peut pas valider $\mathcal{H}_1$ de façon statistiquement significative.

5.1.3 Construction d'un test

1. Se donner un estimateur T de θ.
2. Se donner comme forme de région critique la même que celle décrite par l'hypothèse alternative c. à d. :
   - si $\mathcal{H}_1 : \theta > \theta_0$, $W_\alpha = \{T > C_\alpha\}$
   - si $\mathcal{H}_1 : \theta < \theta_0$, $W_\alpha = \{T < C_\alpha\}$
   - si $\mathcal{H}_1 : \theta \neq \theta_0$, $W_\alpha = \{T > C_\alpha \text{ ou } T < C'_\alpha\}$.
3. Ajuster la quantité C_α pour que
$$\alpha = P_{\mathcal{H}_0}(T \in W_\alpha) = P_{\theta=\theta_0}(T \in W_\alpha).$$
   Pour cela on aura besoin de connaître la loi de T (ou plutôt celle de la fonction $g(T, \theta)$ utilisée dans les constructions d'IC) lorsque l'hypothèse $\mathcal{H}_0$ est vraie.
4. Décider :
   - si $\hat{\theta}$, réalisation de T, est dans W_α, conclure $\mathcal{H}_1$ avec un risque de se tromper de α ;
   - si $\hat{\theta}$, réalisation de T, n'est pas dans W_α, conclure $\mathcal{H}_0$ avec un risque de se tromper de $\beta = 1 - \alpha$. Et on donnera plutôt comme conclusion littérale : les données ne permettent pas de réfuter l'hypothèse nulle si on souhaite garantir cette conclusion avec un risque de se tromper α.

Nous allons décliner cette méthodologie dans les trois cas qui nous intéressent : tests sur la moyenne et la variance d'une population dans un modèle gaussien et test sur une proportion.

5.2 Test sur la moyenne μ

Modèle : $X_1, ..., X_n$, échantillon aléatoire de la variable X de loi $\mathcal{N}(\mu, \sigma^2)$ avec $\theta = \mu$ inconnu, σ^2 éventuellement connue et $\theta_0 = \mu_0$ fixé.

Pour les tests 1, 2 et 3 on notera W_α^j la région de rejet du test de type j de niveau α. Les régions de rejet dépendent de la connaissance ou non de σ^2 et sont données par :

σ^2 **connue**	σ^2 **inconnue**
$W_\alpha^1 = \left\{\bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}}\phi^{-1}(1-\alpha)\right\} = \left\{\bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right\}$	$W_\alpha^1 = \left\{\bar{X} > \mu_0 + \frac{S'}{\sqrt{n}}t_{n-1,1-\alpha}\right\}$
$W_\alpha^2 = \left\{\bar{X} > \mu_0 - \frac{\sigma}{\sqrt{n}}\phi^{-1}(1-\alpha)\right\} = \left\{\bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right\}$	$W_\alpha^2 = \left\{\bar{X} > \mu_0 - \frac{S'}{\sqrt{n}}t_{n-1,1-\alpha}\right\}$
$W_\alpha^3 = \left\{\lvert\bar{X} - \mu_0\rvert > \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right\} =$ $\left\{\bar{X} < \mu_0 - \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}} \text{ ou } \bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}\right\}$	$W_\alpha^3 = \left\{\lvert\bar{X} - \mu_0\rvert > \frac{S'}{\sqrt{n}}t_{n-1,1-\frac{\alpha}{2}}\right\} =$ $\left\{\bar{X} < \mu_0 - \frac{S'}{\sqrt{n}}t_{n-1,1-\frac{\alpha}{2}} \text{ ou } \bar{X} > \mu_0 + \frac{S'}{\sqrt{n}}t_{n-1,1-\frac{\alpha}{2}}\right\}$

5.3 Test sur la variance σ^2

Modèle : $X_1, ..., X_n$, échantillon aléatoire de la variable X de loi $\mathcal{N}(\mu, \sigma^2)$ avec $\theta = \sigma^2$ inconnu, μ éventuellement connue et $\theta_0 = \sigma_0^2$ fixé.

Pour les tests 1, 2 et 3 on notera W_α^j la région de rejet de niveau α du test de type j de niveau α (avec $0 \leq \alpha \leq 1$). Les régions de rejet dépendent de la connaissance ou non de μ et sont données par :

μ **connue**	μ **inconnue**
$W_\alpha^1 = \left\{V^2 > \sigma_0^2 \frac{z_{n,1-\alpha}}{n}\right\}$	$W_\alpha^1 = \left\{S'^2 > \sigma_0^2 \frac{z_{n-1,1-\alpha}}{n-1}\right\}$
$W_\alpha^2 = \left\{V^2 < \sigma_0^2 \frac{z_{n,\alpha}}{n}\right\}$	$W_\alpha^2 = \left\{S'^2 < \sigma_0^2 \frac{z_{n-1,\alpha}}{n-1}\right\}$
$W_\alpha^3 = \left\{V^2 > \sigma_0^2 \frac{z_{n,1-\frac{\alpha}{2}}}{n} \text{ ou } V^2 < \sigma_0^2 \frac{z_{n,\frac{\alpha}{2}}}{n}\right\}$	$W_\alpha^3 = \left\{S'^2 > \sigma_0^2 \frac{z_{n-1,1-\frac{\alpha}{2}}}{n-1} \text{ ou } S'^2 < \sigma_0^2 \frac{z_{n-1,\frac{\alpha}{2}}}{n-1}\right\}$

5.4 Test sur une proportion p

Modèle : $X_1, ..., X_n$, échantillon aléatoire de la variable X de loi de Bernoulli $\mathcal{B}(1, p) = \mathcal{B}(p)$ avec $\theta = p$ inconnu et $\theta_0 = p_0$ fixé. Pour ce modèle, on supposera n assez grand, c'est-à-dire : $np_0 > 10$ et $n(1 - p_0) > 10$. On notera W_α^j la région de rejet du test de type j, de niveau approximatif α . On a défini $F_n = \bar{X}$.

Les régions de rejet de niveau approximativement α sont données par :

$W_\alpha^1 = \left\{F_n > p_0 + \sqrt{\frac{p_0(1-p_0)}{n}}\phi^{-1}(1-\alpha)\right\} = \left\{F_n > p_0 + \sqrt{\frac{p_0(1-p_0)}{n}}u_{1-\alpha}\right\}$
$W_\alpha^2 = \left\{F_n < p_0 - \sqrt{\frac{p_0(1-p_0)}{n}}\phi^{-1}(1-\alpha)\right\} = \left\{F_n < p_0 - \sqrt{\frac{p_0(1-p_0)}{n}}u_{1-\alpha}\right\}$
$W_\alpha^3 = \left\{F_n < p_0 - \sqrt{\frac{p_0(1-p_0)}{n}}u_{1-\frac{\alpha}{2}} \text{ ou } F_n > p_0 + \sqrt{\frac{p_0(1-p_0)}{n}}u_{1-\frac{\alpha}{2}}\right\} = \left\{\lvert F_n - p_0\rvert > \sqrt{\frac{p_0(1-p_0)}{n}}u_{1-\frac{\alpha}{2}}\right\}$

5.5 p-valeur

Revenons au problème de santé publique portant sur la variable A :

Modèle : $A_1, ..., A_n$ échantillon de la variable A de loi $\mathcal{B}(7, p)$. Pour les données observées $a_1, ..., a_{32}$ on a $\hat{p} = \bar{a}/7 = 0.603$. On sait de plus que, si $7p < 5$, une campagne sera lancée. On traduit alors mathématiquement "il faut lancer une campagne " par $\mathcal{H}_1 : p < 5/7$ et "pas besoin de campagne" par $\mathcal{H}_0 : p = 5/7$. En appliquant le test de type 2 pour une proportion avec des niveaux $\alpha = 0.01\%, ..., 30\%$ on obtient les régions de rejets, caractérisées par leur bord donné par $C_\alpha = 5/7 - u_{2\alpha}\sqrt{(5/7) * (2/7)/224}$ et les décisions suivantes :

α	0.01%	0.1%	1%	10%	20%	30%
C_α	0.602	0.621	0.644	0.676	0.689	0.698
décision	non rejet de $\mathcal{H}_0$	rejet de $\mathcal{H}_0$	rejet de $\mathcal{H}_0$	rejet de $\mathcal{H}_0$	rejet de $\mathcal{H}_0$	rejet de $\mathcal{H}_0$

Il est clair d'après ce tableau que la décision de rejeter $\mathcal{H}_0$ (soit d'accepter $\mathcal{H}_1$) dépend du risque de se tromper que l'on est prêt à prendre. En effet si le risque de rejet à tort de $\mathcal{H}_0$ α est inférieur à 0.01% les données conduisent à ne pas la rejeter (en effet on a alors $\hat{p} = 0.603 \geq C_\alpha$). Par contre si $\alpha \geq 0.1\%$ on rejette $\mathcal{H}_0$ et on accepte $\mathcal{H}_1$ avec un risque de se tromper de $\alpha \geq 0.1\%$. Il y a donc une valeur α^* comprise entre 0.01% et 0.1% au delà de laquelle le jeu de données conduit à rejeter $\mathcal{H}_0$ dans un test de niveau α.

Cette valeur α^* est une fonction de n et du jeu de données (au travers de $\hat{\theta}$). C'est la valeur renvoyée par un logiciel de statistique, lorsqu'on lui fournit la forme de l'alternative (ici $<$), la cible θ_0 (ici $5/7$) et les observations $x_1, ..., x_n$ (ici $a_1, ..., a_{32}$).

Définition : La p-valeur d'un test de type j ($j = 1, 2, 3$) est la valeur la plus grande du risque de première espèce, pour lequel on ne rejette pas $\mathcal{H}_0$. Autrement dit la p-valeur d'un test, est la quantité α^* qui satisfait :

- si $\alpha \leq \alpha^*$ au niveau α on ne rejette pas $\mathcal{H}_0$.
- si $\alpha > \alpha^*$ au niveau α on rejette $\mathcal{H}_0$.

Remarques :

- En pratique, on calcule cette p-valeur comme le α^* pour lequel $C_{\alpha^*} = \hat{\theta}$.

- Lorsqu'une p-valeur est proche de 0, cela signifie que l'on accepte $\mathcal{H}_1$ presque sans risque de se tromper. La proximité à 0 de la p-valeur indique donc un grand degré de fiabilité de $\mathcal{H}_1$.

- Au contraire une p-valeur proche de 1 indique un grand degré de fiabilité de $\mathcal{H}_0$.

Dans l'exemple ci-dessus α^* satisfait $5/7 + u_{2\alpha^*}\sqrt{(5/7) * (2/7)/224} = 0.603$ soit $\alpha^* = 0.0108\%$.

Ainsi les données collectées conduisent à conclure qu'il est nécéssaire de lancer une campagne de sensibilisation dès qu'on accepte un risque de se tromper $> 0.0108\%$. Ou encore on valide $\mathcal{H}_1$ pour tout risque de se tromper $> 0.0108\%$.

Chapter 6

Tests de comparaison d'échantillons

On se demande, à présent, si le poids d'un étudiant de l'amphi dépend de son sexe et si ce poids dépend de son âge. Pour cela, on dispose des poids observés sur 32 étudiants dont $n_1 = 24$ sont des filles et $n_2 = n - n_1 = 8$ sont des garçons, à 15 ans et à 20 ans. On va donc comparer des échantillons pour répondre à ces deux questions.

Remarquons que pour évaluer l'influence du sexe sur le poids à 20 ans, on dispose de l'observation d'une même variable (le poids à 20 ans) sur des individus différents : d'une part $n_1 = 24$ filles et d'autres part $n_2 = 8$ garçons. Les échantillons de données seront dans ce cas dits *indépendants*.

En revanche, pour juger d'une différence éventuelle du poids entre l'âge de 15 ans et l'âge de 20 ans, on dispose de l'observation de deux variables différentes, (c.-à-d. le poids à 15 ans et le poids à 20 ans) sur les mêmes individus. Dans ce cas, on parlera *d'échantillons appariés*.

On se propose dans cette partie de construire des tests de comparaison de moyennes, de proportions ou de variances. Nous allons d'abord traiter le cas le plus simple, celui des échantillons appariés.

6.1 Echantillons appariés

Modèle : Soient Y et Y deux variables aléatoires de moyennes inconnues μ_X et μ_Y. Soit $D = X - Y$. Nous supposerons que D est une variable aléatoire gaussienne de moyenne $\mu_D = \mu_X - \mu_Y$ et de variance inconnue σ_D^2 : $\mathcal{N}(\mu_D, \sigma_D^2)$.

Données : on a observé le couple (X, Y) pour n individus et obtenu : $x_1, ..., x_n$ et $y_1, ..., y_n$. On dispose donc de n observations de D : $d_1 = x_1 - y_1, ..., d_n = x_n - y_n$.

On veut tester $\mathcal{H}_0 : \mu_D = 0$ contre $\mathcal{H}_1 : \mu_D > 0$, ou $\mathcal{H}_0 : \mu_D = 0$ contre $\mathcal{H}_1 : \mu_D \neq 0$.

Pour cela, il suffit d'appliquer les tests paramétriques (unilatéral et bilatéral) d'égalité d'une moyenne à 0, pour un modèle gaussien avec variance inconnue. Les deux régions de rejet W_α^1 (test unilatéral supérieur) et W_α^3 (test bilatéral) sont donc données par

$$W_\alpha^1 = \left\{ \bar{D} > \frac{S'_D}{\sqrt{n}} t_{n-1,1-\alpha} \right\} = \{T > t_{n-1,1-\alpha}\} \ ; \ W_\alpha^3 = \left\{ |\bar{D}| > \frac{S'_D}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \right\} = \left\{ |T| > t_{n-1,1-\frac{\alpha}{2}} \right\},$$

où $T = \bar{D}/(S'_D/\sqrt{n})$.

Décisions :

- si $t_{calc} = \bar{d}/(s'_D/\sqrt{n}) > t_{n-1,1-\alpha}$, alors on accepte $\mathcal{H}_1 : \mu_D > 0$ avec un risque de se tromper de α.

- si $|t_{calc}| = |\bar{d}|/(s'_D/\sqrt{n}) > t_{n-1,1-\alpha/2}$, alors on accepte $\mathcal{H}_1 : \mu_D \neq 0$ avec un risque de se tromper de α.

Remarque : Pour faire le test unilatéral inférieur c'est à dire pour $\mathcal{H}_1 : \mu_D < 0$ on utilisera la région de rejet $W_\alpha^2 = \{T < -t_{n-1,1-\alpha}\}$.

Exemple : X le poids à 15 ans et Y le poids à 20 ans. On veut tester $\mathcal{H}_0 : \mu_D = 0$ contre $\mathcal{H}_1 : \mu_D < 0$. L'hypothèse alternative traduit qu'en moyenne on est plus lourd à 20 ans qu'à 15 ans. On obtient $t_{calc} = \bar{x} - \bar{y}/(s'_D/\sqrt{n}) = (52,62 - 58,47)\sqrt{32}/3,62 = -9,13$. Attention s'_D est l'écart type estimé sur la série des observations $d_i = x_i - y_i$. Et la p-valeur de ce test est donnée par $P(T < t_{calc}) = \texttt{pt}(t_{calc}, 31) = 1,3 \cdot 10^{-10}$. On concluera donc avec un risque de se tromper $\alpha > 1,3 \cdot 10^{-10}$ qu'en moyenne le poids à 20 ans est supérieur au poids à 15 ans.

6.2 Echantillons indépendants

6.2.1 Comparaisons de moyennes

Modèle : Soient X_1 et X_2 deux variables aléatoires indépendantes et de loi respectives $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$.

Données : on a observé un échantillon pour X_1 de taille n_1 et un échantillon pour X_2 de taille n_2. On notera $\bar{x}_1$, $\bar{x}_2$, s_1^2 et s_2^2 les moyennes et variances empiriques de ces deux échantillons de données.

En général, X_1 et X_2 décrivent un même caractère sur deux populations différentes $\mathcal{P}_1$ et $\mathcal{P}_2$.

Par exemple, sur les données de poids on observe d'une part les poids d'un échantillon prélevé dans la population des filles et d'autre part les poids d'un second échantillon prélevé dans la population des garçons. Ainsi μ_1 (resp. μ_2) est le poids moyen dans la population $\mathcal{P}_1$ (resp. $\mathcal{P}_2$) et σ_1^2 (resp. σ_2^2) la variance du poids dans la population $\mathcal{P}_1$ (resp. $\mathcal{P}_2$). Dans ce cas, si on veut montrer que le poids moyen d'une personne dépend de son sexe, on choisira $\mathcal{H}_1 : \mu_1 - \mu_2 \neq 0$ et si on veut montrer que le poids d'une femme est en moyenne moins important que celui d'un homme on prendra : $\mathcal{H}_1 : \mu_1 - \mu_2 < 0$.

On veut donc tester $\mathcal{H}_0 : \mu_1 - \mu_2 = 0$ contre $\mathcal{H}_1 : \mu_1 - \mu_2 > 0$ (ou $\mu_1 - \mu_2 < 0$), ou $\mathcal{H}_0 : \mu_1 - \mu_2 = 0$ contre $\mathcal{H}_1 : \mu_1 - \mu_2 \neq 0$.

Selon les tailles des échantillons et les informations dont on dispose sur les paramètres σ_1^2 et σ_2^2, on utilisera des tests différents.

1. ***Echantillons de petites tailles :*** $n_1 < 100$ ***et*** $n_2 < 100$ ***:***

   Dans ce cas, on ne sait proposer un test que dans les situations où σ_1^2 et σ_2^2 sont connues ou dans celle où elles sont inconnues mais *supposées égales*.

   1-a) Les variances de la population σ_1^2 et σ_2^2 sont connues :

   Comme sous l'hypothèse $\mathcal{H}_0$, la variable $U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ suit une loi $\mathcal{N}(0,1)$, les régions de rejet des tests de type 1 (unilatéral supérieur), 2 (unilatéral inférieur) et 3 (bilatéral) sont données par :

   $$W_\alpha^1 = \{U > u_{1-\alpha}\}; \quad W_\alpha^2 = \{U < -u_{1-\alpha}\} \quad ; \quad W_\alpha^3 = \{|U| > u_{1-\frac{\alpha}{2}}\}.$$

Décisions :

- si $u_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > u_{1-\alpha}$, alors on accepte $\mathcal{H}_1 : \mu_1 - \mu_2 > 0$ avec un risque de se tromper de α ;
- si $|u_{calc}| > u_{1-\alpha/2}$, alors on accepte $\mathcal{H}_1 : \mu_1 - \mu_2 \neq 0$ avec un risque de se tromper de α.

S'il arrive parfois que σ_1^2 et σ_2^2 soient connues, cela est cependant peu fréquent, et dans le cas où σ_1^2 et σ_2^2 sont inconnues, on ne sait traiter le problème que si elles sont **égales** :

1-b) Les variances de la population σ_1^2 et σ_2^2 sont inconnues mais égales à σ^2 :

La variable U précédente n'est plus utilisable pour effectuer le test puisqu'elle dépend des inconnues $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Elle sera donc remplacée par la variable T qui est l'analogue de U avec σ_1^2 et σ_2^2 remplacées par l'estimateur de σ^2 noté Σ et défini par:

$$\Sigma^2 = \frac{(n_1-1)S_1'^2 + (n_2-1)S_2'^2}{n_1+n_2-2} = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1+n_2-2} \quad \text{et} \quad T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\Sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}.$$

Sous l'hypothèse $\mathcal{H}_0$ et si de plus $\sigma_1^2 = \sigma_2^2 = \sigma^2$, la variable $(n_1+n_2-2)\Sigma^2/\sigma^2$ suit une loi du Chi2 à n_1+n_2-2 degrés de liberté : $\mathcal{X}^2_{n_1+n_2-2}$. On en déduit alors que T suit une loi de Student à n_1+n_2-2 degrés de liberté. Les régions de rejet des tests de type 1 (unilatéral supérieur), 2 (unilatéral inférieur) et 3 (bilatéral) sont données par :

$$W_\alpha^1 = \{T > t_{n_1+n_2-2,1-\alpha}\} \ ; \ W_\alpha^2 = \{T < -t_{n_1+n_2-2,1-\alpha}\} \quad \text{et} \quad W_\alpha^3 = \{|T| > t_{n_1+n_2-2,1-\frac{\alpha}{2}}\}.$$

Remarque : L'estimation $\hat{\sigma^2}$ de σ^2 est le résultat de l'application de Σ^2 aux données, c'est à dire $\hat{\sigma^2} = \Sigma^2_{calc}$.

Décisions :

- si $t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\hat{\sigma^2}}{n_1} + \frac{\hat{\sigma^2}}{n_2}}} > t_{n_1+n_2-2,1-\alpha}$, alors on accepte $\mathcal{H}_1 : \mu_1 - \mu_2 > 0$ avec un risque de se tromper de α.
- si $|t_{calc}| > t_{n_1+n_2-2,1-\alpha/2}$, alors on accepte $\mathcal{H}_1 : \mu_1 - \mu_2 \neq 0$ avec un risque de se tromper de α.

Exemple : Peut-on déclarer que les filles sont en moyenne de plus petite taille que les garçons?

Notons X la taille d'une fille et Y celle d'un garçon et supposons X et Y indépendantes et de lois respectives $\mathcal{N}(\mu_1, \sigma^2)$ et $\mathcal{N}(\mu_2, \sigma^2)$. On fait un test unilatéral inférieur sur $\mu_1 - \mu_2$. Calculons la statistique Σ^2 pour estimer σ^2 puis la statistique de test T : $\Sigma^2_{calc} = \frac{23 \cdot 39{,}4185 + 7 \cdot 21{,}5536}{30} = 35{,}25$ et $T_{calc} = (165{,}13 - 179{,}88)/\sqrt{35{,}25(1/32+1/8)} = -6{,}28$. D'où la p-valeur du test $\alpha^* = \texttt{pt}(T_{calc}, 30) = 3{,}15 \cdot 10^{-7}$. Ces données permettent donc de conclure avec un risque de se tromper $\alpha > 3{,}15 \cdot 10^{-7}$ qu'en moyenne les garçons sont plus grands que les filles.

2. ***Echantillons de grandes tailles :*** $n_1 \geq 100$ ***et*** $n_2 \geq 100$

Dans ce cas, que les variances σ_1^2 et σ_2^2 soient égales ou non n'importe pas puisque l'on dispose d'assez de données pour les estimer correctement. Les régions de rejet de $\mathcal{H}_0 : \mu_1 = \mu_2$ contre $\mathcal{H}_1 : \mu_1 > \mu_2$ ou contre $\mathcal{H}_1 : \mu_1 \neq \mu_2$ sont données par :

$$W_\alpha^1 = \{U > u_{1-\alpha}\} \quad \text{et} \quad W_\alpha^3 = \{|U| > u_{1-\frac{\alpha}{2}}\} \quad \text{avec} \quad U = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{{S'_1}^2}{n_1} + \frac{{S'_2}^2}{n_2}}} \approx \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}};$$

car, pour de grands effectifs, la variable U suit approximativement une loi normale centrée réduite. Les tests proposés sont de ce fait approximativement de niveau α. On peut remarquer que les échantillons étant de grandes tailles, dans U, $S_1'^2$ et $S_2'^2$ peuvent être remplacés par S_1^2 et S_2^2.

6.2.2 Comparaison de proportions

Les tests précédents de comparaisons de moyennes peuvent être appliqués à de **grands** échantillons de variables de Bernoulli. Autrement dit, il n'est pas nécessaire de supposer les variables X_1 et X_2 normales pour comparer leurs moyennes, à condition que les échantillons prélevés soient de tailles suffisantes (dans ce cas, les paramètres inconnus testés sont $\mu_1 = p_1 = P(X_1 = 1)$ et $\mu_2 = p_2 = P(X_2 = 1)$). Par contre, il faut encore supposer les variables X_1 et X_2 indépendantes. La statistique de test U peut s'exprimer en fonction de $\bar{X}_1$ et $\bar{X}_2$ car dans le modèle de Bernoulli on peut montrer que $S^2 = \bar{X}(1 - \bar{X})$ puisque $X \in \{0, 1\}$. Ainsi on a :

$$U \approx \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\bar{X}_1(1-\bar{X}_1)}{n_1} + \frac{\bar{X}_2(1-\bar{X}_2)}{n_2}}}$$

6.2.3 Comparaisons de variances

Certaines questions posées seront traduites par un test de comparaison de variances d'un même caractère sur deux populations différentes. On peut également souhaiter évaluer s'il est raisonnable, dans le cas de petits échantillons, de supposer les deux variances σ_1^2 et σ_2^2 égales.

Modèle : Soient X_1 et X_2 deux variables aléatoires indépendantes et de loi respectives $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$ avec μ_1, μ_2, σ_1^2 et σ_2^2 inconnues.

Données : on a observé un échantillon pour X_1 de taille n_1 et un échantillon pour X_2 de taille n_2.

On souhaite effectuer les tests unilatéraux (types 1 et 2) :

$$\mathcal{H}_0 : \sigma_1^2/\sigma_2^2 = 1 \quad \text{contre} \quad \mathcal{H}_1 : \sigma_1^2/\sigma_2^2 > 1$$

$$\mathcal{H}_0 : \sigma_1^2/\sigma_2^2 = 1 \quad \text{contre} \quad \mathcal{H}_1 : \sigma_1^2/\sigma_2^2 < 1$$

ou le test bilatéral (de type 3) :

$$\mathcal{H}_0 : \sigma_1^2/\sigma_2^2 = 1 \quad \text{contre} \quad \mathcal{H}_1 : \sigma_1^2/\sigma_2^2 \neq 1.$$

Pour construire ces tests, on utilisera la variable aléatoire F, obtenue comme le rapport des estimateurs de σ_1^2 et σ_2^2 :

$$F = \frac{\frac{n_1 S_1^2}{n_1 - 1}}{\frac{n_2 S_2^2}{n_2 - 1}} = \frac{S_1'^2}{S_2'^2}.$$

Rappelons que $\frac{n_1 S_1^2}{\sigma_1^2}$ et $\frac{n_2 S_1^2}{\sigma_2^2}$ suivent respectivement des lois $\mathcal{X}^2_{n_1-1}$ et $\mathcal{X}^2_{n_2-1}$. Ainsi, sous l'hypothèse $\mathcal{H}_0$, la variable F est égale au rapport de deux Chi2 divisés par leur degrés de libertés respectifs. Sous $\mathcal{H}_0$, la loi de F est donc celle d'une variable de Fisher-Snedecor à (n_1-1, n_2-1) degrés de liberté $\mathcal{F}_{n_1-1,n_2-1}$. Remarquons également que $1/F$ suit aussi une loi de Fisher-Snedecor de degrés (n_2-1, n_1-1). On notera $f_{\nu_1,\nu_2,p}$ le quantile d'ordre p d'une variable de loi $\mathcal{F}_{\nu_1,\nu_2,p}$. Par définition de la variable de Fisher-Snedecor on montre que $f_{\nu_1,\nu_2,p} = 1/f_{\nu_2,\nu_1,1-p}$.

Les régions de rejet suivantes permettent d'effectuer les trois tests au niveau α :

$$W^1_\alpha = \{F > f_{n_1-1,n_2-1,1-\alpha}\} \ ; W^2_\alpha = \{F < 1/f_{n_2-1,n_1-1,1-\alpha}\} = \{F < f_{n_1-1,n_2-1,\alpha}\} \text{ et}$$

$$W^3_\alpha = \{F > f_{n_1-1,n_2-1,1-\alpha/2} \text{ ou } F < f_{n_1-1,n_2-1,\alpha/2}\}.$$

Décisions :

- si $f_{calc} = s_1'^2/s_2'^2 > f_{n_1-1,n_2-1,1-\alpha}$, alors on accepte $\mathcal{H}_1 : \sigma_1^2 > \sigma_2^2$ avec un risque de se tromper de α.
- si $f_{calc} = s_1'^2/s_2'^2 < f_{n_1-1,n_2-1,\alpha}$, alors on accepte $\mathcal{H}_1 : \sigma_1^2 < \sigma_2^2$ avec un risque de se tromper de α.
- si $f_{calc} = s_1'^2/s_2'^2 > f_{n_1-1,n_2-1,1-\alpha/2}$ ou si $f_{calc} = s_1'^2/s_2'^2 < f_{n_1-1,n_2-1,\alpha/2}$ alors on accepte $\mathcal{H}_1 : \sigma_1^2 \neq \sigma_2^2$ avec un risque de se tromper de α.
- si $f_{n_1-1,n_2-1,\alpha/2} \leq s_1'^2/s_2'^2 \leq f_{n_1-1,n_2-1,1-\alpha/2}$, alors on accepte $\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2$ avec un risque de se tromper de $1-\alpha$.

Remarque : Lorsque l' on souhaite vérifier l'égalité des variances afin de le poser en hypothèse pour effectuer une comparaison de moyenne à l'aide de petits échantillons, on se donne un niveau α grand. En effet, rappelons que lorsque la conclusion d'un test est d'accepter $\mathcal{H}_0$, elle est donnée avec un risque d'erreur de $1-\alpha$. En pratique si aucune valeur de α n'est indiquée on calcule la p-valeur du test et on ne rejette pas l'égalité si cette p-valeur est supérieure à 10%. En fait, on jugera d'autant plus raisonnable d'accepter l'égalité des variances que la $p-$valeur du test sera proche de 1.

Exemple : comparaison de la variance de la taille d'une fille et de celle d'un garçon. Dans le test de comparaison des tailles moyennes chez les filles ou chez les garçons on a supposé que les variances σ_1^2 et σ_2^2 étaient égales. Effectuons un test d'égalité des variances pour s'assurer qu'il s'agissait d'une hypothèse raisonnable. On fait ici le test bilatéral, car ce qui importe est de savoir si les variances sont ou pas différentes et pas si l'une est plus grande que l'autre. La statistique de test est $F_{calc} = s_1'^2/s_2'^2 = 39,4/21,6 = 1,8$. Si on fait le test au niveau $\alpha = 5\%$ on utilise les quantiles $f_{31,7,0.975}$ et $f_{31,7,0.025}$. Le premier de ces deux quantiles se lit dans la table page 6 colonne 30 (il n'y a pas de colonne 31) et ligne 7 : $f_{31,7,0.975} \approx f_{30,7,0.975} = 4,36$ et le second quantile se calcule comme $f_{31,7,0.025} = 1/f_{7,31,0.975} \approx 1/f_{7,30,0.975} = 1/2,75 = 0,36$ (lecture page 5). Comme $0,36 < F_{calc} < 4,36$ on ne rejette pas l'égalité des variance dans un test de niveau 5%. Si on veut faire le test au niveau $\alpha = 20\%$ on ne dispose pas des quantiles d'ordre 0.1 ou 0.9 sur les tables mais on peut calculer dans ce cas les quantiles avec R ou avec la fonction quantile d'une calculatrice. Avec R on a `qf(0.9,31,7)=2,6` et `qf(0.1,31,7)=0,5`. Comme $0,5 < F_{calc} < 2,6$, au niveau $\alpha = 20\%$ on ne refuse pas l'égalité des variances.

Si on veut à présent calculer la p-valeur de ce test, on l'obtient ici lorsque la statistique de test coïncide avec la borne supérieure de la région d'acceptation de $\mathcal{H}_0$ car $F_{calc} > 1$ soit $\alpha^* = 2 * P(F > F_{calc}) = 2(1-$`pf(1.8,31,7)`$) = 42\%$. On peut aussi obtenir cette p-valeur en utilisant la fonction `var.test()` de R avec `var.test(x,y)` où `x` (resp. `y`) contiennent les tailles des filles (resp. des garçons).

Chapter 7

Tests du Chi2

A l'aide des tests proposés dans le chapitre précédent, nous avons pu répondre à la question : "le poids moyen d'une fille est-il inférieur à celui d'un garçon ?". Nous avons conclu, à l'aide des données observées, qu'avec un risque d'erreur de α tel que $\alpha > \alpha^*$ la réponse était oui. Cela suffit pour dire que la variable poids d'une fille modélisée par X_1 n'a pas la même distribution que la variable poids d'un garçon modélisée par X_2. Par conséquent on conclut que le poids est un caractère qui dépend du sexe pour tout risque $> \alpha^*$. Par contre, si la réponse avait été "le poids moyen d'une fille est le même que celui d'un garçon", nous n'aurions pu directement conclure à l'indépendance entre le poids et le sexe. En effet, il faudrait pour cela comparer la distribution observée du poids d'une fille à celle du poids d'un garçon. Si ces deux distributions observées, sont proches l'une de l'autre alors on pourra conclure à l'indépendance entre le poids et le sexe. Un des tests permettant de répondre, statistiquement, à cette question s'appelle *test d'indépendance du Chi2*.

D'autre part, dans tous les tests rencontrés jusqu'ici, une hypothèse sur la loi de probabilité de la variable modélisant le caractère étudié est posée. En effet, en général on a supposé les variables de lois normales. De même que l'hypothèse d'égalité des variances peut être statistiquement vérifiée, celle posée sur la distribution théorique de la variable étudiée peut aussi être validée à l'aide d'un test. Nous étudierons ici un test construit à l'aide d'une variable du Chi2, appelé *test d'adéquation du Chi2*. Nous commencerons par l'étude du test d'adéquation, qui a déjà été rencontré lorsque nous avons calculé le d^2 sur les données "alimentation" : $a_1, ..., a_{32}$.

7.1 Test d'adéquation du Chi2

Nous allons d'abord développer ce test en détails dans le cas de données discrètes ou quantitatives. L'extension à une variable aléatoire continue en découlera simplement moyennant quelques modifications.

7.1.1 Variable discrète ou qualitative

Modèle : $X_1, ..., X_n$, échantillon aléatoire d'une variable X de fonction de répartition inconnue F. On suppose que l'on connait l'ensemble des valeurs possibles de X, $\mathcal{X} = \{m_1, ..., m_q\}$. Notons N_k la variable aléatoire définie comme le nombre de variables X_i de l'échantillon aléatoire qui prennent la valeur m_k. On construit ainsi une suite de variables aléatoires $N_1, ..., N_q$.
Données : $x_1, ..., x_n$ qui représentent les réalisations de $X_1, ..., X_n$ sont souvent remplacées par les réalisations de $N_1, ..., N_q$ notées $n_1, ..., n_q$. En effet lorsque l'on observe un échantillon d'assez grande taille $n \geq 20$, le tableau de données en effectifs est plus court à décrire que la suite des n observations. Rappelons que l'on a évidemment $\sum_{k=1}^{q} n_k = n$.

Nous voulons donc répondre à la question "la variable X suit-elle une distribution donnée F^* ?". Cette distribution F^* est décrite par des probabilités fixées $p_1^*, ..., p_q^*$. D'autre part nous noterons $p_k = P(X = m_k)$.

Nous allons construire un test de

$$\mathcal{H}_0 : X \text{ suit la loi } F^* \qquad \text{contre} \qquad \mathcal{H}_1 : X \text{ ne suit pas la loi } F^*;$$

qui s'écrit aussi

$$\mathcal{H}_0 : \quad \text{pour tout } k, p_k = p_k^* \qquad \text{contre} \qquad \mathcal{H}_1 : \text{il y a une valeur de } \ k \ \text{ telle que } \ p_k \neq p_k^*.$$

L'hypothèse $\mathcal{H}_0$ traduit l'adéquation de la variable étudiée à la loi F^*.

En posant $\delta^2 = \sum_{k=1}^{q}(np_k - np_k^*)^2/(np_k^*)$, l'hypothèse $\mathcal{H}_0$ s'écrit $\delta^2 = 0$. Ainsi notre problème peut être ramené au test suivant :

$$\mathcal{H}_0 : \delta^2 = 0 \qquad \text{contre} \qquad \mathcal{H}_1 : \delta^2 > 0.$$

On va donc construire un estimateur de δ^2, D^2 puis à l'aide de la loi de D^2 on ajuste la région de rejet $\{D^2 > C_\alpha\}$ pour que la probabilité de rejet de $\mathcal{H}_0$ alors qu'elle est vraie soit α.

Les p_k étant inconnues on les estime à l'aide de l'échantillon $X_1, ..., X_n$ qui fournit la suite de variables $N_1, ..., N_q$. En effet, on montre que N_k/n est un estimateur sans biais et convergent de p_k. Ainsi $(N_k - np_k^*)^2/(np_k^*)$ est un estimateur de $(np_k - np_k^*)^2/(np_k^*)$. On utilisera donc la variable aléatoire suivante pour construire le test :

$$D^2 = \sum_{k=1}^{q} \frac{(N_k - np_k^*)^2}{np_k^*}.$$

Sous l'hypothèse $\mathcal{H}_0$, et si de plus $np_k^* \geq 5$, alors D^2 suit une loi du Chi2 à $q-1$ degrés de liberté $\mathcal{X}^2_{q-1}$. La région de rejet du test de niveau approximatif α est donc donnée par :

$$W_\alpha = \{D^2 > z_{q-1,1-\alpha}\}.$$

L'application de D^2 aux données fournit une estimation du paramètre inconnu $\hat{\delta^2}$ et elle est notée d^2 (c.-à d. $\hat{\delta^2} = d^2$).

Décision :

- si $d^2 > z_{q-1,1-\alpha}$ on refuse l'adéquation à la distribution F^* avec un risque d'erreur de α;
- sinon on accepte l'adéquation avec un risque d'erreur de $1 - \alpha$.

Le test d'adéquation donne un exemple de situation où c'est l'hypothèse $\mathcal{H}_0$ que l'on souhaite accepter. Par conséquent si l'on souhaite valider l'adéquation avec peu de risque de se tromper, on choisira α proche de 1. Si aucun risque d'erreur n'est indiqué pour faire le test, on jugera l'adéquation satisfaisante si la p-valeur, α^* de ce test est au moins supérieure à 10% (ce qui est peu satisfaisant mais permet juste de dire que si on rejette le modèle proposé ce sera avec un risque de se tromper de 10%). Bien sure plus la p-valeur du test sera grande, plus la conclusion $\mathcal{H}_0$ (qui affirme l'adéquation) sera fiable. La p-valeur satisfait $d^2 = z_{q-1,1-\alpha^*}$ et se calcule comme $P(D^2 > d^2|\mathcal{H}_0)$=$P(Chi2 > d^2)$ (remarque : ici D^2 calculé sur les données est noté d^2, on aurait pu également le noter D^2_{calc}).

Exemple de l'alimentation :

On souhaite répondre à la question : "le nombre de jours par semaine où le nombre de fruits et légumes consommés est d'au moins quatre, suit-il une répartition uniforme ?". Nous avons déjà calculé le d^2 dans le chapitre 1 et obtenu $\hat{\delta^2} = d^2 = 23$, d'où $\alpha^* = 0.17\%$. Par exemple, pour $\alpha = 5\%$ on refuse $\mathcal{H}_0$ (c.à d. l'adéquation) et on conclut avec un risque d'erreur de 5% : il n'y a pas adéquation des données à la distribution uniforme. Nous avions également comparé la distribution observée avec une loi binomiale $\mathcal{B}(7, 0.5)$ et obtenu dans ce cas $d^2 = 11.276$. Si c'est un peu mieux que l'adéquation précédente, on obtient cependant une p-valeur assez loin de 1 puisqu'elle est de $\alpha^* = 12.7\%$. On se

propose maintenant de tester l'adéquation à une loi binomiale $\mathcal{B}(7,p)$ mais où l'on ne fixe pas, p à priori. Dans ce cas on estime p, à partir des données et on obtient $\hat{p} = 0.603$ (voir chap. 4). On obtient le tableau suivant :

$\mathcal{X}$	0	1	2	3	4	5	6	7	
eff. obs.	0	1	2	6	9	9	4	1	d^2
p. th.$\mathcal{B}(7,\hat{p})$	0.16%	1.66%	7.55%	19.09%	28.96%	26.36%	13.33%	2.89%	
eff. th.$\mathcal{B}(7,\hat{p})$	0.05	0.53	2.42	6.11	9.27	8.43	4.26	0.92	0.606

Dans ce cas la variable D^2 qui fournit l'estimation $\hat{\delta^2} = d^2$ suit une loi du Chi2 à 6 degrés de libertés (au lieu de 7 pour les deux lois précédemment testées) car il a fallu estimer un paramètre pour calculer les effectifs théoriques. Ainsi on obtient dans ce cas une p-valeur α^* qui satisfait $0.606 = z_{6,\alpha^*}$ soit $\alpha^* = 99.63\%$. L'adéquation à la $\mathcal{B}(7, 0.6027)$ est bien meilleure que celle aux deux autres distributions envisagées. Elle est même très satisfaisante puisqu'à moins de prendre un risque, de rejeter à tort l'adéquation, supérieur à 99.63% les données ne permettent pas de rejeter l'hypothèse indiquant la distribution $\mathcal{B}(7, 0.6027)$.

Remarquons qu'ici, l'échantillon est de trop petite taille pour que les conditions $np_k^* \geq 5$ soient vérifiées ce qui nous interdit en principe d'appliquer le test d'adéquation en envisageant de faire autant de catégories que de modalités. Le test est ainsi effectué pour conserver les mêmes classes que celles étudiées comme exemple à la fin du premier chapitre, lors du calcul de d^2.

Afin de ne rencontrer aucune catégorie ayant un effectif théorique inférieur à cinq nous devrions réunir les quatres premières classes en une seule et les deux dernières en une seule. Ainsi la classe réunissant les modalités $0, 1, 2$ et 3 aurait un effectif théorique de 9.11 à comparer avec l'effectif observé 9 et la dernière classe englobant les modalités 6 et 7 aurait un effectif théorique de 5.19 et un effectif observé de 5. Une répartition des modalités en quatre classes conduit à $d^2 = 0.054$ et donne une p-valeur de $\alpha^* = 97.35\%$

Souvent, on cherche à valider l'adéquation à une distribution partiellement connue. Par exemple, "les observations sont-elles issues d'une loi gaussienne ou binomiale ?" sans en préciser les paramètres.

Cas de F^* partiellement donnée :

Lorsque la loi F^* est connue à r paramètres près (dans l'exemple précédent un paramètre est estimé) ces paramètres sont estimés et si $q \geq r+2$, la variable D^2 suit alors une loi du Chi2 à $q-1-r$ degrés de libertés et la région de rejet du test d'ádéquation de niveau α est donnée par :

$$W_\alpha = \{D^2 > z_{q-1-r,1-\alpha}\}.$$

7.1.2 Variable continue

Lorsque la variable observée est à valeurs dans un intervalle de $\mathbb{R}$. On partitionne cet intervalle en q classes, notées $[e_0, e_1], ...,]e_{q-1}, e_q]$. On procède ensuite comme précédemment en remplaçant les modalités m_k par les classes $]e_{k-1}, e_k]$. Les probabilités p_k^* qui "caractérisent" la distribution F^* à laquelle on souhaite vérifier l'adéquation sont définies comme :

$$p_k^* = F^*(e_k) - F^*(e_{k-1}) = P(e_{k-1} \leq X \leq e_k | \mathcal{H}_0 \text{vraie}).$$

L' effectif observé n_k est le nombre d'éléments de l'échantillon observé, $x_1, ..., x_n$, qui sont tombés dans la classe k.

Exemple du "poids d'une fille" :

Dans le chapitre 6 pour comparer le poids moyen d'une fille (à 20 ans) à celui d'un garçon (à 20 ans), nous supposons que la variable poids d'une fille X_1 a une distribution normale (idem pour le poids d'un garçon). Sur l'échantillon de taille $n_1 = 24$ nous avons observé les poids suivants : $49, 53, 50, ..., 55, 68$. Ces valeurs sont toutes dans l'intervalle $[40, 70]$ que l'on peut, par exemple, découper en quatre classes : $[40, 49],]49, 54],]54, 60]$ et $]60, 70]$. Nous voulons tester l'adéquation de la distribution observée à une

loi normale de moyenne μ et de variance σ^2 inconnues. On a déja vu que ce jeu de données fournit les estimations $\hat{\mu} = 54.19$ et $\hat{\sigma^2} = 54.4229$ et on a

$$p_k^* = \Phi_{\hat{\mu},\hat{\sigma}}(e_k) - \Phi_{\hat{\mu},\hat{\sigma}}(e_{k-1}) = \Phi\left(\frac{e_k - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{e_{k-1} - \hat{\mu}}{\hat{\sigma}}\right).$$

On obtient le tableau suivant :

$[40, 70]$	[40,49]	]49,54]	]54,59]	]59,70]	
eff. obs.	7	8	4	5	d^2
pr. th. $\mathcal{N}(\hat{\mu}, \hat{\sigma^2})$	21.44%	24.91%	25.28%	24.03%	
eff. th. $\mathcal{N}(\hat{\mu}, \hat{\sigma^2})$	6.86	7.97	8.1	7.69	3.011

La p-valeur, α^* qui satisfait $z_{1,1-\alpha^*} = 3.011$ vaut 8.27%. Le résultat du test d'adéquation est ici très peu satisfaisant puisque l'on accepte l'adéquation à la $\mathcal{N}(\hat{\mu}, \hat{\sigma^2})$ qu'à condition de limiter le risque de la rejeter à tort à 8.27%.

7.2 Test d'indépendance du Chi2

Comme dans la partie précédente, nous développons ce test en détails dans le cas de données discrètes ou qualitatives. L'extension à des variables aléatoires continues en découlera simplement moyennant les mêmes modifications que dans le test d'adéquation.

7.2.1 Variables discrètes ou qualitatives

Modèle : $(X_1, Y_1), ..., (X_n, Y_n)$, échantillon aléatoire d'un couple de variables aléatoires X et Y à valeurs dans $\mathcal{X} = \{m_1, ..., m_q\}$ et $\mathcal{Y} = \{\tilde{m}_1, ..., \tilde{m}_p\}$.

Notons N_{ij} la variable aléatoire définie comme le nombre de couples (X_k, Y_k) de l'échantillon aléatoire $(X_1, Y_1)....,(X_n, Y_n)$ qui prennent les valeur $(m_i, \tilde{m}_j)$. On construit aussi les variables aléatoires $N_{1.}, ..., N_{q.}$ et $N_{.1}, ..., N_{.p}$ définies comme :

$$N_{.k} = \sum_{j=1}^{p} N_{kj} \qquad \text{et} \qquad N_{k.} = \sum_{i=1}^{q} N_{ik}.$$

$N_{.k}$ indique le nombre de couples dans l'échantillon aléatoire pour lesquels X prend la valeur m_k et $N_{k.}$ le nombre de couples dans l'échantillon aléatoire pour lesquels Y prend la valeur $\tilde{m}_k$.

Données : $(x_1, y_1), ..., (x_n, y_n)$ qui représentent les réalisations de $(X_1, Y_1), ..., (X_n, Y_n)$ sont en général remplacées par le tableau à double entrées des effectifs n_{ij} réalisations des N_{ij}. Ce tableau est appelé tableau de contingence et a la forme suivante :

Y / X	$\tilde{m}_1$	...	$\tilde{m}_j$	...	$\tilde{m}_p$	Total
m_1	n_{11}	...	n_{1j}	...	n_{1p}	$n_{1.}$
$\vdots$						
m_i	n_{i1}	...	n_{ij}	...	n_{ip}	$n_{i.}$
$\vdots$						
m_q	n_{q1}	...	n_{qj}	...	n_{qp}	$n_{q.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.p}$	$n_{..} = n$

L'objectif est ici de répondre à la question : "X et Y sont-elles indépendantes ? ".
L'indépendance entre X et Y est mathématiquement définie par :

$$P(X = m_i, Y = \tilde{m}_j) = P(X = m_i)P(X = \tilde{m}_j), \quad \text{pour tout} \quad (i, j).$$

De façon équivalente l'indépendance entre X et Y s'écrit aussi $\delta^2 = 0$ avec δ^2 défini comme :

$$\delta^2 = n \cdot \sum_{i,j} \frac{(P(X = m_i, Y = \tilde{m}_j) - P(X = m_i)P(X = \tilde{m}_j))^2}{P(X = m_i)P(X = \tilde{m}_j)}.$$

En plaçant l'hypothèse d'indépendance sous $\mathcal{H}_0$ on est amené à poser le test suivant :

$$\mathcal{H}_0 : \delta^2 = 0 \qquad \text{contre} \qquad \mathcal{H}_1 : \delta^2 > 0.$$

Comme N_{ij}/n (resp. $N_{i.}/n$, $N_{.j}/n$) est un bon estimateur de $P(X = m_i, Y = \tilde{m}_j)$ (resp. $P(X = m_i)$, $P(Y = \tilde{m}_j)$), on utilisera l'estimateur de δ^2 suivant :

$$D^2 = \sum_{i,j} \frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{\frac{N_{i.}N_{.j}}{n}}.$$

Sous l'hypothèse $\mathcal{H}_0$ et si $p \geq 2$ et $q \geq 2$, D^2 suit une loi du Chi2 à $(p-1)(q-1)$ degrés de liberté $\mathcal{X}^2_{(p-1)(q-1)}$. La région de rejet du test de niveau approximatif α est donc donnée par :

$$W_\alpha = \{D^2 > z_{(p-1)(q-1),1-\alpha}\}.$$

L'application de D^2 aux données, qui fournit l'estimation $\hat{\delta^2}$ est notée d^2.
Décision :

- si $d^2 > z_{(p-1)(q-1),1-\alpha}$ on refuse l'indépendance entre X et Y avec un risque d'erreur de α;
- sinon on accepte l'indépendance avec un risque d'erreur de $1 - \alpha$.

La p-valeur de ce test est donnée par α^* tel que $z_{(p-1)(q-1),1-\alpha^*} = d^2$.

7.2.2 Variable(s) continue(s)

Lorsque l'une ou l'autre des deux variables est continue on procède comme ci-dessus en remplaçant modalité m_k par classe $]e_{k-1}, e_k]$.

Exemple poids/sexe :
Pour répondre à la question : "le poids (à 20 ans) d'un individu de l'amphi dépend-il de son sexe ?", on va appliquer le test précédent au couple (sexe, poids). Le sexe, X prend les modalités 1 ou 0. Le poids Y est à valeurs dans $[40, 80]$ décomposé en deux classes : $[40, 60]$ et $]60, 80]$. On a observé les effectifs (premières lignes des cases du tableau) et calculé les $n_{i.}n_{.j}/n$ (seconde ligne du tableau) :

Y / X	$[40, 60]$	$]60, 80]$	Total
1	20 480/32	4 288/32	24
0	0 160/32	8 96/32	8
	20	8	32

On obtient $\hat{\delta^2} = d^2 = 17.78$ et $\alpha^* = 2.48\%$. On refuse donc $\mathcal{H}_0$ pour tout risque d'erreur supérieur à 2.48%. Autrement dit on conclut que le poids dépend bien du sexe dès que l'on accepte un risque d'erreur supérieur à 2.48%.