PROBABILITÉS ET STATISTIQUE INFÉRENTIELLE

DUT TC 2 - Module OS 01

Université du Littoral - Côte d'Opale, La Citadelle

Laurent SMOCH

(smoch@lmpa.univ-littoral.fr)

Septembre 2016

Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville
Université du Littoral, zone universitaire de la Mi-Voix, bâtiment H. Poincaré
50, rue F. Buisson, BP 699, F-62228 Calais cedex

UE31	Elargir ses compétences en gestion	Vol. horaire global : 27h **CM 12h TD 15h**
OS 01	**Probabilités et Statistique inférentielle**	Semestre 3

Objectifs du module

Savoir faire des calculs de probabilité, d'intervalle de confiance et de test d'indépendance en rapport avec des situations d'entreprises, avec l'utilisation des tables
Savoir formuler une hypothèse et tester un risque

Compétences visées

L'étudiant doit être capable de :

savoir identifier la loi de probabilité régissant un phénomène et retrouver le paramètre de Poisson
savoir poser des hypothèses
savoir les tester dans des situations classiques rencontrées en études et recherches commerciales

Prérequis

M1208, M2101

Contenus

Lois de probabilités usuelles (binomiale, poisson, normale avec lecture inverse de la table) et approximations
Droite d'HENRY
Test d'ajustement (Khi-2 avec maîtrise des calculs)
Échantillonnage, estimation (moyenne, fréquence)
Détermination de la taille d'un échantillon pour un risque alpha

Modalités de mise en œuvre

Utiliser des exemples de situation d'entreprise ou de marché (notamment en probabilité)
TIC, ERC

Prolongements

Analyse de variance, travail transversal avec Qualité

Mots clés

lois de probabilités, échantillon, intervalle de confiance, estimation, tests

Table des matières

1 Lois de probabilités discrètes usuelles **1**

1.1 Loi et variable de Bernoulli . . . 1
1.1.1 Définition . . . 1
1.1.2 Moments . . . 1
1.2 Loi et variable binomiales . . . 2
1.2.1 Définition . . . 2
1.2.2 Moments . . . 2
1.2.3 Somme de deux variables binomiales indépendantes . . . 3
1.2.4 Loi et variable fréquences . . . 3
1.3 Loi et variable multinomiales . . . 3
1.3.1 Exemple introductif . . . 3
1.3.2 Loi trinomiale . . . 4
1.3.3 Loi multinomiale . . . 4
1.4 Loi et variables hypergéométriques . . . 5
1.4.1 Définition . . . 5
1.4.2 Les moments . . . 5
1.4.3 Limite d'une variable hypergéométrique . . . 6
1.5 Loi et variable de Poisson . . . 6
1.5.1 Définition . . . 6
1.5.2 Les moments . . . 6
1.5.3 Somme de deux variables de Poisson indépendantes . . . 7
1.5.4 Limite d'une variable binomiale . . . 7
1.6 Loi et variable géométriques . . . 7
1.6.1 Définition . . . 7
1.6.2 Moments . . . 8
1.7 Exercices . . . 8

2 Lois de probabilités continues usuelles **15**

2.1 Loi et variable uniformes . . . 15
2.1.1 Définition . . . 15
2.1.2 Fonction de répartition . . . 15
2.1.3 Moments . . . 16
2.2 Loi exponentielle . . . 17
2.2.1 Définition . . . 17
2.2.2 Fonction de répartition . . . 17
2.2.3 Les moments . . . 17
2.3 Loi de Laplace-Gauss ou loi normale . . . 18
2.3.1 Définition . . . 18
2.3.2 Représentation graphique . . . 18

2.3.3 Moments . . . 19
2.3.4 Variable normale centrée réduite . . . 19
2.3.5 Fonction de répartition . . . 20
2.3.6 Table de l'écart réduit . . . 23
2.3.7 Exemples . . . 23
2.3.8 Remarques . . . 24
2.3.9 Relation entre la fonction de répartition et la densité de probabilité des loi normale et loi normale centrée réduite . . . 24
2.3.10 Propriétés . . . 25
2.3.11 Somme de deux variables normales indépendantes . . . 25
2.3.12 Approximation d'une loi binomiale par une loi normale . . . 25
2.3.13 Résumé sur les approximations de lois . . . 26
2.4 Loi et variable du χ^2 (Khi-deux) de Pearson . . . 26
2.4.1 Distribution du χ^2 . . . 26
2.5 Loi de Student-Fischer . . . 28
2.5.1 Définition . . . 28
2.5.2 Courbes . . . 28
2.5.3 Moments . . . 28
2.5.4 Tables . . . 29
2.6 Loi de Fischer-Snedecor . . . 30
2.6.1 Définition . . . 30
2.6.2 Courbes . . . 31
2.6.3 Moments . . . 32
2.6.4 Tables . . . 32
2.7 Exercices . . . 32

3 Échantillonnage et estimation 41
3.1 Introduction . . . 41
3.2 Estimation ponctuelle . . . 41
3.2.1 Introduction . . . 41
3.2.2 Estimateur sans biais . . . 43
3.3 Estimations pas intervalle de confiance . . . 46
3.3.1 Préliminaires . . . 46
3.3.2 Intervalle de confiance pour une proportion . . . 46
3.3.3 Intervalle de confiance pour l'espérance . . . 49
3.3.4 Intervalle de confiance pour la variance . . . 51
3.4 Exercices . . . 51

4 Le test du χ^2 55
4.1 Les données du problème . . . 55
4.2 Ajustement d'une distribution observée à une distribution théorique . . . 55
4.2.1 Construction du test . . . 55
4.3 Comparaison de distributions observées. Test d'indépendance. Test d'homogénéité . . . 62
4.3.1 Présentation du test . . . 62
4.3.2 Construction du test . . . 62
4.3.3 Exemples . . . 64
4.4 Comparaison de proportions . . . 66
4.4.1 Comparaison d'une proportion observée à une proportion théorique . . . 66
4.4.2 Comparaison de deux proportions . . . 67
4.5 Exercices . . . 68

Chapitre 1

Lois de probabilités discrètes usuelles

1.1 Loi et variable de Bernoulli

1.1.1 Définition

Soit une épreuve aléatoire comportant deux issues, deux événements élémentaires appelés souvent **succès** et **échec** dont les probabilités respectives sont p et q avec $p+q=1$.
On définit alors

$$\Omega = \{S, E\}$$

avec $p(S)=p$ et $p(E)=q$. Soit la variable aléatoire $X : \Omega \mapsto \{0,1\}$ telle que

$$X(S)=1 \quad ; \quad X(E)=0$$

La variable X est appelée **variable de Bernoulli** dont la loi de probabilité est

x_i	0	1	Total
p_i	q	p	1
$p_i x_i$	0	p	$E(X)=p$
$p_i x_i^2$	0	p	$E(X^2)=p$

On note cette loi

$$\boxed{X \rightsquigarrow \mathcal{B}(p)}$$

Remarque 1.1.1 Cette loi ne dépend que d'un paramètre p, la probabilité de succès.

1.1.2 Moments

1. Espérance : $\boxed{E(X)=p}$
2. Variance : $\boxed{V(X)=p-p^2=p(1-p)=pq}$
3. Écart-type : $\boxed{\sigma(X)=\sqrt{pq}}$

Exemple 1.1.1 Un entreprise possède 10 chaînes de fabrication C_1, C_2,...,C_{10}. Elle sait qu'une chaîne possède un problème mais elle ignore laquelle, elle choisit alors une chaîne au hasard. On considère la variable aléatoire X prenant la valeur 1 si la chaîne testée est la chaîne défaillante et 0 sinon. Dans ce cas X est une variable de Bernoulli de paramètre $\dfrac{1}{10}$. L'espérance et la variance de cette variable valent respectivement $E(X)=0,1$ et $V(X)=0,09$.

1.2 Loi et variable binomiales

1.2.1 Définition

Soit une épreuve de Bernoulli. À n répétitions indépendantes de cette épreuve de Bernoulli sont associées n variables aléatoires X_1, X_2,...,X_n indépendantes.
On considère la variable aléatoire $X = X_1 + X_2 + \ldots + X_n$. Cette variable X désigne le nombre de succès lors de n épreuves. L'univers image de la variable X est $\{0, 1, 2, \ldots, n\}$. On a

$$\boxed{p(\{X = k\}) = C_n^k p^k q^{n-k}}\ .$$

Preuve : L'événement $\{X = k\}$ est obtenu par le résultat de k succès et $n - k$ échecs. On peut avoir par exemple

$$\underbrace{S \ldots S}_{k \text{ fois}}\ \underbrace{E \ldots E}_{(n-k) \text{ fois}}$$

de probabilité $p^k q^{n-k}$, mais il existe C_n^k événements comportant k succès et $(n - k)$ échecs d'où le résultat.■

La variable ainsi définie suit une **loi binomiale** et on note : $\boxed{X \rightsquigarrow \mathcal{B}(n, p)}$

On a bien défini une loi de probabilité puisque pour $p \in]0; 1[$, $C_n^k p^k (1-p)^{n-k} \geq 0$, $\forall k \in \{0, 1, \ldots, n\}$ et $\sum_{k=1}^{n} p^k (1-p)^{n-k} = p + (1-p) = 1$ d'après la formule du binôme.

Remarque 1.2.1 Si une variable aléatoire X représente le nombre de succès dans une série de n expériences de Bernoulli identiques et indépendantes alors X suit une loi binomiale de paramètres n et p où p représente la probabilité de succès lors d'une épreuve de Bernoulli.

1.2.2 Moments

1. Espérance : $$\boxed{E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = np}$$

2. Variance : les variables X_i étant indépendantes,
$$\boxed{V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = npq}$$

3. Écart-type : $$\boxed{\sigma(X) = \sqrt{npq}}$$

Remarque 1.2.2 On a le rapport $\dfrac{p(\{X = k+1\})}{p(\{X = k\})} = \dfrac{C_n^{k+1} p^{k+1} q^{n-k-1}}{C_n^k p^k q^{n-k}} = \dfrac{n-k}{k+1} \times \dfrac{p}{q}$ d'où

$$p(\{X = k+1\}) = \frac{n-k}{k+1} \times \frac{p}{q} \times p(\{X = k\})$$

relation qui permet de disposer de $p(\{X = k+1\})$ lorsqu'on a déja $p(\{X = k\})$.

Exemple 1.2.1 Dans une population très nombreuse, on estime que la probabilité pour qu'une personne soit atteinte d'une maladie donnée est $0,1$. On choisit au hasard 1000 personnes de cette population (avec l'éventualité de choisir plusieurs fois la même personne). On note X la variable aléatoire représentant le nombre de personnes atteintes de la maladie parmi les 1000. X représente le nombre de succès (c'est-à-dire être atteint par la maladie) dans une suite de 1000 épreuves de Bernoulli (la personne est atteinte ou pas) identiques et indépendantes donc $X \rightsquigarrow \mathcal{B}(1000; 0,1)$.

1.2.3 Somme de deux variables binomiales indépendantes

Soient X_1 et X_2 telles que $X_1 \rightsquigarrow \mathcal{B}(n_1, p)$ et $X_2 \rightsquigarrow \mathcal{B}(n_2, p)$, X_1 et X_2 étant supposées indépendantes. Alors, la variable $Z = X_1 + X_2$ suit une loi binomiale $\mathcal{B}(n_1 + n_2, p)$.

Remarque 1.2.3 On peut généraliser cette propriété à l variables binomiales indépendantes.

1.2.4 Loi et variable fréquences

Soit $X \rightsquigarrow \mathcal{B}(n, p)$. On définit la variable $F_n = \dfrac{X}{n}$.
X désigne le nombre de succès obtenus au cours des n épreuves, F_n le nombre de succès divisé par le nombre d'épreuves soit la fréquence du succès. F_n est la **variable fréquence** associée à X :

$$F_n = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

L'univers image de F_n est $\left\{0, \dfrac{1}{n}, \ldots, \dfrac{k}{n}, \ldots, \dfrac{n}{n}\right\}$. On a $\{X = k\} = \left\{F_n = \dfrac{k}{n}\right\}$ donc

$$\boxed{p\left(\left\{F_n = \frac{k}{n}\right\}\right) = C_n^k p^k q^{n-k}}$$

Concernant les moments de cette variable,

- $E(F_n) = E\left(\dfrac{X}{n}\right) = \dfrac{1}{n}E(X) = \dfrac{np}{n} = p$ donc
$$\boxed{E(F_n) = p}$$
- $V(F_n) = \dfrac{1}{n^2}V(X) = \dfrac{npq}{n^2} = \dfrac{pq}{n}$ donc
$$\boxed{V(F_n) = \frac{pq}{n}} \quad \text{et} \quad \boxed{\sigma(F_n) = \sqrt{\frac{pq}{n}}}$$

Remarque 1.2.4 Soit $X \rightsquigarrow \mathcal{B}(n, p)$, X désigne le nombre de succès et $Y = n - X$ le nombre d'échecs. Par conséquent, $p(\{X = k\}) = p(\{Y = n - k\}) = C_n^k p^k q^{n-k}$.

1.3 Loi et variable multinomiales

1.3.1 Exemple introductif

On jette 50 fois une pièce de monnaie truquée. "Pile" apparaît avec une probabilité $0,3$, "face" avec une probabilité $0,6$, la pièce retombe sur la tranche avec une probabilité de $0,1$.
Quelle est la probabilité d'obtenir 20 "pile", 25 "face", 5 tranches ?
Cet événement peut être obtenu de la façon suivante

$$\underbrace{P\ldots P}_{20 \text{ fois}}\underbrace{F\ldots F}_{25 \text{ fois}}\underbrace{T\ldots T}_{5 \text{ fois}}$$

et sa probabilité vaut $(0,3)^{20}(0,6)^{25}(0,1)^5$. Le nombre de ces 50-uplets est égal au nombre de façons de disposer 20 fois la lettre P, 25 fois la lettre F et 5 fois la lettre T dans un mot de longueur 50. Ce nombre est $C_{50}^{20}C_{30}^{25}C_5^5 = C_{50}^{20}C_{30}^{25}$, la probabilité de cet événement vaut par conséquent

$$C_{50}^{20}C_{30}^{25}(0,3)^{20}(0,6)^{25}(0,1)^5.$$

1.3.2 Loi trinomiale

Soit une épreuve aléatoire à 3 issues A de probabilité p, B de probabilité q et C de probabilité r avec $p+q+r=1$. Pour n répétitions indépendantes de cette épreuve, on cherche la probabilité d'obtenir k fois A, i fois B et donc $n-k-i$ fois C. Cette probabilité vaut :

$$C_n^k C_{n-k}^i C_{n-k-i}^{n-k-i} p^k q^i r^{n-k-i}$$

or $C_{n-k-i}^{n-k-i}=1$ et $C_n^k C_{n-k}^i = \dfrac{n!}{k!i!(n-k-i)!}$. Conclusion, cette probabilité vaut :

$$\boxed{p = \frac{n!}{k!i!(n-k-i)!} p^k q^i r^{n-k-i}}$$

Exemple 1.3.1 Une équipe de football gagne un match avec une probabilité de $0,3$, le perd avec une probabilité de $0,4$, fait match nul avec une probabilité de $0,3$. Sur les 36 matchs joués dans l'année, on cherche la probabilité d'obtenir 15 succès, 18 échecs et 3 nuls. Cette probabilité vaut

$$p = C_{36}^{15} C_{21}^{18} C_3^3 (0,3)^{15}(0,4)^{18}(0,3)^3 = \frac{36!(0,3)^{15}(0,4)^{18}(0,3)^3}{15!18!3!}.$$

1.3.3 Loi multinomiale

Supposons qu'il y ait dans une urne N boules de r couleurs distinctes C_1, C_2,…,C_r. Soit n_i le nombre de boules de couleur C_i et $p_i = \dfrac{n_i}{N}$ la proportion de boules de la couleur C_i dans l'urne. On a

$$N = \sum_{i=1}^{r} n_i = n_1 + n_2 + \ldots + n_r \quad \text{et} \quad \sum_{i=1}^{r} p_i = 1.$$

Supposons que l'on effectue un tirage de n boules, chaque boule étant remise dans l'urne avant le tirage de la boule suivante ; les tirages répétés des boules sont des épreuves indépendantes. On cherche la probabilité d'obtenir l'événement A défini par

- m_1 boules de la couleur C_1
- m_2 boules de la couleur C_2

$\vdots$

- m_i boules de la couleur C_i

$\vdots$

- m_r boules de la couleur C_r

avec $m_1 + m_2 + \ldots + m_r = n$.

Cet événement est réalisé par exemple avec le n-uplet

$$\underbrace{C_1 \ldots C_1}_{m_1 \text{ boules } C_1} \quad \underbrace{C_2 \ldots C_2}_{m_2 \text{ boules } C_2} \quad \ldots \quad \underbrace{C_r \ldots C_r}_{m_r \text{ boules } C_r}$$

de probabilité $p_1^{m_1} p_2^{m_2} \ldots p_r^{m_r}$. Le nombre de ces n-uplets est égal au nombre de façons de disposer m_1 fois la lettre C_1, m_2 fois la lettre C_2,…, m_r fois la lettre C_r dans un mot de longueur $n = m_1 + m_2 + \ldots + m_r$ d'où la probabilité de l'événement A :

$$p(A) = C_n^{m_1} C_{n-m_1}^{m_2} \ldots C_{m_r}^{m_r} (p_1)^{m_1} (p_2)^{m_2} \ldots (p_r)^{m_r}.$$

On a la relation $C_n^{m_1} C_{n-m_1}^{m_2} \ldots C_{m_r}^{m_r} = \dfrac{n!}{m_1! m_2! \ldots m_r!}$, par conséquent,

$$\boxed{p(A) = \frac{n!}{m_1!m_2!\dots m_r!}(p_1)^{m_1}(p_2)^{m_2}\dots(p_r)^{m_r}}$$

Exemple 1.3.2 Une urne est composée de 10% de boules rouges, 20% de boules blanches, 40% de boules vertes, 30% de noires. Le nombre de boules de l'urne est $N > 20$. On effectue un tirage avec remise de 12 boules. Quelle est la probabilité d'obtenir 3 boules rouges, 5 boules blanches, 3 boules vertes et une boule noire ?
Il suffit d'appliquer la formule précédente :

$$p = \frac{12!}{3!5!3!1!}(0,1)^3(0,2)^5(0,4)^3(0,3)^1.$$

1.4 Loi et variables hypergéométriques

1.4.1 Définition

Soit une urne contenant N boules dont a boules blanches et b boules noires avec $a+b=N$. On effectue n tirages d'une boule sans remise (ou on prélève simultanément n boules) avec $n \leq N$. Le tirage sans remise est dit **exhaustif**. Soit X la variable aléatoire représentant le nombre de boules blanches obtenues.

La variable X est dite **hypergéométrique**. On utilise la notation

$$\boxed{X \rightsquigarrow \mathcal{H}(N,a,n)}$$

Cette loi dépend de trois paramètres et

$$\boxed{p(\{X=k\}) = \frac{C_a^k C_b^{n-k}}{C_N^n}}$$

En effet, $\{X=k\}$ est l'ensemble des parties à k éléments parmi a donc $Card(\{X=k\}) = C_a^k C_b^{n-k}$.

Remarque 1.4.1 Si p est la proportion des boules blanches de l'urne, q celle des noires, on a $p = \frac{a}{N}$ et $q = \frac{b}{N}$ avec $p+q=1$ donc $a=pN$, $b=qN$ et $p(\{X=k\}) = \frac{C_{pN}^k C_{qN}^{n-k}}{C_N^n} = \frac{C_{pN}^k C_{(1-p)N}^{n-k}}{C_N^n}$. La loi est notée $\mathcal{H}(N,p,n)$.

1.4.2 Les moments

On admettra les propriétés suivantes :

— L'espérance mathématique est donnée par :

$$\boxed{E(X) = np}$$

la formule est identique à celle de la loi binomiale.

— La variance est définie par :

$$\boxed{V(X) = npq\frac{N-n}{N-1} = npq\rho}$$

avec $\rho = \frac{N-n}{N-1}$ définissant le **coefficient d'exhaustivité**.

Remarque 1.4.2 Généralement $n > 1$ donc $\rho < 1$. La variance d'une variable hypergéométrique (tirages sans remise) est inférieure à la variance de la variable binomiale (tirages avec remise).

Exemple 1.4.1 Chaque matin, un professeur interroge 4 étudiants pour tester leur connaissance du cours. Une indiscrétion lui permet de savoir que dans la classe composée de 45 étudiants, 10 ne connaissent pas le cours. On se trouve dans la situation d'un ensemble E comprenant 45 éléments dont une proportion $\frac{10}{45}$ est de type 1 (les étudiants ne connaissent pas le cours), le professeur interroge 4 étudiants successivement, sans interroger deux fois le même (ce qui correspond à 4 tirages successifs sans remise d'un élément de E) alors la variable aléatoire X représentant le nombre d'éléments de type 1 obtenu suit une loi hypegéométrique $\mathcal{H}\left(45, 4, \frac{10}{45}\right)$.

1.4.3 Limite d'une variable hypergéométrique

Soit X une variable hypergéométrique, $X \rightsquigarrow \mathcal{H}(N, p, n)$. Lorsque N tend vers $+\infty$,

$$\mathcal{H}(N, p, n) \to \mathcal{B}(n, p).$$

En effet, le nombre N de boules étant infiniment grand, la non-remise de la boule tirée ne modifie presque pas la proportion de boules blanches. Dans la pratique,

$$\mathcal{H}(N, p, n) \xrightarrow{\text{si } N>10n} \mathcal{B}(n, p).$$

1.5 Loi et variable de Poisson

On rappelle la formule

$$\boxed{\lim_{n\to+\infty} \sum_{k=0}^{n} \frac{x^k}{k!} = \sum_{k\in\mathbb{N}} \frac{x^k}{k!} = e^x}$$

1.5.1 Définition

On dit que la variable aléatoire X suit une **loi de Poisson** de paramètre λ ($\lambda \in \mathbb{R}^{+*}$) si et seulement si l'univers image de X est $\mathbb{N}$ et

$$\boxed{p(\{X = k\}) = \frac{e^{-\lambda}\lambda^k}{k!}}$$

On note cette variable

$$\boxed{X \rightsquigarrow \mathcal{P}(\lambda)}$$

Puisque $\sum_{i\in\mathbb{N}} p_i = \sum_{i\in\mathbb{N}} e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{i\in\mathbb{N}} \frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1$ et $\forall k \in \mathbb{N}$, $\frac{e^{-\lambda}\lambda^k}{k!} \geq 0$, on a bien une loi de probabilité.

Remarque 1.5.1 La loi de Poisson est encore appelée **loi des événements rares**. On peut admettre que le nombre de pannes survenant sur une machine donnée au cours d'une période donnée suit une loi de Poisson, ou encore le nombre d'accidents qui se produisent à un carrefour donné pendant une période donnée.

1.5.2 Les moments

— Espérance : on a la relation

$$\boxed{E(X) = \lambda}$$

Preuve : Soit $X \rightsquigarrow \mathcal{P}(\lambda)$ alors $E(X) = \sum_{k\in\mathbb{N}} ke^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k\in\mathbb{N}} k\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{+\infty} \frac{\lambda^k}{(k-1)!} = \lambda e^{-\lambda}\sum_{k=1}^{+\infty} \frac{\lambda^{k-1}}{(k-1)!}$.

En réalisant un changement d'indice, $E(X) = \lambda e^{-\lambda}\sum_{i=0}^{+\infty} \frac{\lambda^i}{i!} = \lambda e^{-\lambda}e^{\lambda} = \lambda$. ■

— Variance et écart-type : on a les résultats suivants

$$\boxed{V(X) = \lambda} \text{ et } \boxed{\sigma(X) = \sqrt{\lambda}}$$

Preuve : Déterminons $E(X(X-1)) = E(X^2) - E(X)$.
$E(X(X-1)) = \sum_{k\in\mathbb{N}} \frac{k(k-1)\lambda^k e^{-\lambda}}{k!} = \sum_{k\in\mathbb{N}/\{0,1\}} \frac{\lambda^k e^{-\lambda}}{(k-2)!} = e^{-\lambda}\lambda^2 \sum_{k\in\mathbb{N}/\{0,1\}} \frac{\lambda^{k-2}}{(k-2)!}$. On effectue un changement d'indice dans la somme et $E(X(X-1)) = e^{-\lambda}\lambda^2 \sum_{i\in\mathbb{N}} \frac{\lambda^i}{i!} = e^{-\lambda}\lambda^2 e^{\lambda} = \lambda^2$.
Ainsi $E(X^2) - E(X) = \lambda^2 \Leftrightarrow E(X^2) = \lambda^2 + \lambda$. Par conséquent, $V(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$ donc $V(X) = \lambda$ et $\sigma(X) = \sqrt{\lambda}$. ■

Remarque 1.5.2

- $\frac{p_k}{p_{k-1}} = \frac{e^{-\lambda}\lambda^k (k-1)!}{e^{-\lambda}\lambda^{(k-1)} k!} = \frac{\lambda}{k}$ ce qui signifie que $p_k = \frac{\lambda}{k} p_{k-1}$.
- Il existe des tables (voir annexe A1) donnant $p_k = p(\{X = k\})$ pour différentes valeurs de k et des tables donnant $\sum_{k=0}^{n} p_k$ (voir annexes A2 et A3) c'est-à-dire la valeur de la fonction de répartition de X en n pour $X \rightsquigarrow \mathcal{P}$. Pour le calcul de $p(\{X > k\})$, il suffit d'utiliser la relation $p(A) + p(\overline{A}) = 1$ avec $A = \{X > k\}$ et $\overline{A} = \{X \leq k\}$ ce qui permet de calculer $p(\{X > k\}) = 1 - p(\{X \leq k\}) = 1 - \sum_{j=0}^{k} p_j$.

Exemple 1.5.1 Dans une entreprise. on admet que le nombre de pièces défectueuses produites par minute est une variable aléatoire X avec $X \rightsquigarrow \mathcal{P}(3)$. Déterminons la probabilité de l'événement A : "le nombre de pièces défectueuses produites en 1 minute est supérieur à 3".
Par définition, $p(A) = p(\{X > 3\}) = 1 - p(\{X \leq 3\}) = 1 - 0,6472 = 0,3528$, ceci en utilisant l'annexe A2.

1.5.3 Somme de deux variables de Poisson indépendantes

Soient $X_1 \rightsquigarrow \mathcal{P}(\lambda_1)$ et $X_2 \rightsquigarrow \mathcal{P}(\lambda_2)$, X_1 et X_2 étant indépendantes. La variable $X = X_1 + X_2$ suit alors une loi de Poisson de paramètre $\lambda = \lambda_1 + \lambda_2$

$$X = X_1 + X_2 \rightsquigarrow \mathcal{P}(\lambda_1 + \lambda_2)$$

Remarque 1.5.3 On peut généraliser ce résultat à n variables de Poisson indépendantes.

1.5.4 Limite d'une variable binomiale

— Soit X une variable binomiale $\mathcal{B}(n,p)$. Si p tend vers zéro lorsque n tend vers $+\infty$ de telle sorte que np ait une limite finie λ, la loi binomiale $\mathcal{B}(n,p)$ tend vers une loi de Poisson $\mathcal{P}(\lambda)$.
— Dans la pratique on approxime la loi binomiale $\mathcal{B}(n,p)$ par une loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda = np$ si $n \geq 30$, $p < 0,1$ et $np \leq 10$.
— L'intérêt de cette approximation est l'utilisation des tables de la loi de Poisson, plus commodes que celles de la loi binomiale.

1.6 Loi et variable géométriques

1.6.1 Définition

Soit une épreuve aléatoire à deux issues, succès et échec, de probabilités respectives p et q avec $p+q=1$. On répète cette épreuve (épreuves indépendantes) jusqu'à obtenir le premier succès. On considère la variable

aléatoire X donnant le rang du premier succès ou encore le nombre d'épreuves nécessaires à l'obtention d'un premier succès

Lorsque X est une variable géométrique, l'univers image est $\mathbb{N}^*$.

L'événement $\{X = k\}$ est obtenu par la réalisation de $k-1$ échecs puis d'un succès.

$$\boxed{p(\{X = k\}) = q^{k-1}p}$$

On notera alors :

$$\boxed{X \rightsquigarrow \mathcal{G}(p)}$$

On peut vérifier qu'on a bien défini une loi de probabilité. On a besoin pour cela du rappel suivant :

Rappel : Si $|x| \leq 1$ alors $\lim\limits_{n\to+\infty} \sum\limits_{k=0}^{n} x^k = \dfrac{1}{1-x}$.

On a par conséquent $\forall k \in \mathbb{N}^\star$, $q^{k-1}p \geq 0$ et $\sum\limits_{k=1}^{n} q^{k-1}p = p\sum j = 0^{n-1}q^j \xrightarrow{n\to+\infty} p\dfrac{1}{1-q} = 1$ donc $\sum\limits_{k=1}^{+\infty} q^{k-1}p = 1$.

Exemple 1.6.1 On considère une urne contenant des boules blanches en proportion p et des boules noires en proportion $q = 1-p$, on tire une infinité de fois une boule avec remise. Avec l'exact formalisme précédent, la variable aléatoire X représente le rang où on obtient une boule blanche pour la première fois.

1.6.2 Moments

— L'espérance mathématique est définie par

$$\boxed{E(X) = \frac{1}{p}}$$

Preuve : $E(X) = X$ admet une espérance si et seulement si la série $\sum\limits_{k} kp(\{X = k\})$ est absolument convergente. La série étant à termes positifs, la convergence absolue est éqivalente à la convergence.

Rappel : Si $x \in]-1;1[$ alors $\lim\limits_{n\to} \sum\limits_{k=1}^{n} kx^{k-1} = \dfrac{1}{(1-x)^2}$.

On a par conséquent $\sum\limits_{k=0}^{n} kp(\{X = k\}) = \sum\limits_{k=0}^{n} kq^{k-1}p = p\sum\limits_{k=0}^{n} kq^{k-2} \xrightarrow{n\to+\infty} p\dfrac{1}{(1-q)^2} = \dfrac{p}{p^2} = \dfrac{1}{p}$ ■

— La variance et l'écart-type sont respectivement définis par

$$\boxed{V(X) = \frac{q}{p^2}} \quad \text{et} \quad \boxed{\sigma(X) = \frac{\sqrt{q}}{p}}$$

On admettra ce résultat.

1.7 Exercices

Exercice 1 Pour réaliser le montage d'un système électronique, on dispose de résistances issues d'une production importante, où l'on sait que le pourcentage P de résistances défectueuses est de 5%. On doit utiliser 4 résistances.

1. Quelle est la probabilité d'en avoir 3 de mauvaises ?

2. Quelle est la probabilité d'en avoir un nombre inférieur ou égal à 3 de mauvaises ?

Exercice 2 Une boîte contient 4 boules rouges, 3 boules vertes et 7 boules jaunes. On tire simultanément 2 boules de la boîte et on suppose que les tirages sont équiprobables.

1. On considère les événements suivants :
   A : "obtenir 2 boules de la même couleur",
   B : "obtenir 2 boules de couleurs différentes".
   Calculez les probabilités $p(A)$ et $p(B)$.
2. On répète 10 fois l'épreuve précédente en remettant les 2 boules tirées dans la boîte, après chaque tirage. Les 10 épreuves aléatoires élémentaires sont donc indépendantes. On note X la variable aléatoire qui, à chaque partie de 10 épreuves, associe le nombre de fois où l'événement A est réalisé.
   (a) Expliquez pourquoi X suit une loi binomiale. Donner les paramètres de cette loi.
   (b) Donnez une loi de probabilité de X en complétant, après l'avoir reproduit, le tableau suivant, dans lequel on fera figurer des valeurs approchées arrondies avec un seul chiffre différent de zéro.

k	0	$\cdots$
$p(\{X = k\})$	$\cdots$	$\cdots$

   (c) Calculez l'espérance mathématique $E(X)$ de X. Que représente $E(X)$?

Exercice 3 On envisage l'installation d'une pompe à chaleur en relève de chaudière dans un hôtel "deux étoiles" en construction.
On se propose d'étudier si le contrat de maintenance forfaitaire annuel proposé par l'installateur, après la période de garantie d'un an, est plus avantageux que la facturation au prix réel des interventions ponctuelles.
Une étude statistique permet au constructeur d'affirmer que la probabilité de l'événement "la pompe à chaleur tombe en panne une fois pendant un mois donné" est $0,125$.
Dans un but de simplification, on admet que, pendant un mois donné, la pompe à chaleur ne peut tomber en panne qu'au plus une fois et que les pannes éventuelles survenues deux mois d'une même année sont indépendantes. On note X la variable aléatoire qui, à chaque année (de douze mois), associe le nombre de pannes survenues à la pompe.

1. Expliquez pourquoi X suit une loi binomiale. Donner les paramètres de cette loi.
2. Calculez la probabilité des événements suivants :
   A : "il n'y a pas de panne dans l'année",
   B : "il y a au plus deux pannes dans l'année".
3. Calculez l'espérance mathématique, notée $E(X)$, de la variable aléatoire X. Que représente $E(X)$?
4. Les résultats d'une étude statistique menée auprès de nombreux utilisateurs de ce modèle de pompe à chaleur n'ayant souscrit de contrat de maintenance annuel permettent d'admettre que le coût d'une intervention est de 320 euros. Soit Y la variable aléatoire qui à chaque année associe le montant total en euros des frais de réparation de la pompe à chaleur.
   (a) Écrivez une relation entre les variables Y et X.
   (b) Déterminez l'espérance mathématique, notée $E(Y)$, de la variable Y. Que représente $E(Y)$?
   (c) Le contrat de maintenance forfaitaire annuel de la pompe à chaleur est proposé par l'installateur au prix de 685 euros TTC.
   Quelle est la solution de maintenance la plus intéressante sur une longue période ?
5. On approche la loi binomiale du 1. par une loi de Poisson de paramètre $\lambda = np$ où n et p sont les paramètres de cette loi binomiale.
   En utilisant la loi de Poisson, déterminez les probabilités respectives de deux événements A et B de la question 2.

6. On considère que, pour un événement, l'approximation d'une loi binomiale par une loi de Poisson est justifiée lorsque l'erreur relative $\frac{p-p'}{p}$ est, en valeur absolue, inférieure à 10% (p étant la probabilité de cet événement mesurée avec la loi de Poisson). Pour chacun des deux événements précédents, déterminez si l'approximation de la loi binomiale du 1. par la loi de Poisson du 5. est justifiée.

Exercice 4 La probabilité qu'une imprimante de modèle PRINT ne puisse transcrire correctement un caractère est $0,0005$; on suppose que les qualités de transmission des caractères sont indépendantes.
On désigne par X la variable aléatoire qui à tout lot de 10000 caractères associe le nombre de caractères transcrits incorrectement par l'imprimante.

1. Quelle est la loi de probabilité de X ?
2. On admet que la loi de probabilité suivie par X peut être approchée par une loi de Poisson dont on déterminera le paramètre.
   Quelle est la probabilité que, parmi 10000 caractères à imprimer,
   (a) tous soient transcrits correctement ?
   (b) au moins 9998 soient transcrits correctement ?
   (c) plus de 5 caractères soient transcrits incorrectement ?

Exercice 5 Dans une urne, il y a 10 boules blanches et 18 boules rouges indiscernables au toucher. On considère l'épreuve qui consiste à extraire, au hasard, l'une après l'autre et sans remise, deux boules de l'urne. On est dans une situation d'équiprobabilité.

1. Déterminer la probabilité de l'événement suivant :
   E : "la première boule tirée est blanche".
2. On répète 5 fois de suite l'épreuve précédente. Après chaque épreuve, les 2 boules tirées sont remises dans l'urne : les 5 épreuves élémentaires précédentes sont donc indépendantes.
   Soit X la variable aléatoire qui, à chaque partie de 5 épreuves, associe le nombre de fois que se produit l'événement E.
   (a) Expliquer pourquoi X suit une loi binomiale et préciser les paramètres de cette loi.
   (b) Calculer la probabilité de l'événement
   F : "E se produit exactement 2 fois".

Exercice 6 Soit X une variable aléatoire qui suit une loi de Poisson de paramètre 4. Déterminez la probabilité d'avoir $7 \leq X \leq 9$.

Exercice 7 3% des bouteilles d'eau fabriquées par une usine sont défectueuses. On appelle X la variable aléatoire qui, à tout lot de 100 bouteilles prises au hasard, associe le nombre de bouteilles défectueuses. On admet que X suit une loi de Poisson de paramètre 3.
Trouvez la probabilité de chacun des 3 événements suivants :

1. "Un tel lot n'a aucune bouteille défectueuse"
2. "Un tel lot a deux bouteilles défectueuses"
3. "Un tel lot a trois bouteilles défectueuses"

Exercice 8 Une urne contient 6 boules blanches et 4 boules noires.

1. On tire dans cette urne trois fois 1 boule avec remise de cette boule après tirage. On note X le nombre de boules blanches obtenues.
   Déteminez la loi de X puis donner les valeurs de $E(X)$ et $V(X)$.

2. On tire dans cette urne trois fois une boule sans remise de cette boule après tirage. On note Y le nombre de boules blanches obtenues.
   Déterminez la loi de Y puis donner les valeurs de $E(Y)$ et $V(Y)$.

Exercice 9 On considère une urne contenant 3 boules blanches et 7 boules noires. On tire successivement et sans remise les dix boules de l'urne. On note X_1 le numéro du tirage où l'on obtient une boule blanche pour la première fois. Déterminez la loi de X_1.

Exercice 10 Le système de navigation d'un navire comporte un équipement E dont la fiabilité est exponentielle avec un MTBF (mean time between failures - temps moyen entre pannes) théorique égal à 500 heures. Ce navire doit effectuer une mission de 2500 heures sans ravitaillement technique.

1. Si l'on veut la certitude pratique (98%) de la continuité de la fonction assurée par E, combien au départ doit-on emporter d'équipements E ?
2. On suppose maintenant que l'on puisse réparer à bord l'équipement E et l'on estime que le temps d'indisponibilité pour la réparation n'excède pas 250 heures. Combien doit-on emporter d'équipements E pour avoir la certitude (à 100% près par hypothèse) de la continuité de la fonction assurée par E ?
   N.B. On estime que la réparation ne dégrade pas le $MTBF = 500$ heures.

ANNEXE A1 - Probabilités individuelles de la loi de Poisson $\mathcal{P}(\lambda)$.

Cette table donne $p(\{X = k\}) = \dfrac{e^{-\lambda}\lambda^k}{k!}$ pour $X \rightsquigarrow \mathcal{P}(\lambda)$:

k \ λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494
4		0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111
5				0,0001	0,0002	0,0004	0,0007	0,0012	0,0020
6							0,0001	0,0002	0,0003

k \ λ	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0
0	0,3679	0,2231	0,1353	0,0821	0,0492	0,0302	0,0183	0,0111	0,0067
1	0,3679	0,3347	0,2707	0,2052	0,1454	0,1057	0,0733	0,0500	0,0337
2	0,1839	0,2510	0,2707	0,2565	0,2240	0,1850	0,1465	0,1125	0,0842
3	0,0613	0,1255	0,1804	0,2132	0,2240	0,2158	0,1954	0,1898	0,1404
4	0,0153	0,0471	0,0902	0,1336	0,1680	0,1888	0,1954	0,1898	0,1755
5	0,0031	0,0141	0,0361	0,0668	0,1008	0,1322	0,1563	0,1708	0,1755
6	0,0005	0,0035	0,0120	0,0278	0,0504	0,0771	0,1042	0,1281	0,1462
7	0,0001	0,0008	0,0034	0,0099	0,0216	0,0385	0,0595	0,0824	0,1044
8		0,0001	0,0009	0,0031	0,0081	0,0169	0,0298	0,0463	0,0653
9			0,0002	0,0009	0,0027	0,0066	0,0132	0,0232	0,0363
10				0,0002	0,0008	0,0023	0,0053	0,0104	0,0181
11					0,0002	0,0007	0,0019	0,0043	0,0082
12					0,0001	0,0002	0,0006	0,0016	0,0034
13						0,0001	0,0002	0,0006	0,0013
14							0,0001	0,0002	0,0005
15								0,0001	0,0002
16									0,0001

ANNEXE A2 - Probabilités individuelles et cumulées de la loi de Poisson $\mathcal{P}(\lambda)$.

Cette table donne $p(\{X = k\}) = \dfrac{e^{-\lambda}\lambda^k}{k!}$ pour $X \rightsquigarrow \mathcal{P}(\lambda)$ et $F(k) = \displaystyle\sum_{l=0}^{k} e^{-\lambda}\frac{\lambda^l}{l!}$:

λ	1		2		3		4		5	
k	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$
0	0,3679	0,3679	0,1353	0,1353	0,0498	0,0498	0,0183	0,0183	0,0067	0,0067
1	0,3679	0,7358	0,2707	0,4060	0,1494	0,1991	0,0733	0,0916	0,0337	0,0404
2	0,1839	0,9197	0,2707	0,6767	0,2240	0,4232	0,1465	0,2381	0,0842	0,1247
3	0,0613	0,9810	0,1804	0,8571	0,2240	0,6472	0,1954	0,4335	0,1404	0,2650
4	0,0153	0,9963	0,0902	0,9473	0,1680	0,8152	0,1954	0,6288	0,1755	0,4405
5	0,0031	0,9994	0,0361	0,9834	0,1008	0,9161	0,1563	0,7851	0,1755	0,6160
6	0,0005	0,9999	0,0120	0,9955	0,0504	0,9665	0,1042	0,8893	0,1462	0,7622
7	0,0001	1,0000	0,0034	0,9989	0,0216	0,9881	0,0595	0,9489	0,1044	0,8666
8			0,0009	0,9998	0,0081	0,9962	0,0298	0,9786	0,0653	0,9319
9			0,0002	1,0000	0,0027	0,9989	0,0132	0,9919	0,0363	0,9682
10					0,0008	0,9997	0,0053	0,9972	0,0181	0,9863
11					0,0002	0,9999	0,0019	0,9991	0,0082	0,9945
12					0,0001	1,0000	0,0006	0,9997	0,0036	0,9980
13							0,0002	0,9999	0,0013	0,9993
14							0,0001	1,0000	0,0005	0,9998
15									0,0002	0,9999
16									0,0001	1,0000

λ	6		7		8		9		10	
k	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$
0	0,0025	0,0025	0,0009	0,0009	0,0003	0,0003	0,0001	0,0001		
1	0,0149	0,0174	0,0064	0,0073	0,0027	0,0030	0,0011	0,0012	0,0005	0,0005
2	0,0446	0,0620	0,0223	0,0296	0,0107	0,0138	0,0050	0,0062	0,0023	0,0028
3	0,0892	0,1512	0,0521	0,0818	0,0286	0,0424	0,0150	0,0212	0,0076	0,0104
4	0,1339	0,2851	0,0912	0,1730	0,0573	0,0996	0,0337	0,0550	0,0189	0,0293
5	0,1606	0,4457	0,1277	0,3007	0,0916	0,1912	0,0607	0,1157	0,0378	0,0671
6	0,1606	0,6063	0,1490	0,4497	0,1221	0,3134	0,0911	0,2068	0,0631	0,1302
7	0,1377	0,7440	0,1490	0,5987	0,1396	0,4530	0,1171	0,3239	0,0901	0,2203
8	0,1033	0,8472	0,1304	0,7291	0,1396	0,5925	0,1318	0,4557	0,1126	0,3329
9	0,0688	0,9161	0,1014	0,8305	0,1241	0,7166	0,1318	0,5874	0,1251	0,4580
10	0,0413	0,9574	0,0710	0,9015	0,0993	0,8159	0,1186	0,7060	0,1251	0,5381
11	0,0225	0,9799	0,0452	0,9466	0,0722	0,8881	0,0970	0,8030	0,1137	0,6968
12	0,0113	0,9912	0,0264	0,9730	0,0481	0,9362	0,0728	0,8758	0,0948	0,7916
13	0,0052	0,9964	0,0142	0,9872	0,0296	0,9658	0,0504	0,9261	0,0729	0,8645
14	0,0022	0,9986	0,0071	0,9943	0,0169	0,9827	0,0324	0,9585	0,0521	0,9166
15	0,0009	0,9995	0,0033	0,9976	0,0090	0,9918	0,0194	0,9780	0,0347	0,9513
16	0,0003	0,9998	0,0014	0,9990	0,0045	0,9963	0,0109	0,9889	0,0217	0,9730
17	0,0001	1,0000	0,0006	0,9996	0,0021	0,9984	0,0058	0,9947	0,0128	0,9857
18			0,0002	0,9999	0,0009	0,9993	0,0029	0,9976	0,0071	0,9928
19			0,0001	1,0000	0,0004	0,9997	0,0014	0,9989	0,0037	0,9965
20					0,0002	0,9999	0,0006	0,9996	0,0019	0,9984
21					0,0001	1,0000	0,0003	0,9998	0,0009	0,9993
22							0,0001	0,9999	0,0004	0,9997
23								1,0000	0,0002	0,9999
24									0,0001	1,0000

ANNEXE A3 - Probabilités individuelles et cumulées de la loi de Poisson $\mathcal{P}(\lambda)$.

Cette table donne $p(\{X = k\}) = \dfrac{e^{-\lambda}\lambda^k}{k!}$ pour $X \rightsquigarrow \mathcal{P}(\lambda)$ et $F(k) = \displaystyle\sum_{l=0}^{k} e^{-\lambda}\frac{\lambda^l}{l!}$:

λ	11		12		13		14		15	
k	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$	$p(k,\lambda)$	$F(k)$
0										
1	0,0002	0,0002	0,0001	0,0001						
2	0,0010	0,0012	0,0004	0,0005	0,0002	0,0001	0,0001			
3	0,0037	0,0049	0,0018	0,0023	0,0008	0,0010	0,0004	0,0005	0,0002	0,0002
4	0,0102	0,0151	0,0053	0,0076	0,0027	0,0037	0,0013	0,0018	0,0007	0,0009
5	0,0224	0,0375	0,0127	0,0203	0,0070	0,0107	0,0037	0,0055	0,0019	0,0028
6	0,0411	0,0786	0,0255	0,0458	0,0152	0,0259	0,0087	0,0142	0,0048	0,0076
7	0,0646	0,1432	0,0437	0,0895	0,0281	0,0540	0,0174	0,0316	0,0104	0,0180
8	0,0888	0,2320	0,0655	0,1550	0,0457	0,0997	0,0304	0,0620	0,0194	0,0374
9	0,1085	0,3405	0,0874	0,2424	0,0661	0,1658	0,0473	0,1093	0,0324	0,0698
10	0,1194	0,4599	0,1048	0,3472	0,0859	0,2517	0,0663	0,1756	0,0486	0,1184
11	0,1194	0,5793	0,1144	0,4616	0,1015	0,3532	0,0844	0,2600	0,0663	0,1847
12	0,1094	0,6887	0,1144	0,5760	0,1099	0,4631	0,0984	0,3584	0,0829	0,2676
13	0,0926	0,7813	0,1056	0,6816	0,1099	0,5730	0,1060	0,4644	0,0956	0,3622
14	0,0728	0,8541	0,0905	0,7721	0,1021	0,6751	0,1060	0,5704	0,1024	0,4656
15	0,0534	0,9075	0,0724	0,8445	0,0885	0,7636	0,0989	0,6693	0,1024	0,5680
16	0,0367	0,9442	0,0543	0,8988	0,0719	0,8355	0,0866	0,7559	0,0960	0,6640
17	0,0237	0,9679	0,0383	0,9371	0,0550	0,8905	0,0713	0,8272	0,0847	0,7487
18	0,0145	0,9824	0,0255	0,9626	0,0397	0,9302	0,0554	0,8826	0,0706	0,8193
19	0,0084	0,9908	0,0161	0,9787	0,0272	0,9574	0,0409	0,9235	0,0558	0,8751
20	0,0046	0,9954	0,0097	0,9884	0,0177	0,9751	0,0286	0,9521	0,0418	0,9169
21	0,0024	0,9978	0,0055	0,9939	0,0109	0,9680	0,0191	0,9712	0,0299	0,9468
22	0,0012	0,9990	0,0030	0,9969	0,0065	0,9925	0,0121	0,9833	0,0204	0,9672
23	0,0006	0,9996	0,0016	0,9985	0,0037	0,9962	0,0074	0,9907	0,0133	0,9805
24	0,0003	0,9999	0,0008	0,9993	0,0020	0,9982	0,0043	0,9950	0,0083	0,9888
25	0,0001	1,0000	0,0004	0,9997	0,0010	0,9992	0,0024	0,9974	0,0050	0,9938
26			0,0002	0,9999	0,0005	0,9997	0,0013	0,9987	0,0029	0,9967
27			0,0001	1,0000	0,0002	0,9999	0,0007	0,9994	0,0016	0,9983
28					0,0001	1,0000	0,0003	0,9997	0,0009	0,9992
29							0,0002	0,9999	0,0004	0,9996
30							0,0001	1,0000	0,0002	0,9998
31									0,0001	0,9999
32									0,0001	1,0000

Chapitre 2

Lois de probabilités continues usuelles

2.1 Loi et variable uniformes

2.1.1 Définition

On dit que la loi de probabilité d'une variable aléatoire réelle est **uniforme** sur un segment $[a;b]$, avec $0 \leq a < b$, si sa densité de probabilité f est définie par

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{pour } x \in [a;b] \\ 0 & \text{pour } x < a \text{ ou } x > b \end{cases}$$

On note alors $X \rightsquigarrow \mathcal{U}([a;b])$. f admet la représentation graphique de la Figure 2.1.

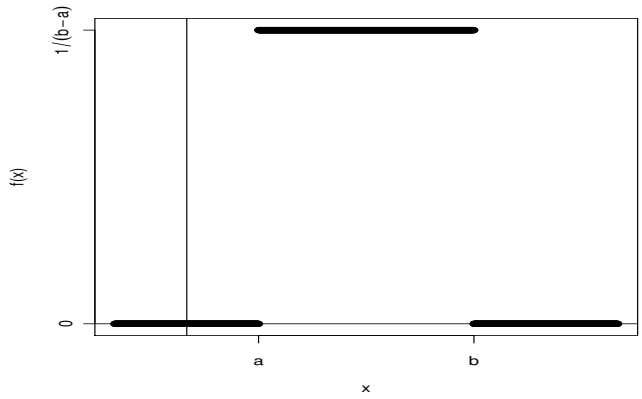

FIGURE 2.1 –

On a bien une densité de probabilité puisque

- $f(x) \geq 0 \ \forall x \in \mathbb{R}$,
- f est continue sur $]-\infty;a[\cup]a;b[\cup]b;+\infty[$,
- $\displaystyle\int_{-\infty}^{+\infty} f(t)dt = \int_{-\infty}^{a} f(t)dt + \int_{a}^{b} f(t)dt + \int_{b}^{+\infty} f(t)dt = 0 + 1 + 0 = 1.$

2.1.2 Fonction de répartition

1. On sait que $F(x) = p(\{X \leq x\}) = \displaystyle\int_{-\infty}^{x} f(t)dt$ donc si $X \rightsquigarrow \mathcal{U}([a;b])$,

$$F(x) = \begin{cases} 0 & \text{pour } x < a \\ \dfrac{x-a}{b-a} & \text{pour } a \leq x \leq b \\ 1 & \text{pour } x > b \end{cases}$$

Preuve : On distingue trois cas :

— si $x < a$, $F(x) = \displaystyle\int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} 0dt = 0$,

— si $a \leq x \leq b$, $F(x) = \displaystyle\int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{a} 0dt + \int_{a}^{x} \frac{1}{b-a}dt = \frac{x-a}{b-a}$,

— si $x > b$, $F(x) = \displaystyle\int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{a} 0dt + \int_{a}^{b} \frac{1}{b-a}dt + \int_{b}^{x} 0dt = 0 + 1 + 0 = 1$.

■

2. Représentation graphique

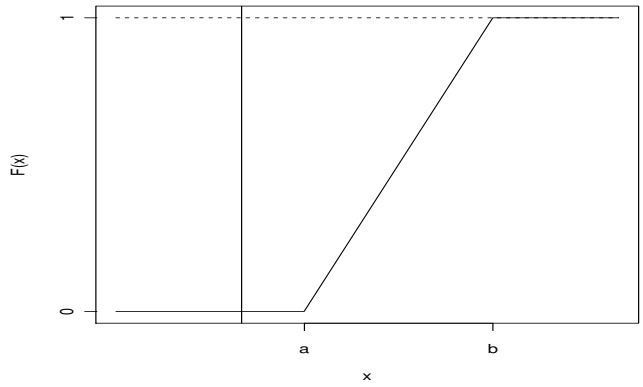

FIGURE 2.2 –

2.1.3 Moments

Si $X \rightsquigarrow \mathcal{U}([a;b])$ alors

- $E(X) = \displaystyle\int_{-\infty}^{+\infty} tf(t)dt = \int_{a}^{b} \frac{t}{b-a}dt = \left[\frac{t^2}{2(b-a)}\right]_a^b = \frac{a+b}{2}$.
- $V(X) = \dfrac{a^2+ab+b^2}{3} - \left(\dfrac{a+b}{2}\right)^2 = \dfrac{(b-a)^2}{12}$.

  En effet, on a $E(X^2) = \displaystyle\int_{a}^{b} \frac{t^2}{b-a}dt = \left[\frac{t^3}{3(b-a)}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}$.

Exemple 2.1.1 On considère la fonction f définie par :

$$f(x) = \begin{cases} 0 & \text{pour } x < 0 \text{ ou } x > 1 \\ 1 & \text{pour } 0 \leq x \leq 1 \end{cases}.$$

Il apparaît en intégrant que

$$F(x) = \begin{cases} 0 & \text{pour } x < 0 \\ x & \text{pour } 0 \leq x \leq 1 \\ 1 & \text{pour } x > 1 \end{cases}$$

On trouve $E(X) = \dfrac{1}{2}$ et $V(X) = \dfrac{1}{12}$.

Remarque 2.1.1 Dans l'exemple précédent, comme dans toute variable aléatoire absolument continue, on a $p(\{X = 0\}) = 0$. En effet, $\forall a \in \mathbb{R}$, $\int_a^a f(t)dt = 0$.

2.2 Loi exponentielle

2.2.1 Définition

Une variable aléatoire X suit une loi **exponentielle** de paramètre λ ($\lambda \in \mathbb{R}^{+\star}$) si X est une variable aléatoire absolument continue dont la densité de probabilité est définie par

$$f(x) = \begin{cases} 0 & \text{pour } x < 0 \\ \lambda e^{-\lambda x} & \text{pour } x \geq 0 \end{cases}$$

On note alors $X \rightsquigarrow \mathcal{E}(\lambda)$. La fonction f admet la représentation graphique de la Figure 2.3.

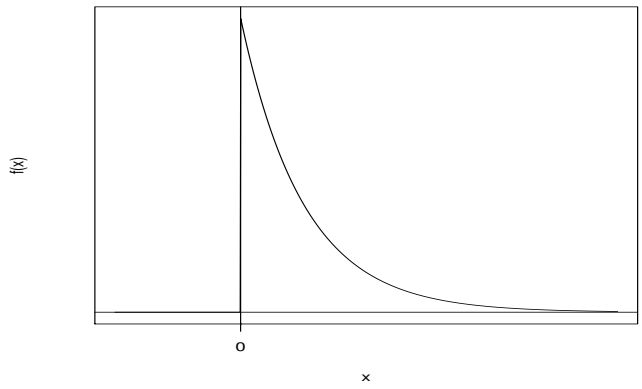

FIGURE 2.3 –

On a bien une densité de probabilité puisque

- $f(x) \geq 0 \ \forall x \in \mathbb{R}$,
- f est continue sur $\mathbb{R}^+$ et $\mathbb{R}^{-\star}$,
- $\int_{-\infty}^{+\infty} f(t)dt = \int_{-\infty}^{0} 0dt + \int_0^{+\infty} \lambda e^{-\lambda t}dt$. Or, si $A > 0$, $\int_0^A \lambda e^{-\lambda t}dt = [-e^{\lambda t}]_0^A = 1 - e^{-\lambda A} \to 1$ quand $A \to +\infty$ donc $\int_{-\infty}^{+\infty} f(t)dt = 1$.

La loi exponentielle peut être considérée comme l'équivalent en continu de la loi géométrique dans le cas discret. En effet, elle modélise un temps d'attente du premier succès dans un processus de Poisson.

2.2.2 Fonction de répartition

Si $X \rightsquigarrow \mathcal{E}(\lambda)$, on a

$$F(x) = \begin{cases} 0 & \text{pour } x < 0 \\ 1 - \lambda e^{-\lambda x} & \text{pour } x \geq 0 \end{cases}$$

dont la représentation graphique est donnée à la Figure 2.4.

2.2.3 Les moments

Si $X \rightsquigarrow \mathcal{E}(\lambda)$,

1. $E(X) = \int_{-\infty}^{+\infty} tf(t)dt = \int_0^{+\infty} \lambda t e^{-\lambda t}dt = \frac{1}{\lambda}$ à l'aide d'une intégration par parties.

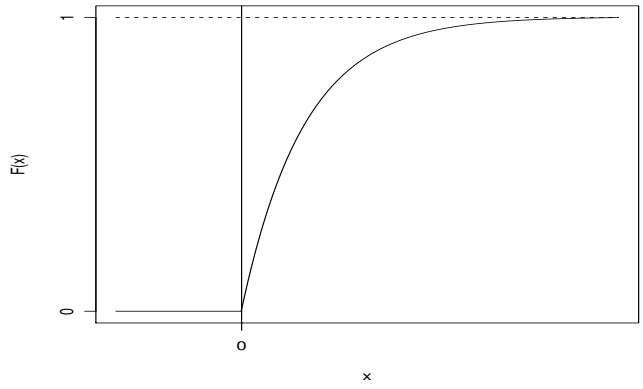

FIGURE 2.4 –

2. $V(X) = \dfrac{1}{\lambda^2}$ et $\sigma(X) = \dfrac{1}{\lambda}$.

   En effet, on sait que $V(X) = E(X^2) - (E(X))^2$. Comme $E(X^2) = \displaystyle\int_0^{+\infty} \lambda t^2 e^{-\lambda t} dt$, à l'aide de deux intégrations par parties, on obtient $E(X^2) = \dfrac{2}{\lambda^2}$.

2.3 Loi de Laplace-Gauss ou loi normale

2.3.1 Définition

On appelle **variable aléatoire normale** ou **gaussienne** toute variable aléatoire absolument continue dont la densité de probabilité f est définie par

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-m)^2}{2\sigma^2}}}$$

m étant une constante réelle, σ une constante réelle strictement positive. On utilise la notation suivante

$$\boxed{X \rightsquigarrow \mathcal{N}(m, \sigma)}$$

Remarque 2.3.1 On admettra que f est bien une densité de probabilité (la difficulté étant de montrer que $\displaystyle\int_{-\infty}^{+\infty} f(t)dt = 1$).

2.3.2 Représentation graphique

La courbe représentative de f est donnée par la Figure 2.5.

Remarque 2.3.2

— La courbe, dite courbe en cloche, a un axe de symétrie qui est la droite d'équation $x = m$.

— La densité f a un maximum atteint pour $x = m$ valant $\dfrac{1}{\sigma\sqrt{2\pi}}$.

— La courbe est d'autant plus pointue que σ est petit.

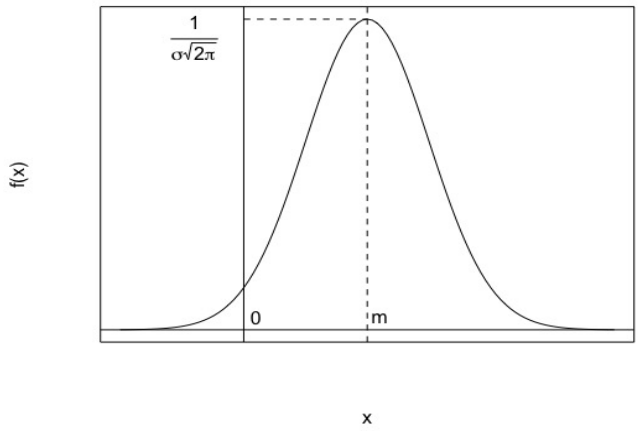

FIGURE 2.5 –

Rappels : Pour déterminer le(s) **point(s) d'inflexion** d'une fonction, on calcule sa dérivée seconde et on détermine le signe de cette dernière. Si le signe change pour une abscisse particulière, la fonction y admet un point d'inflexion.
Déterminons le(s) point(s) d'inflexion de f. Le calcul de la dérivée première donne :

$$f'(x) = \frac{1}{\sigma\sqrt{2\pi}}\frac{-2}{2\sigma^2}(x-m)e^{-\frac{(x-m)^2}{2\sigma^2}} = -\frac{1}{\sigma^3\sqrt{2\pi}}(x-m)e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Le calcul de la dérivée seconde donne :

$$f''(x) = -\frac{1}{\sigma^3\sqrt{2\pi}}\left[1-\frac{(x-m)^2}{\sigma^2}\right]e^{-\frac{(x-m)^2}{2\sigma^2}} = \frac{1}{\sigma^5\sqrt{2\pi}}[(x-m-\sigma)(x-m+\sigma)]e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

On a le tableau de signes suivant :

x	$-\infty$		$m-\sigma$		$m+\sigma$		$+\infty$
$f''(x)$		$+$		$-$		$+$	

Par conséquent, la courbe admet deux points d'inflexion pour $x = m+\sigma$ et $x = m-\sigma$.

Exemple 2.3.1 Les variables normales sont très fréquentes, par exemple la variable aléatoire réelle "poids" d'un français adulte, la variable aléatoire "quotient intellectuel" d'une population donnée.

2.3.3 Moments

L'espérance et la variance d'une variable normale sont respectivement données par :

$$\boxed{E(X) = m} \text{ et } \boxed{V(X) = \sigma^2}$$

2.3.4 Variable normale centrée réduite

Si $m = 0$ et $\sigma = 1$, la variable normale est appelée **variable normale centrée réduite** et est notée $\mathcal{Z}$ ou Γ, on note alors

$$\boxed{\mathcal{Z} \rightsquigarrow \mathcal{N}(0,1)}$$

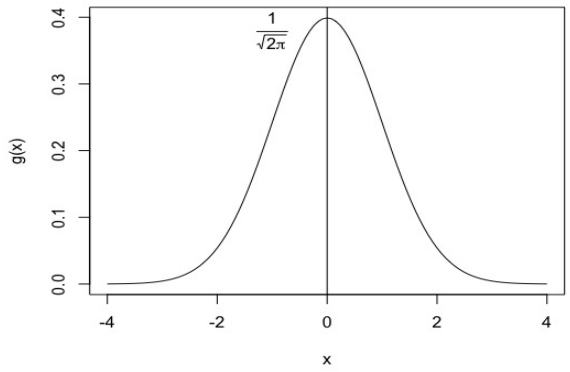

FIGURE 2.6 –

La courbe représentative de la fonction g est donnée par la Figure 2.6. Sa densité de probabilité est la fonction g définie par $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

- Cette fonction g est paire, la courbe a un axe de symétrie qui est la droite des ordonnées. En $x = 0$, la fonction g vaut $g(0) = \frac{1}{\sqrt{2\pi}}$.
- Les points d'inflexion de la fonction g se trouvent en $x = -1$ et $x = 1$.
- En général, les valeurs de $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ sont données à l'aide d'une table pour $x \geq 0$.

2.3.5 Fonction de répartition

Si on note Π la fonction de répartition de la loi normale centrée réduite $\mathcal{Z}$ associée à X,

$$\Pi(z) = \int_{-\infty}^{z} g(x)dx = p(\{\mathcal{Z} \leq z\})$$

cette fonction est représentée graphiquement à la Figure 2.7. Il existe des tables qui donnent la valeur de

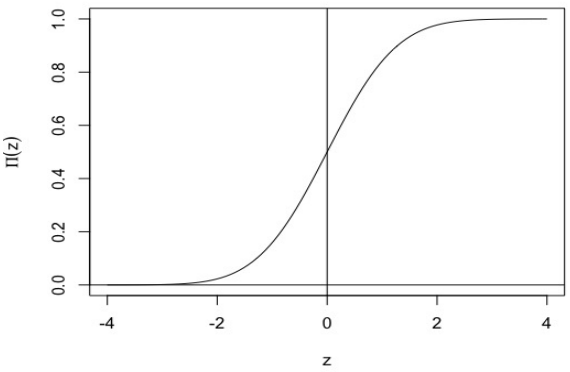

FIGURE 2.7 –

$\Pi(z)$ pour $z \geq 0$ (voir annexe A). $\Pi(z)$ désigne l'aire du domaine plan en jaune (voir Figure 2.8).

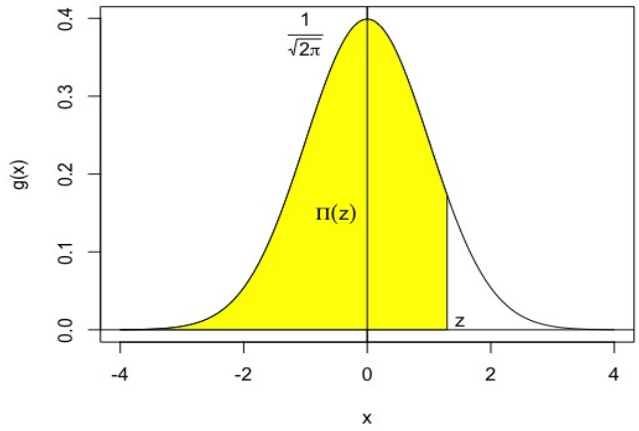

FIGURE 2.8 –

Exemple 2.3.2 $\Pi(0) = 0,5$ ce qui est évident puisqu'on considère exactement la moitié de l'aire totale (qui vaut 1). On peut également trouver à l'aide de la table $\Pi(1) = 0,8413$.

— Pour $z > 0$ on a la relation $\boxed{\Pi(z) + \Pi(-z) = 1}$

On peut observer cette propriété à l'aide de la Figure 2.9.

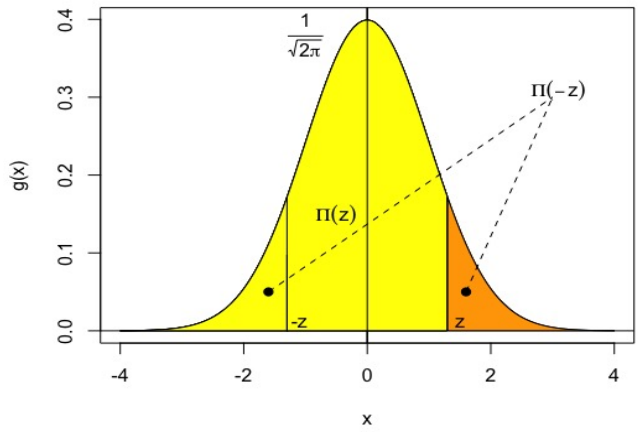

FIGURE 2.9 –

Exemple 2.3.3 $p(\{\mathcal{Z} \leq -1\}) = \Pi(-1) = 1 - \Pi(1) = 1 - 0,8413 = 0,1587$.

— Soit $z \in \mathbb{R}$, on a la relation $\boxed{p(\{\mathcal{Z} > z\}) = 1 - p(\{\mathcal{Z} \leq z\}) = 1 - \Pi(z)}$

On peut observer cette propriété à l'aide de la Figure 2.10.
Exemple 2.3.4 $p(\{Z > 1\}) = 1 - \Pi(1) = 0,1587$.

— Soient $a, b \in \mathbb{R}$ vérifiant $a < b$ alors $\boxed{p(\{a \leq \mathcal{Z} \leq b\}) = \Pi(b) - \Pi(a)}$

On peut observer cette propriété à l'aide de la Figure 2.11.
Exemple 2.3.5 $p(\{1 \leq \mathcal{Z} \leq 2\}) = \Pi(2) - \Pi(1) = 0,9772 - 0,8413 = 0,1359$.

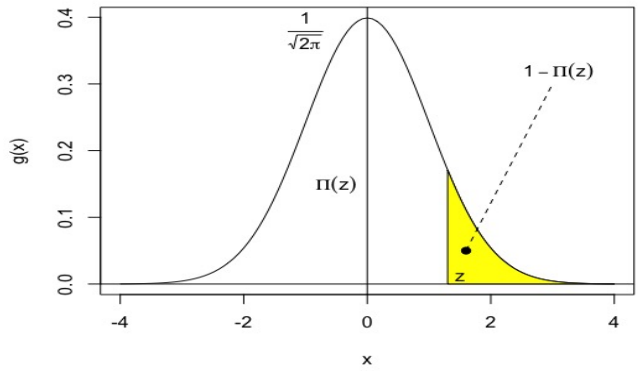

FIGURE 2.10 –

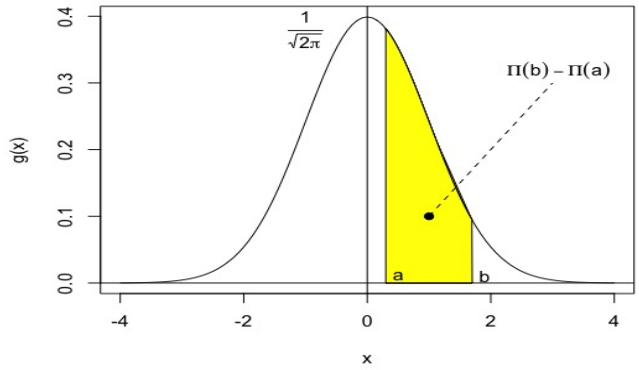

FIGURE 2.11 –

— Soit $Z_\alpha > 0$ alors $\boxed{p(\{-Z_\alpha \leq \mathcal{Z} \leq Z_\alpha\}) = \Pi(Z_\alpha) - \Pi(-Z_\alpha) = 2\Pi(Z_\alpha) - 1}$

On peut observer cette propriété à l'aide de la Figure 2.12.

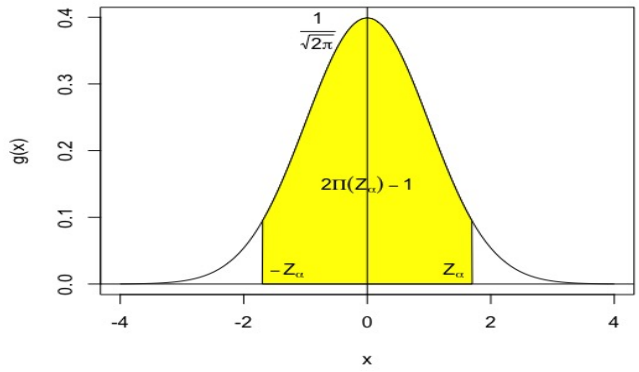

FIGURE 2.12 –

Exemple 2.3.6 $p(\{-1 \leq \mathcal{Z} \leq 1\}) = 2\Pi(1) - 1 = 2 \times 0,8413 - 1 = 0,6826.$

2.3.6 Table de l'écart réduit

Soit un intervalle centré en 0 de probabilité $1 - \alpha$, on note $-Z_\alpha$ et Z_α ses bornes. Alors

$$p(\{-Z_\alpha \leq z \leq Z_\alpha\}) = 1 - \alpha = 2\Pi(Z_\alpha) - 1$$

ou encore

$$\boxed{\Pi(Z_\alpha) = 1 - \frac{\alpha}{2}}$$

Par conséquent,

$$\boxed{p(\{z > Z_\alpha\}) = p(\{z < -Z_\alpha\}) = \frac{\alpha}{2}}$$

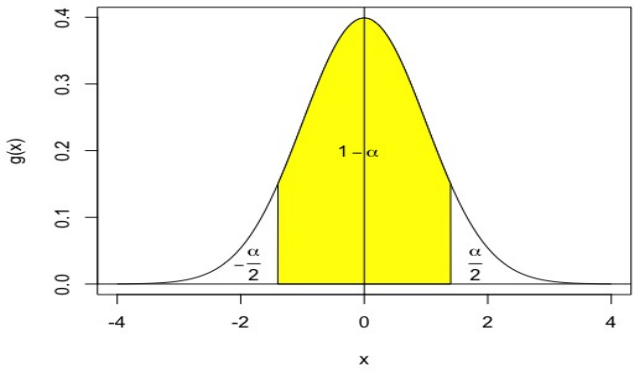

FIGURE 2.13 –

Il existe une table (ne figurant pas dans les annexes) qui pour α fixé donne Z_α, ce qui permet d'obtenir les bornes d'un intervalle centré en 0 dont la probabilité est connue.

Exemple 2.3.7 Pour $\alpha = 0,03$ on obtient $Z_\alpha = 2,170090$ ce qui signifie

- $p(\{-2,17090 \leq Z \leq 2,17090\}) = 0,97$
- $p(\{Z > 2,17090\}) = 0,015$
- $p(\{Z < -2.17090\}) = 0,015$

2.3.7 Exemples

1. Soient $X \rightsquigarrow \mathcal{N}(m = 30, \sigma = 3)$ et sa variable Z centrée réduite associée vérifiant $Z = \dfrac{X - 30}{3} \rightsquigarrow \mathcal{N}(0,1)$.
   Déterminons les probabilités $p(\{X = 28\})$, $p(\{X \leq 33\})$, $p(\{X \leq 27\})$, $p(\{27 \leq X \leq 33\})$ et $p(\{X > 33\})$.

   - $p(\{X = 28\}) = 0$.
   - $X \leq 33 \Leftrightarrow Z \leq \dfrac{33 - 30}{3} = 1$ donc $p(\{X \leq 33\}) = p(\{Z \leq 1\}) = \Pi(1) = 0,8413$.
   - $\{X \leq 27\} = \left\{Z \leq \dfrac{27 - 30}{3} = -1\right\}$ donc $p(\{X \leq 27\}) = p(\{Z \leq -1\}) = \Pi(-1) = 1 - \Pi(1) = 0,1587$.

- $\{27 \leq X \leq 33\} = \{-1 \leq Z \leq 1\}$ par conséquent $p(\{27 \leq X \leq 33\}) = p(\{-1 \leq Z \leq 1\}) = \Pi(1) - \Pi(-1) = 2\Pi(1) - 1$ ce qui donne $p(\{27 \leq X \leq 33\}) = 2 \times 0,8413 - 1 = 0,6826$.
- $p(\{X > 33\}) = 1 - p(\{X \leq 33\}) = 1 - \Pi(1) = 0,1587$.

2. Soit une variable aléatoire réelle dont on sait qu'elle suit une loi normale. On sait de plus que $p(\{X \leq 3\}) = 0,5517$ et $p(\{X > 7\}) = 0,0166$.
Déterminons les paramètres m et σ de la loi normale que suit X.

On a $X \rightsquigarrow \mathcal{N}(m, \sigma) \Leftrightarrow Z = \dfrac{X - m}{\sigma} \rightsquigarrow \mathcal{N}(0, 1)$ donc

- $X \leq 3 \Leftrightarrow Z \leq \dfrac{3 - m}{\sigma}$ et $p(\{X \leq 3\}) = p\left(\left\{Z \leq \dfrac{3 - m}{\sigma}\right\}\right) = 0,5517$. Alors $\Pi\left(\dfrac{3 - m}{\sigma}\right) = 0,5517$ or $\Pi(0,13) = 0,5517$ donc $\dfrac{3 - m}{\sigma} = 0,13$.
- $X > 7 \Leftrightarrow Z > \dfrac{7 - m}{\sigma}$ et $p(\{X \leq 7\}) = 0,9834 = p\left(\left\{Z \leq \dfrac{7 - m}{\sigma}\right\}\right)$. Ainsi $\Pi\left(\dfrac{7 - m}{\sigma}\right) = 0,9834$ or $\Pi(2,13) = 0,9834$ donc $\dfrac{7 - m}{\sigma} = 2,13$.

Afin de déterminer m et σ, on résout le système

$$\left\{ \begin{array}{rcl} 7 - m & = & 2,13\sigma \\ 3 - m & = & 0,13\sigma \end{array} \right. \Leftrightarrow \left\{ \begin{array}{rcl} 4 & = & 2\sigma \\ m & = & 3 - 0,13\sigma \end{array} \right. \Leftrightarrow \left\{ \begin{array}{rcl} \sigma & = & 2 \\ m & = & 3 - 0,26 = 2,74 \end{array} \right. .$$

Conclusion, $X \rightsquigarrow \mathcal{N}(m = 2,74; \sigma = 2)$.

2.3.8 Remarques

On sait que $X \rightsquigarrow \mathcal{N}(m, \sigma) \Leftrightarrow Z = \dfrac{X - m}{\sigma} \rightsquigarrow \mathcal{N}(0, 1)$ donc :

- $\{m - \sigma \leq X \leq m + \sigma\} = \{-1 \leq Z \leq 1\}$ et $p(\{m - \sigma \leq X \leq m + \sigma\}) = \Pi(1) - \Pi(-1) = 2\Pi(1) - 1 = 0,6826$. On peut alors affirmer que $68,26\%$ de la population étudiée appartient à l'intervalle $[m - \sigma; m + \sigma]$.
- $\{m - 2\sigma \leq X \leq m + 2\sigma\} = \{-2 \leq Z \leq 2\}$ et $p(\{m - 2\sigma \leq X \leq m + 2\sigma\}) = 2\Pi(2) - 1 = 0,9544$. On peut alors affirmer que $95,44\%$ de la population étudiée appartient à l'intervalle $[m - 2\sigma; m + 2\sigma]$.
- $\{m - 3\sigma \leq X \leq m + 3\sigma\} = \{-3 \leq Z \leq 3\}$ et $p(\{m - 3\sigma \leq X \leq m + 3\sigma\}) = 2\Pi(3) - 1 = 0,9973$. On peut alors affirmer que $99,73\%$ de la population étudiée appartient à l'intervalle $[m - 3\sigma; m + 3\sigma]$.

2.3.9 Relation entre la fonction de répartition et la densité de probabilité des loi normale et loi normale centrée réduite

— Soit $X \rightsquigarrow \mathcal{N}(m, \sigma)$ de densité f. La fonction de répartition F est définie par $F(x) = p(\{X \leq x\})$ et vérifie alors la relation

$$\boxed{F' = f}$$

— Soit $Z \rightsquigarrow \mathcal{N}(0, 1)$ de fonction de répartition Π et de densité g alors

$$\boxed{\Pi' = g}$$

— On a l'égalité $\{X \leq x\} = \left\{Z \leq \dfrac{x - m}{\sigma}\right\}$ ainsi que la relation $F(x) = \Pi\left(\dfrac{x - m}{\sigma}\right)$. Par dérivation,

$$F'(x) = \frac{1}{\sigma}\Pi'\left(\frac{x - m}{\sigma}\right) \Leftrightarrow f(x) = \frac{1}{\sigma} g\left(\frac{x - m}{\sigma}\right),$$

ce qui permet l'utilisation de la table de densité de probabilité de la loi $\mathcal{N}(0,1)$ pour calculer les valeurs de la densité de probabilité de $\mathcal{N}(m,\sigma)$.

2.3.10 Propriétés

Soit $X \rightsquigarrow \mathcal{N}(m,\sigma)$ et k une constante. On a les résultats suivants :

— la variable kX suit une loi normale $\mathcal{N}(km, |k|\sigma)$,

— la variable $k+X$ suit une loi normale $\mathcal{N}(k+m,\sigma)$.

2.3.11 Somme de deux variables normales indépendantes

Soient deux variables $X_1 \rightsquigarrow \mathcal{N}(m_1,\sigma_1)$ et $X_2 \rightsquigarrow \mathcal{N}(m_2,\sigma_2)$ indépendantes. Alors,

$$\boxed{X_1 + X_2 \rightsquigarrow \mathcal{N}(m_1+m_2, \sqrt{\sigma_1^2+\sigma_2^2})}$$

et

$$\boxed{X_1 - X_2 \rightsquigarrow \mathcal{N}(m_1-m_2, \sqrt{\sigma_1^2+\sigma_2^2})}$$

Plus généralement, soient n variables aléatoires indépendantes deux à deux telles que $X_i \rightsquigarrow \mathcal{N}(m_i,\sigma_i)$, $\forall i \in \{1,2,\ldots,n\}$, alors

$$\boxed{\text{la variable } X = \sum_{i=1}^{n} a_i X_i \text{ suit une loi normale } \mathcal{N}(m,\sigma)}$$

de moyenne $m = \sum_{i=1}^{n} a_i m_i$ et d'écart-type $\sigma = \sqrt{\sum_{i=1}^{n} a_i^2 \sigma_i^2}$.

Remarque 2.3.3 Si les X_i suivent la même loi $\mathcal{N}(m,\sigma)$,

— la variable $X = X_1 + X_2 + \ldots + X_n$ suit une loi normale de moyenne nm et d'écart-type $\sqrt{\sigma^2+\sigma^2+\ldots+\sigma^2} = \sqrt{n\sigma^2} = \sigma\sqrt{n}$.

— La variable $Y = \dfrac{X_1+X_2+\ldots+X_n}{n}$ suit une loi normale de moyenne $\dfrac{nm}{n} = m$, d'écart-type $\dfrac{\sigma}{\sqrt{n}}$.

2.3.12 Approximation d'une loi binomiale par une loi normale

Soit une loi binomiale $\mathcal{B}(n,p)$ de moyenne $m = np$, d'écart-type $\sigma = \sqrt{npq}$.
On montre que l'on peut approximer la loi binomiale $\mathcal{B}(n,p)$ par la loi normale $\mathcal{N}(m = np, \sigma = \sqrt{npq})$ si $n \geq 15$, p et q étant non voisins de 0. Dans la pratique, l'approximation est admise si $n \geq 20$, $np \geq 10$, $nq \geq 10$.

Exemple 2.3.8 Soit $X \rightsquigarrow \mathcal{B}(n = 100; p = 0,4)$ avec $E(X) = 40$ et $\sigma(X) = \sqrt{40 \times 0,6} = \sqrt{24}$. On a dans ce cas

$$\begin{cases} n = 100 \geq 20 \\ np = 40 \geq 10 \\ nq = 60 > 10 \\ npq = 24 \end{cases}$$

ce qui implique que $X \rightsquigarrow \mathcal{N}(m = 40, \sigma = \sqrt{24})$. Calculons $p(\{X < 38\})$.

— On a $p(\{X = 38\})_{\mathcal{B}} = p(\{37,5 \leq X \leq 38,5\})_{\mathcal{N}}$. Posons $Z = \dfrac{X-40}{\sqrt{24}}$ alors

$$\{37,5 \leq X \leq 38,5\} = \left\{ \frac{-2,5}{\sqrt{24}} \leq Z \leq \frac{-1,5}{\sqrt{24}} \right\} = \{0,510 \leq Z \leq 0,306\}.$$

Ainsi, $p(\{X = 38\})_{\mathcal{B}} = \Pi(-0,306) - \Pi(-0,510) = \Pi(-0,510) - \Pi(-0,306) = 0,6950 - 0,6202 = 0,0748$.

Montrons que $\pi(0,306) = 0,6202$ à l'aide de l'interpolation linéaire : on a $\pi(0,30) = 0,6179$ et $\pi(0,31) = 0,6217$ et le tableau suivant

$0,6179$	x	$0,6217$
$0,30$	$0,306$	$0,31$

$$\Leftrightarrow \frac{x - 0,6179}{0,6217 - 0,6179} = \frac{0,306 - 0,30}{0,31 - 0,30} \Leftrightarrow x = 0,6179 + 0,0038 \times \frac{0,006}{0,01} = 0,6202.$$

— Ensuite, $p(\{X > 38\})_{\mathcal{B}} = p(\{X > 38,5\})_{\mathcal{N}}$ et $\{X > 38,5\} = \left\{ Z > \frac{38,5 - 40}{\sqrt{24}} \right\}$. Par conséquent, $p(\{X > 38\})_{\mathcal{B}} = p(\{X > 38,5\})_{\mathcal{N}} = 1 - p(\{X \leq 38,5\})_{\mathcal{N}}$ ou encore $p(\{X > 38\})_{\mathcal{B}} = 1 - \Pi(-0,306) = \Pi(0,306) = 0,6202$.

— Enfin, $p(\{X \leq 38\})_{\mathcal{B}} = p\left(\{X \leq 38,5\}\right)_{\mathcal{N}}$ et $\{X \leq 38,5\} = \left\{ Z \leq \frac{40 - 38,5}{\sqrt{24}} \right\}$. Par conséquent, $p(\{X \leq 38\})_{\mathcal{B}} = p\left(\left\{ Z \leq \frac{-1,5}{\sqrt{24}} \right\} \right)_{\mathcal{B}} = \Pi(-0,306)$.

Finalement, $p(\{X \leq 38\})_{\mathcal{B}} = 1 - \Pi(0,306) = 1 - 0,6202 = 0,3797$.

2.3.13 Résumé sur les approximations de lois

- $\boxed{\mathcal{H}(N,p,n) \sim \mathcal{B}(n,p)}$ pour $N > 10n$,
- $\boxed{\mathcal{B}(n,p) \sim \mathcal{P}(\lambda = np)}$ pour $n \geq 30,\ p \leq 0,1$ et $np \leq 10$,
- $\boxed{\mathcal{B}(n,p) \sim \mathcal{N}(m = np, \sigma = \sqrt{npq})}$ avec $\left\{ \begin{array}{l} n \geq 20 \\ 0,1 < p < 0,9 \end{array} \right.$ ou $\left\{ \begin{array}{l} np \geq 10 \\ nq \geq 10 \end{array} \right.$ ou $npq > 3$.
- $\boxed{\mathcal{P}(\lambda = np) \sim \mathcal{N}(m = np, \sigma = \sqrt{npq})}$ pour $np \geq 10$.

2.4 Loi et variable du χ^2 (Khi-deux) de Pearson

2.4.1 Distribution du χ^2

1. On considère n variables indépendantes d'une loi normale centrée réduite T_1, T_2,...,T_n. La quantité

$$T_1^2 + T_2^2 + \ldots + T_n^2 = \sum_{i=1}^{n} T_i^2$$

est une variable aléatoire dont la distribution est celle d'un χ^2 **à** n **degrés de liberté** de moyenne et variance respectives,

$$\boxed{E(\chi_n^2) = n} \text{ et } \boxed{V(\chi_n^2) = 2n}$$

Lorsque n augmente, la densité f d'une loi du χ^2 ressemble de plus en plus à la densité d'une loi normale (voir la Figure 2.14) : La variable χ^2 est tabulée en fonction du nombre n de degrés de liberté. La table (voir annexe B) donne pour différentes valeurs de α, la valeur de x telle que :

$$P(\{\chi_n^2 < x\}) = 1 - \alpha$$

2. Graphiquement, cette valeur est égale à la surface grisée de la Figure 2.15 :

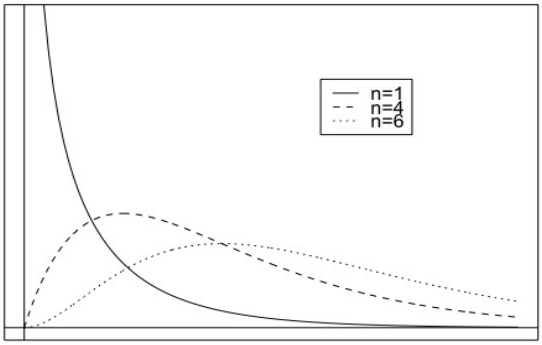

FIGURE 2.14 –

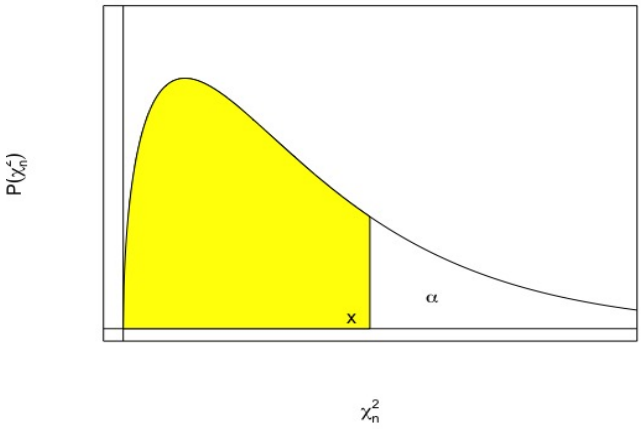

FIGURE 2.15 –

Exemple 2.4.1 Calculer $p(\{\chi^2_{10} > 20,5\})$. On récupère à l'aide de la table, la probabilité $p(\{\chi^2_{10} < 20,5\}) = 0,975$. Par conséquent, la probabilité recherchée $p(\{\chi^2_{10} > 20,5\})$ est égale à $1 - 0,975 = 0,025$.

Remarque 2.4.1 Attention, d'autres tables donnent la probabilité α, en fonction du nombre de degrés de liberté ν pour qu'une variable aléatoire X suivant une loi de χ^2_ν soit supérieure ou égale à une valeur donnée x : $\alpha = p(\{X \geq x\})$.

On a la propriété suivante :

$$\boxed{\chi^2_m + \chi^2_n = \chi^2_{m+n}}$$

Ce χ^2 admet

— une moyenne $E(\chi^2_{m+n}) = m + n$,

— une variance $\sigma^2(\chi^2_{m+n}) = 2(m+n)$

et ceci par application directe du théorème sur l'addition de variables aléatoires indépendantes.

2.5 Loi de Student-Fischer

2.5.1 Définition

La **loi de Student** est une loi continue qui comme la loi du χ^2 dépend d'un seul paramètre qu'on appellera également degré de liberté et qu'on note ν ($\nu \in \mathbb{N}^\star$). La variable X distribuée selon cette loi qu'on note

$$\boxed{X \rightsquigarrow t_\nu}$$

prend toutes ses valeurs dans $\mathbb{R}$. Si $Y \rightsquigarrow \mathcal{N}(0,1)$ et $Z \rightsquigarrow \chi^2_\nu$, Y et Z étant indépendantes, la variable $X = \dfrac{Y}{\sqrt{\dfrac{Z}{\nu}}}$ suit une loi de Student à ν degrés de liberté.

On dit qu'une variable aléatoire réelle à densité X a une loi de probabilité de Student à ν degrés de liberté (n entier > 0) si, et seulement si, sa densité de probabilité est donnée par la formule :

$$f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu}\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}}.$$

Dans cette formule, Γ est la fonction Gamma d'Euler définie, lorsque la partie réelle de x est positive, par :

$$\Gamma(x) = \int_0^{+\infty} e^u u^{x-1} du.$$

La loi de Student à ν degrés de liberté est la loi de probabilité du quotient d'une variable normale centrée réduite par la racine carrée de la somme des carrés de ν variables normales centrées réduites indépendantes entre elles et indépendantes de la première variable.

Pour $\nu = 1$, la loi de Student s'appelle **loi de Cauchy**, ou **loi de Lorentz**. C'est la loi du rapport de deux variables normales centrées réduites indépendantes.

2.5.2 Courbes

La courbe est unimodale, centrée, symétrique et plus plate que la courbe d'une loi normale. Lorsque le nombre de degrés de liberté augmente, la loi de Student tend vers la loi normale $\mathcal{N}(0,1)$ (voir Figure 2.16).

2.5.3 Moments

Soit $X \rightsquigarrow t_\nu$, on a

$$\boxed{E(X) = 0} \text{ et } \boxed{V(X) = \frac{\nu}{\nu-2}} \text{ pour } \nu > 2.$$

Remarque 2.5.1

— Lorsque l'espérance existe, elle est nulle, puisque la loi est symétrique autour de 0.

— Lorsque $\nu = 1$ ou $\nu = 2$, la variance n'est pas déterminée.

— Lorsque ν tend vers l'infini, la variance tend vers 1.

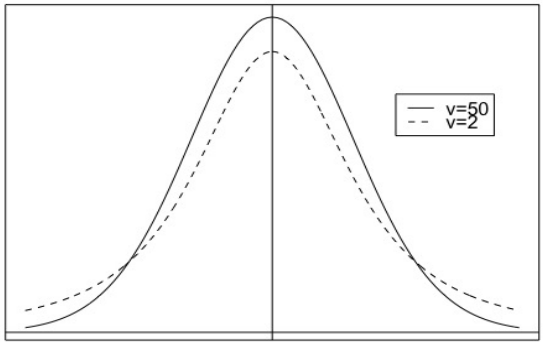

FIGURE 2.16 –

2.5.4 Tables

Soit $X \rightsquigarrow t_\nu$. Il existe une table (voir annexe C1) qui fournit les valeurs $t_{\nu,1-\alpha}$ pour ν et α donnés, telles que

$$p(\{X < t_{\nu,1-\alpha}\}) = 1 - \alpha.$$

Graphiquement, cette probabilité est donnée par la surface grisée de la Figure 2.17 :

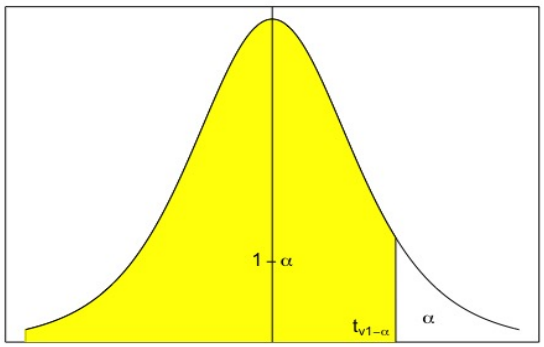

FIGURE 2.17 –

Exemple 2.5.1

- $\nu = 10$, $\alpha = 0,1$ et $X \rightsquigarrow t_{10}$ donc $p(\{X < 1,372\}) = 0,9$ et $t_{10;0,9} = 1,372$.
- $\nu = 20$, $\alpha = 0,05$ et $X \rightsquigarrow t_{20}$ donc $p(\{X < 1,725\}) = 0,95$ et $t_{20;0,95} = 1,725$.

Il existe une autre table (voir annexe C2) qui fournit pour ν et α donnés la valeur $t_{\nu,\alpha}$ telle que

$$p(\{-t_{\nu,\alpha} < X < t_{\nu,\alpha}\}) = 1 - \alpha.$$

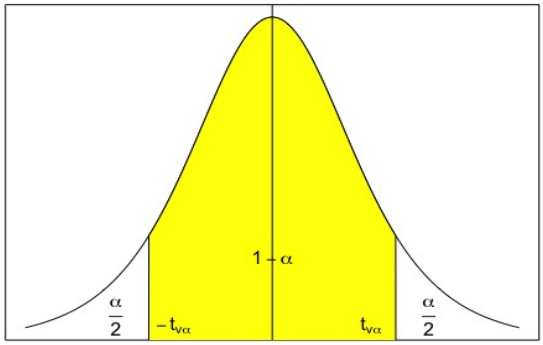

FIGURE 2.18 –

Graphiquement, cette probabilité est donnée par la surface grisée de la Figure 2.18 :
On remarque alors que $p(\{X < -t_{\nu,\alpha}\}) = p(\{X > t_{\nu,\alpha}\}) = \dfrac{1-(1-\alpha)}{2} = \dfrac{\alpha}{2}$.

Exemple 2.5.2 $\nu = 12$, $\alpha = 0,4$ et $t_{12;0,4} = 0,873$ donc $p(\{-0,873 < X < 0,873\}) = 0,6$ et $p(\{X < -0,873\}) = p(\{X > 0,873\}) = 0,2$.

2.6 Loi de Fischer-Snedecor

2.6.1 Définition

1. La **loi de Fischer-Snedecor** est une loi continue dépendant de deux paramètres notés ν_1 et ν_2, entiers naturels non nuls. La variable X distribuée selon cette loi prend toutes ses valeurs dans $\mathbb{R}^{+\star}$ ou dans $\mathbb{R}^{+}$.

   Si $Y \rightsquigarrow \chi^2_{\nu_1}$ et $Z \rightsquigarrow \chi^2_{\nu_2}$, Y et Z étant indépendantes, la variable $X = \dfrac{\dfrac{Y}{\nu_1}}{\dfrac{Z}{\nu_2}}$ suit une loi de Fischer-Snedecor. On note

   $$\boxed{X \rightsquigarrow F_{(\nu_1,\nu_2)}}$$

   La loi F de Fischer-Snedecor à (ν_1, ν_2) degrés de liberté est la loi de probabilité du rapport de deux variables de khi-deux indépendantes divisées par leurs nombres de degrés de liberté (ν_1 pour le numérateur, ν_2 pour le dénominateur).
   Pour $\nu_1 = 1$, la loi F de Fischer-Snedecor à $(1,\nu_2)$ degrés de liberté est la loi de probabilité du carré d'une variable de Student à ν_2 degrés de liberté.

2. La densité de probabilité est, par définition :

   $$f_{(\nu_1,\nu_2)}(x) = \nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}} \frac{\Gamma\left(\dfrac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\dfrac{\nu_1}{2}\right)\Gamma\left(\dfrac{\nu_2}{2}\right)} \frac{x^{\frac{\nu_1}{2}-1}}{(\nu_1 x+\nu_2)^{\frac{\nu_1+\nu_2}{2}}} \text{ pour } x > 0,\ \nu_1 \text{ et } \nu_2 \in \mathbb{N}^{\star}.$$

   Dans cette formule, Γ est la fonction Gamma d'Euler définie, lorsque la partie réelle de x est positive, par :

$$\Gamma(x) = \int_0^{+\infty} e^u u^{x-1} du.$$

La fonction $f_{(\nu_1,\nu_2)}$ est bien une densité de probabilité sur $]0;+\infty[$, car :

— ses valeurs sont positives,

— la fonction est intégrable et son intégrale est donnée par :

$$\int_0^{+\infty} f_{(\nu_1,\nu_2)}(x)dx = \nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \int_0^{+\infty} \frac{x^{\frac{\nu_1}{2}-1}}{(\nu_1 x+\nu_2)^{\frac{\nu_1+\nu_2}{2}}} dx.$$

Pour calculer l'intégrale $I = \displaystyle\int_0^{+\infty} \frac{x^{\frac{\nu_1}{2}-1}}{(\nu_1 x+\nu_2)^{\frac{\nu_1+\nu_2}{2}}} dx$, on pose $t = \dfrac{\nu_1 x}{\nu_1 x+\nu_2} \Rightarrow dx = \dfrac{n_2}{n_1}\dfrac{dt}{1-t^2}$. De plus, $\nu_1 x + \nu_2 = \nu_2 \times \dfrac{1}{1-t}$ ce qui implique que lorsque $x = 0$, $t = 0$ et lorsque x tend vers l'infini, t tend vers 1. Par conséquent,

$$I = \int_0^1 \left(\frac{\nu_2}{\nu_1}\frac{t}{1-t}\right)^{\frac{\nu_1}{2}-1} \left(\frac{1-t}{\nu_2}\right)^{\frac{\nu_1+\nu_2}{2}} \frac{\nu_2}{\nu_1}\frac{dt}{(1-t)^2} = \nu_1^{-\frac{\nu_1}{2}} \nu_2^{-\frac{\nu_2}{2}} \int_0^1 t^{\frac{\nu_1}{2}-1}(1-t)^{\frac{\nu_2}{2}-1} dt.$$

Dans l'intégrale, on reconnaît la fonction Beta d'Euler définie, lorsque les parties réelles de x et de y sont positives, par :

$$B(x,y) = \int_0^1 u^{x-1}(1-u)^{y-1} du = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

donc $\displaystyle\int_0^1 t^{\frac{\nu_1}{2}-1}(1-t)^{\frac{\nu_2}{2}-1} dt = B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)$ ce qui implique que

$$\int_0^{+\infty} f_{(\nu_1,\nu_2)}(x)dx = \nu_1^{\frac{\nu_1}{2}} \nu_2^{\frac{\nu_2}{2}} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \nu_1^{-\frac{\nu_1}{2}} \nu_2^{-\frac{\nu_2}{2}} \frac{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)} = 1.$$

L'intégrale de $f_{(\nu_1,\nu_2)}$ est bien égale à 1, ce qui montre que $f_{(\nu_1,\nu_2)}$ est bien une densité de probabilité.

3. Si $X \rightsquigarrow F(\nu_1,\nu_2)$ la variable $\dfrac{1}{X} \rightsquigarrow F(\nu_2,\nu_1)$ donc $F(\nu_1,\nu_2,1-\alpha) = \dfrac{1}{F(\nu_2,\nu_1,\alpha)}$.

2.6.2 Courbes

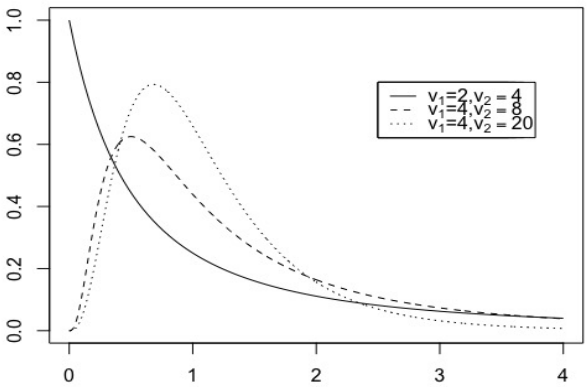

FIGURE 2.19 –

On a représenté ci-dessus (Figure 2.19) la loi F de Fischer-Snedecor pour diverses valeurs de ν_1 et de ν_2.

2.6.3 Moments

Soit $X \rightsquigarrow F(\nu_1, \nu_2)$.

— Pour $\nu_2 > 2$, l'espérance est définie par

$$\boxed{E(X) = \frac{\nu_2}{\nu_2 - 2}}$$

Remarque 2.6.1 Pour $\nu_2 \leq 2$, l'espérance n'est pas déterminée.

— Pour $\nu_2 > 4$, la variance est définie par

$$\boxed{V(X) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}}$$

Remarque 2.6.2 Pour $\nu_2 \leq 4$, la variance n'est pas déterminée.

2.6.4 Tables

Soit $X \rightsquigarrow F(\nu_1, \nu_2)$. La table (voir annexe D) fournit, pour $\alpha = 0,025$, pour ν_1 et ν_2 donnés, les valeurs $F_{\nu_1,\nu_2,1-\alpha}$ telles que $p(\{X < F_{\nu_1,\nu_2,1-\alpha}\}) = 1 - \alpha$. Cette table sert à la comparaison des variances de deux populations à partir de deux échantillons.

2.7 Exercices

Exercice 11 Une entreprise de transport a un parc total de 150 camions. On désigne par X la variable aléatoire qui à chaque camion choisi au hasard dans le parc, associe la distance qu'il a parcourue dans une journée (les distances sont mesurées en kilomètres). Un étude statistique permet d'admettre que cette variable aléatoire X suit une loi normale de moyenne 120 et d'écart-type 14.
Déterminer à 10^{-4} près la probabilité qu'un camion parcourt un jour donné une distance comprise entre 110 et 130 kilomètres (utiliser éventuellement une interpolation affine).

Exercice 12

1. *Statistique* - Avant d'accepter un contrat de livraison de véhicules, une société d'équipements automobiles établit une statistique de production journalière sur 100 jours. Le nombre de véhicules équipés journellement se répartit comme suit :

Production journalière de véhicules équipés	Nombre de jours	Production journalière de véhicules équipés	Nombre de jours
95	1	102	14
96	3	103	9
97	6	104	8
98	8	105	6
99	10	106	2
100	13	107	2
101	18	Total	100

Déterminer la valeur moyenne de la production journalière et une valeur approchée à 10^{-2} près de l'écart-type de cette production.

2. *Probabilités* - La production exigée par le contrat est de 100 véhicules équipés au moins par jour, pendant 100 jours de travail consécutif. À chaque journée on associe le nombre de véhicules équipés que l'on suppose indépendant du nombre obtenu chacun des autres jours. On définit ainsi une variable

aléatoire X. On admet que la variable aléatoire discrète X peut être approchée par la loi normale de paramètres $m = 101$ et $\sigma = 2.59$. On note Y une variable aléatoire suivant la loi $\mathcal{N}(101; 2, 59)$.
Calculer la probabilité de l'événement "le contrat est rempli", c'est-à-dire $p(\{Y \geq 99, 5\})$.

Exercice 13 On jette 10 fois de suite une pièce de monnaie bien équilibrée en notant chaque fois le résultat, ce qui constitue une partie.

1. On note X la variable aléatoire qui à chaque partie associe le nombre de "face" obtenu.
   (a) Justifier que la loi de probabilité suivie par la variable X est une loi binomiale (on précisera les paramètres de cette loi).
   (b) Calculer la probabilité de l'événement E :
   "le nombre de 'face' est compris entre 3 et 6 (bornes incluses)".
2. On décide d'approcher la loi de variable aléatoire discrète X par la loi normale de paramètres m et σ.
   (a) Expliquer pourquoi on prend $m = 5$ et $\sigma = \sqrt{2, 5}$.
   (b) On considère une variable aléatoire Y suivant une la loi $\mathcal{N}(5; \sqrt{2, 5})$. En utilisant cette approximation, calculer la probabilité de l'événement :
   "le nombre de 'face' est compris entre 3 et 6 (bornes incluses)"
   c'est-à-dire $p\{(2, 5 \leq Y \leq 6, 5\})$.

Exercice 14 Une entreprise fabrique des imprimantes de modèle PRINT et constate que le nombre de commandes journalières définit une variable aléatoire Y dont la loi peut être approchée par la loi normale de paramètres $m = 80$ et $\sigma = 60$. On désigne par Z la variable aléatoire qui, à chaque mois de 25 jours ouvrables, associe le nombre d'unités du modèle PRINT demandé. Il y a indépendance entre les commandes journalières.

1. Montrer que la loi de Z peut être approchée par la loi normale $\mathcal{N}(2000, 300)$.
   L'entreprise a en stock, au début du mois, 2300 unités. Quelle est la probabilité qu'elle ne puisse satisfaire à la demande ?
2. On veut que la probabilité qu'elle ne puisse satisfaire à la demande soit inférieure à $0, 05$. Quel doit être le nombre minimal d'unités que l'entreprise doit stocker en début de mois ?

Exercice 15 La variable aléatoire X suit une loi normale $\mathcal{N}(20, 5)$. Calculer

1. $p(\{X \leq 28\})$
2. $p(\{X \geq 28\})$
3. $p(\{X \geq 12\})$
4. $p(\{X \leq 12\})$
5. $p(\{12 \leq X \leq 28\})$

Exercice 16 Pour mesurer l'impact d'un régime amaigrissant, un club a choisi au hasard un échantillon de 5 individus avant le régime, et un échantillon de 5 autres individus après. Les masses corporelles se présentent ainsi :

Avant	84	92	72	91	84
Après	81	88	74	81	90

1. Déterminer un intervalle de confiance à 95% pour :
   (a) la masse corporelle moyenne avant le régime,

(b) la masse corporelle moyenne après le régime,

(c) la perte moyenne de masse corporelle durant le régime.

2. Tout compte fait, on a décidé qu'il aurait peut-être été plus adapté de peser les mêmes individus avant et après le régime. On a obtenu :

Individu	1	2	3	4	5
Avant	84	92	72	91	84
Après	81	88	74	81	90

Sur la base de cet échantillon, déterminer un intervalle de confiance à 95% pour la perte moyenne de masse corporelle durant le régime. Conclusion ?

Exercice 17 Un laboratoire veut fabriquer des pilules se composant de deux substances A et B. Pour chaque pilule de la fabrication, on considère les masses a et b respectivement des 2 substances A et B qui la constituent. On désigne par X et Y respectivement les variables aléatoires qui associent à chaque pilule la masse a et la masse b des substances de cette pilule. On suppose que ces variables sont indépendantes et suivent des lois normales de moyennes respectives $m_X = 8,55$ mg et $m_Y = 5,20$ mg et de même écart-type $\sigma_X = \sigma_Y = 0,05$ mg.

1. Déterminer les probabilités $p(\{8,45 \leq X \leq 8,70\})$ et $p(\{5,07 \leq Y \leq 5,33\})$.
2. Les normes imposées pour la fabrication sont les suivantes : $8,45 \leq a \leq 8,70$ et $5,07 \leq b \leq 5,33$
   (a) Calculer le pourcentage de pilules qui seront hors normes à la sortie de la chaîne de fabrication.
   (b) En déduire que le procédé de fabrication ne peut être retenu si on veut que le pourcentage de pilules défectueuses ne dépasse pas 3%. On modifie alors la fabrication de la substance B. La moyenne de Y ne change pas mais son écart-type est modifié. Trouver la valeur minimum de ce nouvel écart-type pour que le pourcentage de pièces défecteuses soit inférieur à 3%.
3. (a) Déterminer la moyenne et l'écart-type de la variable aléatoire S qui associe à chaque pilule sa masse totale, les variables X et Y gardant leurs caractéristiques de la question 1.
   (b) On admet que S est encore une variable aléatoire normale dont les paramètres sont ceux calculés précédemment. Calculer $p(\{13,6 \leq S \leq 13,8\})$.
4. On assure le conditionnement des pilules par boîtes de 100 unités. Une boîte est constituée à partir d'un tirage au hasard dans un stock assez grand pour qu'on puisse estimer que les tirages successifs se font avec remises. On désigne par Z la variable aléatoire qui, à chaque boîte associe le nombre de pilules hors normes au sens de la question 2.(a). On pourra prendre pour probabilité p d'une pilule hors-norme $p = 0,01$.
   (a) Dans ces conditions, montrer que Z est une variable binomiale dont on précisera les paramètres.
   (b) Dire pourquoi on peut approcher cette variable par une loi de Poisson. En utilisant cette loi, donner une valeur approximative de $p(\{Z \geq 5\})$.
5. On désigne par U la variable aléatoire qui à chaque boîte associe le nombre de pilules dont la masse totale est supérieure à $13,8$. Là aussi, on peut supposer que U est une variable binomiale de paramètres n et p.
   (a) Calculer p.
   (b) Dire pourquoi on peut approcher U par une variable normale. À l'aide de cette approximation, donner une valeur approchée de $p(\{U \in \{70, 71, \ldots, 85\}\})$.

ANNEXE A - Fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$.

Cette table donne $\Pi(x) = p(\{X \leq x\})$ pour $X \rightsquigarrow \mathcal{N}(0,1)$:

x	$0,00$	$0,01$	$0,02$	$0,03$	$0,04$	$0,05$	$0,06$	$0,07$	$0,08$	$0,09$
$0,0$	$0,5000$	$0,5040$	$0,5080$	$0,5120$	$0,5160$	$0,5199$	$0,5239$	$0,5279$	$0,5319$	$0,5359$
$0,1$	$0,5398$	$0,5438$	$0,5478$	$0,5517$	$0,5557$	$0,5596$	$0,5636$	$0,5675$	$0,5714$	$0,5753$
$0,2$	$0,5793$	$0,5832$	$0,5871$	$0,5910$	$0,5948$	$0,5987$	$0,6026$	$0,6064$	$0,6103$	$0,6141$
$0,3$	$0,6179$	$0,6217$	$0,6255$	$0,6293$	$0,6331$	$0,6368$	$0,6406$	$0,6443$	$0,6480$	$0,6517$
$0,4$	$0,6554$	$0,6591$	$0,6628$	$0,6664$	$0,6700$	$0,6736$	$0,6772$	$0,6808$	$0,6844$	$0,6879$
$0,5$	$0,6915$	$0,6950$	$0,6985$	$0,7019$	$0,7054$	$0,7088$	$0,7123$	$0,7157$	$0,7190$	$0,7224$
$0,6$	$0,7257$	$0,7291$	$0,7324$	$0,7357$	$0,7389$	$0,7422$	$0,7454$	$0,7486$	$0,7517$	$0,7549$
$0,7$	$0,7580$	$0,7611$	$0,7642$	$0,7673$	$0,7704$	$0,7734$	$0,7764$	$0,7794$	$0,7823$	$0,7852$
$0,8$	$0,7881$	$0,7910$	$0,7939$	$0,7967$	$0,7995$	$0,8023$	$0,8051$	$0,8078$	$0,8106$	$0,8133$
$0,9$	$0,8159$	$0,8186$	$0,8212$	$0,8238$	$0,8264$	$0,8289$	$0,8315$	$0,8340$	$0,8365$	$0,8389$
$1,0$	$0,8413$	$0,8438$	$0,8461$	$0,8485$	$0,8508$	$0,8531$	$0,8554$	$0,8577$	$0,8599$	$0,8621$
$1,1$	$0,8643$	$0,8665$	$0,8686$	$0,8708$	$0,8729$	$0,8749$	$0,8770$	$0,8790$	$0,8810$	$0,8830$
$1,2$	$0,8849$	$0,8869$	$0,8888$	$0,8907$	$0,8925$	$0,8944$	$0,8962$	$0,8980$	$0,8997$	$0,9015$
$1,3$	$0,9032$	$0,9049$	$0,9066$	$0,9082$	$0,9099$	$0,9115$	$0,9131$	$0,9147$	$0,9162$	$0,9177$
$1,4$	$0,9192$	$0,9207$	$0,9222$	$0,9236$	$0,9251$	$0,9265$	$0,9279$	$0,9292$	$0,9306$	$0,9319$
$1,5$	$0,9332$	$0,9345$	$0,9357$	$0,9370$	$0,9382$	$0,9394$	$0,9406$	$0,9418$	$0,9429$	$0,9441$
$1,6$	$0,9452$	$0,9463$	$0,9474$	$0,9484$	$0,9495$	$0,9505$	$0,9515$	$0,9525$	$0,9535$	$0,9545$
$1,7$	$0,9554$	$0,9564$	$0,9573$	$0,9582$	$0,9591$	$0,9599$	$0,9608$	$0,9616$	$0,9625$	$0,9633$
$1,8$	$0,9641$	$0,9649$	$0,9656$	$0,9664$	$0,9671$	$0,9678$	$0,9686$	$0,9693$	$0,9699$	$0,9706$
$1,9$	$0,9713$	$0,9719$	$0,9726$	$0,9732$	$0,9738$	$0,9744$	$0,9750$	$0,9756$	$0,9761$	$0,9767$
$2,0$	$0,9772$	$0,9778$	$0,9783$	$0,9788$	$0,9793$	$0,9798$	$0,9803$	$0,9808$	$0,9812$	$0,9817$
$2,1$	$0,9821$	$0,9826$	$0,9830$	$0,9834$	$0,9838$	$0,9842$	$0,9846$	$0,9850$	$0,9854$	$0,9857$
$2,2$	$0,9861$	$0,9864$	$0,9868$	$0,9871$	$0,9875$	$0,9878$	$0,9881$	$0,9884$	$0,9887$	$0,9890$
$2,3$	$0,9893$	$0,9896$	$0,9898$	$0,9901$	$0,9904$	$0,9906$	$0,9909$	$0,9911$	$0,9913$	$0,9916$
$2,4$	$0,9918$	$0,9920$	$0,9922$	$0,9925$	$0,9927$	$0,9929$	$0,9931$	$0,9932$	$0,9934$	$0,9936$
$2,5$	$0,9938$	$0,9940$	$0,9941$	$0,9943$	$0,9945$	$0,9946$	$0,9948$	$0,9949$	$0,9951$	$0,9952$
$2,6$	$0,9953$	$0,9955$	$0,9956$	$0,9957$	$0,9959$	$0,9960$	$0,9961$	$0,9962$	$0,9963$	$0,9964$
$2,7$	$0,9965$	$0,9966$	$0,9967$	$0,9968$	$0,9969$	$0,9970$	$0,9971$	$0,9972$	$0,9973$	$0,9974$
$2,8$	$0,9974$	$0,9975$	$0,9976$	$0,9977$	$0,9977$	$0,9978$	$0,9979$	$0,9979$	$0,9980$	$0,9981$
$2,9$	$0,9981$	$0,9982$	$0,9982$	$0,9983$	$0,9984$	$0,9984$	$0,9985$	$0,9985$	$0,9986$	$0,9986$
$3,0$	$0,9987$	$0,9987$	$0,9987$	$0,9988$	$0,9988$	$0,9989$	$0,9989$	$0,9989$	$0,9990$	$0,9990$
$3,1$	$0,9990$	$0,9991$	$0,9991$	$0,9991$	$0,9992$	$0,9992$	$0,9992$	$0,9992$	$0,9993$	$0,9993$
$3,2$	$0,9993$	$0,9993$	$0,9994$	$0,9994$	$0,9994$	$0,9994$	$0,9994$	$0,9995$	$0,9995$	$0,9995$
$3,3$	$0,9995$	$0,9995$	$0,9995$	$0,9996$	$0,9996$	$0,9996$	$0,9996$	$0,9996$	$0,9996$	$0,9997$
$3,4$	$0,9997$	$0,9997$	$0,9997$	$0,9997$	$0,9997$	$0,9997$	$0,9997$	$0,9997$	$0,9997$	$0,9998$

ANNEXE B - Probabilités individuelles de la loi du χ^2_ν,

Cette table donne les valeurs (quantiles) $\chi^2_{\nu,1-\alpha}$ telles que $p(\{\chi^2_\nu < \chi^2_{\nu,1-\alpha}\}) = 1-\alpha$:

ν \ $1-\alpha$	0,005	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990	0,995
1	0,0000393	0,000157	0,000982	0,00393	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,60
3	0,072	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,064	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	1,61	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,989	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,96
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73	26,76
12	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,46	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	85,53	90,53	95,02	100,4	104,2
80	51,17	53,54	57,15	60,39	64,28	96,58	101,9	106,6	112,3	116,3
90	59,20	61,75	65,65	69,13	73,29	107,6	113,1	118,1	124,1	128,3
100	67,33	70,06	74,22	77,93	82,36	118,5	124,3	129,6	135,8	140,2

ANNEXE C1 - Probabilités individuelles et cumulées de la loi de Student-Fischer $t_{\nu,\alpha}$,

Cette table donne les valeurs (quantiles) $t_{\nu,1-\alpha}$ telles que $p(\{-t_\nu < t_{\nu,1-\alpha}\}) = 1-\alpha$:

ν \ $1-\alpha$	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	0,975	0,99	0,995	0,9995
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	635,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,150
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,648
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
80	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

ANNEXE C2 - Probabilités individuelles et cumulées de la loi de Student-Fischer $t_{\nu,\alpha}$.

Cette table donne les valeurs $t_{\nu,\alpha}$ telles que $p(\{t_{\nu,\alpha} < t_\nu < +t_{\nu,\alpha}\}) = 1 - \alpha$:

ν \ α	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,387	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,745	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,649
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,656
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
80	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,l26	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,940	2,326	2,576	3,291

ANNEXE D - la loi de Fischer-Snedecor,

Cette table donne, pour $\alpha = 0,025$, pour ν_1 et ν_2 donnés, les valeurs $F_{\nu_1,\nu_2,1-\alpha}$ telles que $p(\{X < F_{\nu_1,\nu_2,1-\alpha}\}) = 1 - \alpha$,

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6	976,7	984,9	993,1	997,2	1001	1006	1010	1014	1018
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	5,90	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

Chapitre 3

Échantillonnage et estimation

3.1 Introduction

La **théorie de l'échantillonnage** étudie les liens entre une population et des échantillons de cette population. À partir d'informations relatives à la loi d'une variable X pour une population donnée, on en déduit le comportement d'échantillons aléatoires simples relatifs à cette variable.
Dans la pratique c'est le problème inverse qui se pose. En général on ne connaît pas la loi de X, on ne connaît pas tous ses paramètres et on souhaite obtenir des informations à partir de l'observation d'un échantillon. Ce problème fait partie de la **théorie de l'estimation**.

Souvent on s'intéresse à la valeur d'un paramètre bien précis de la loi de X, espérance, variance, proportion. Ce paramètre noté θ est appelé **paramètre d'intérêt**, c'est un nombre dont la valeur est inconnue. On cherche à évaluer ce nombre à partir de l'observation d'un échantillon. À partir des données de l'observation d'un échantillon, on détermine une valeur numérique $\hat{\theta}$ qu'on appelle **estimation ponctuelle du paramètre d'intérêt**.

On peut aussi définir un **intervalle de confiance** c'est-à-dire déterminer un intervalle $[\theta_1; \theta_2]$ qui contient la vraie valeur du paramètre θ inconnu avec une grande probabilité fixée à priori.

Exemple 3.1.1 On veut estimer l'espérance mathématique de la variable X, "note des étudiants à l'examen", vérifiant $X \rightsquigarrow \mathcal{N}(m, \sigma)$. On prélève 50 copies dans la population, on les corrige, on obtient 50 notes x_1, x_2,...,x_{50} et on détermine la moyenne de cet échantillon $\overline{x} = \dfrac{x_1 + x_2 + \ldots + x_{50}}{50}$, on obtient 9,1. Intuitivement on peut estimer m par 9,1. On dit que $9,1$ est une estimation ponctuelle de m. On remarque que si on avait pris un autre échantillon, l'estimation serait différente. On pourrait aussi conclure que la moyenne m appartiendrait à l'intervalle $[8,4; 9,8]$ avec une probabilité de 0,9 par exemple. L'intervalle $[8,4; 9,8]$ est alors un intervalle de confiance au risque d'erreur 0,1.

3.2 Estimation ponctuelle

3.2.1 Introduction

L'ensemble des hypothèses relatives au problème d'estimation de paramètre est appelé **modèle statistique**. Celui-ci comprend :

— des hypothèses relatives à la loi de la variable X, par exemple $X \rightsquigarrow \mathcal{N}(m, \sigma)$, m et σ étant inconnus, ou X suit une loi inconnue. Le paramètre θ doit être défini, par exemple $\theta = E(X)$, $\theta = \sigma(X)$, $\theta = p$. On écrira $X \rightsquigarrow l(x, \theta)$ où x est la réalisation de X.

— La méthode de construction de l'échantillon doit être précisée, échantillon aléatoire simple par exemple. On n'utilisera dans ce cours que des échantillons aléatoires simples.

Rappel sur le choix d'un échantillon : Les échantillons étudiés sont tous aléatoires, le hasard intervient dans le choix de leurs éléments. Cependant deux procédures sont possibles pour construire un échantillon aléatoire :

— échantillon non exhaustif : pour construire un échantillon de taille n, on procède par n tirages au hasard avec remise (remise de l'individu dans la population après chaque tirage),

— échantillon exhaustif : pour construire un échantillon de taille n, on procède par n tirages au hasard sans remise ou par le tirage simultané de n individus.

Si la population est très grande, on peut considérer un échantillon exhaustif comme non exhaustif.

Rappel sur les échantillons aléatoire simples : On considère l'exemple suivant.

Exemple 3.2.1 Considérons un économiste chargé de réaliser un étude de marché pour une entreprise qui souhaite lancer une nouvelle marque de fromage. Il commence par analyser la consommation de fromage en France. Il doit réaliser un sondage et demander aux personnes interrogées combien de fois elles ont consommé de fromage la semaine dernière. La consommation de fromage est extrêmement variable et incertaine. Certaines n'en mangent jamais, d'autres en mangent plusieurs fois par jour. On a donc un grand nombre de réalisations possibles. À chacune de ces réalisations potentielles est associée une probabilité, la consommation hebdomadaire de fromage est donc une variable aléatoire. Notons X la quantité consommée et plus précisément le nombre de fois par semaine qu'un individu mange du fromage. Cette variable X a une distribution de probabilité, une loi qu'on note $l(x)$. L'espérance et la variance de X sont deux paramètres de cette loi. $X \rightsquigarrow l(x, m, \sigma)$ avec $m = E(X)$ et $\sigma = \sigma(X)$. À priori, la loi de X, m et σ sont inconnus. Considérons un prélévement au hasard de n individus avec remise dans la population. Observer les quantités consommées de fromage pour ces n individus revient à observer la réalisation de la variable X pour ces n individus, choisis au hasard, indépendamment les uns des autres et avec remise. Les consommations de ces n individus peuvent être considérées comme n variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes et de même loi que X c'est-à-dire $l(x, m, \sigma)$.

Les n variables aléatoires indépendantes $X_1, X_2, \ldots, X_n$ constituent un échantillon aléatoire simple de la variable X si et seulement si

$$E(X_1) = E(X_2) = \ldots = E(X_n) = E(X) = m,$$
$$\sigma(X_1) = \sigma(X_2) = \ldots = \sigma(X_n) = \sigma(X) = \sigma.$$

Une fois ces n personnes interrogées, on dispose de n valeurs numériques des quantités consommées. On appelle ces n valeurs numériques observations ou encore réalisations. Ce sont des nombres réels qu'on notera $x_1, x_2, \ldots, x_n$.

On considère donc un modèle statistique $X \rightsquigarrow l(x, \theta)$ et un échantillon aléatoire simple $X_1, X_2, \ldots, X_n$. On recherche une statistique fonction des variables $X_1, X_2, \ldots, X_n$ susceptible de fournir la meilleure estimation possible du paramètre d'intérêt. Cette statistique est appelée **estimateur**.

Population	**Échantillon aléatoire simple**	**Observations**
$X \rightsquigarrow l(x, m)$	$X_1, X_2, \ldots, X_n$ estimateur $\overline{X}$ (par exemple)	$x_1, x_2, \ldots, x_n$ $\overline{x}$ estimation ponctuelle de m

Remarque 3.2.1 Dans le cas de la variable "note", on pourrait prendre comme estimation de m :

$$\frac{x_1 + x_{50}}{2}, \frac{x_1 + x_3 + x_5 + \ldots + x_{49}}{25}, \frac{x_2 + x_4 + \ldots + x_{50}}{25}, \ldots.$$

Dès lors le problème est celui du choix d'un estimateur. Comment va t-on décider quelle statistique utiliser en fonction du paramètre θ recherché ?

3.2.2 Estimateur sans biais

Définition 3.2.1 *Soit $X \rightsquigarrow l(x,\theta)$ un modèle statistique et soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire simple de X. On appelle **estimateur sans biais** du paramètre θ toute statistique $T = T(X_1, X_2, \ldots, X_n)$ telle que $E(T) = \theta$.*

Définition 3.2.2 *Si $E(T) \neq \theta$, T est **biaisé** et le biais vaut $E(T-\theta) = E(T) - \theta$.*

Considérons différentes statistiques ainsi que des tirages non exhaustifs (les tirages ont lieu avec remise) :

1. Prenons l'exemple de la statistique moyenne échantillon.
   Supposons que nous nous intéressons par exemple à l'espérance de la consommation hebdomadaire de fromage. On constitue un échantillon aléatoire simple en tirant au hasard n personnes de la population. Un enquêteur les interroge et obtient les réalisations numériques $x_1, x_2, \ldots, x_n$ des n variables aléatoires $X_1, X_2, \ldots, X_n$. La variable aléatoire "consommation moyenne" est $\overline{X} = \dfrac{\sum_{i=1}^n X_i}{n} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ et la réalisation de la variable $\overline{X}$ est $\overline{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$.
   On remarquera que la consommation moyenne $\overline{x}$ de l'échantillon varie en fonction de l'échantillon, c'est-à-dire que pour des échantillons différents, on obtient des moyennes d'échantillons différentes.

   Définition 3.2.3 *La variable*

   $$\boxed{\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}}$$

   est appelée variable moyenne échantillon.

   Si l'on considère par exemple 20 échantillons de taille n, on obtient la moyenne $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_{20}$ de chacun de ces échantillons. On peut s'attendre à ce que ces 20 valeurs soient proches de l'espérance m de la consommation hebdomadaire.
   — Espérance de la variable moyenne : soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire simple relatif à la variable X. Pour $i = 1, 2, \ldots, n$ on a $E(X_i) = E(X) = m$. Donc $E(\overline{X}) = E\left(\dfrac{\sum_{i=1}^n X_i}{n}\right) = \dfrac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \dfrac{1}{n}\sum_{i=1}^n E(X_i) = \dfrac{nm}{n} = m$.
   — Variance de la variable moyenne : pour i variant de 1 à n, $V(X_i) = V(X)$, les variables $X_1, X_2, \ldots, X_n$ sont indépendantes donc $V(\overline{X}) = V\left(\dfrac{\sum_{i=1}^n X_i}{n}\right) = \dfrac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) = \dfrac{1}{n^2}\sum_{i=1}^n V(X_i) = \dfrac{nV(X)}{n^2} = \dfrac{V(X)}{n} \Leftrightarrow \sigma(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$.

   Proposition 3.2.1 $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ *est un estimateur sans biais de $\theta = m$ car $E(\overline{X}) = m$.*

2. Prenons l'exemple de la statistique variance échantillon.
   Reprenons l'exemple sur la consommation de fromage. La variabilité des comportements individuels de la consommation est mesurée par l'écart-type σ de la consommation X. On considère un n-échantillon aléatoire simple $X_1, X_2, \ldots, X_n$ de X et la statistique

   $$\boxed{\Sigma^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\overline{X})^2}$$

   La réalisation de cette variable aléatoire Σ^2 est la variance de l'échantillon, notée

$$\sigma'^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2.$$

Déterminons l'espérance de la variable variance : on a $E(\Sigma^2) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - E\left(\overline{X}^2\right)$. Donc, $E(\Sigma^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E\left(\overline{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left(V(X_i) + E(X_i)^2\right) - (V(\overline{X}) + E(\overline{X})^2)$. En utilisant les formules précédentes relatives à la variable moyenne échantillon, on trouve $E(\Sigma^2) = \frac{1}{n}\sum_{i=1}^{n}\left(V(X) + E(X)^2\right) - \left(\frac{V(X)}{n} + E(X)^2\right) = V(X) - \frac{V(X)}{n} = \left(1 - \frac{1}{n}\right)\sigma^2$.

Proposition 3.2.2 $\Sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$ *est un estimateur biaisé de* $\theta = \sigma^2$ *car* $E(\Sigma^2) = \left(1 - \frac{1}{n}\right)\sigma^2$, *le biais est alors* $E(\Sigma^2) - \sigma^2 = -\frac{\sigma^2}{n} < 0$

On remarque que l'espérance de Σ^2 ne donne pas une image parfaite de σ^2, variance de X dans la population. Elle est systématiquement plus petite que σ^2 et lorsque n tend vers $+\infty$, $E(\Sigma^2)$ tend vers σ^2. Pour remédier à cet inconvénient de la la statistique Σ^2, on introduit la statistique S^2.

3. Prenons l'exemple de la statistique S^2.
Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire simple de la variable X. On définit la variable

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

La réalisation de S^2 est notée

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

On remarque alors que $n\Sigma^2 = (n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2$ et $n\sigma'^2 = (n-1)s^2 = \sum_{i=1}^{n} x_i^2 - n(\overline{x})^2$.

Déterminons l'espérance de la variable S^2 : $E(S^2) = E\left(\frac{n}{n-1}\Sigma^2\right) = \frac{n}{n-1}E(\Sigma^2) = \frac{n}{n-1}\left(1 - \frac{1}{n}\right)\sigma^2 = \sigma^2$.

Proposition 3.2.3 *On déduit du résultat sur l'espérance de la variable* S^2 *que* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ *est un estimateur sans biais de* $\theta = \sigma^2$ *car* $E(S^2) = \sigma^2$.

4. Prenons l'exemple de la statistique fréquence F.
On considère une population où une certaine proportion, un certain pourcentage d'individus ont une caractéristique donnée. Dans toute la population, avoir ou non une caractéristique donnée est une épreuve de Bernoulli. Par exemple, à l'issue d'une chaîne de fabrication, un article est défectueux avec la probabilité p, non défectueux avec la probabilité $1 - p = q$. Pour chaque article fabriqué, on peut définir la variable aléatoire X : lorsque l'article est défectueux, X prend la valeur 1 et $p(\{X = 1\}) = p$, lorsque l'article n'est pas défectueux, X prend la valeur 0 et $p(\{X = 0\}) = q$. X suit une loi de Bernoulli de paramètre p. Considérons un n-échantillon aléatoire simple de cette variable X soit $X_1, X_2, \ldots, X_n$ de réalisation $x_1, x_2, \ldots, x_n$. Ces n variables aléatoires indépendantes suivent toutes la même loi, celle de X, c'est-à-dire $\mathcal{B}(p)$. Leur somme $Y = X_1 + X_2 + \ldots + X_n$ suit une loi binomiale de paramètres n et p, $Y \rightsquigarrow \mathcal{B}(n, p)$ et $p(\{Y = k\}) = C_n^k p^k (1-p)^{n-k}$ pour k variant de 0 à

n.
On définit ensuite la variable fréquence

$$\boxed{F = \frac{Y}{n} = \frac{X_1 + X_2 + \dots X_n}{n}}$$

Cette variable fréquence est une variable aléatoire dont l'univers image est $\left\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\right\}$ dont on connaît la distribution de probabilité : $p\left(\left\{F = \frac{k}{n}\right\}\right) = p(\{Y = k\}) = C_n^k p^k q^{n-k}$.
La réalisation de cette variable fréquence est $f = \frac{x_1 + x_2 + \dots + x_n}{n}$, c'est-à-dire la fréquence de l'échantillon ou encore fréquence (ou pourcentage) des articles défectueux dans l'échantillon.

— Espérance de la variable fréquence F : on sait que $F = \frac{Y}{n}$ et $E(Y) = np$ donc $E(F) = \frac{1}{n}E(Y) = \frac{np}{n} = p$.

— Variance de la variable F : on sait que $F = \frac{Y}{n}$ et $V(Y) = npq$ donc $V(F) = \frac{1}{n^2}V(Y) = \frac{npq}{n^2} = \frac{pq}{n}$ et $\sigma(F) = \sqrt{\frac{pq}{n}}$.

On déduit du résultat sur l'espérance de la variable F que

Proposition 3.2.4 *$F = \frac{X_1 + X_2 + \dots + X_n}{n}$ est un estimateur sans biais de $\theta = p$ car $E(F) = p$.*

Dans le cas de tirages exhaustifs (les tirages ont lieu sans remise), si l'on désigne par N la taille de la population et par n la taille de l'échantillon, on obtient les résultats suivants, en faisant intervenir les facteurs d'exhaustivité :

1. $E(\overline{X}) = m$
2. $V(\overline{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$
3. $E(\Sigma^2) = \sigma^2\left(1 - \frac{1}{n}\right)\frac{N}{N-1}$
4. $E(S^2) = \sigma^2\frac{N}{N-1}$
5. $E(F) = p$
6. $V(F) = \frac{pq}{n}\frac{N-n}{N-1}$

Jusqu'ici, nous avons toujours parlé d'échantillons formés avec remise, situation où après avoir été choisie, une unité statistique de la population est remise avec les autre unités de la population, faisant en sorte qu'elle pourrait être choisie à nouveau pour faire partie de l'échantillon. Par le fait même, le choix se faisait toujours à partir d'un ensemble identique, donc les variables X_k étaient toutes identiquement distribuées et indépendantes. Dans notre situation, au fur et à mesure que les unités statistiques sont choisies, cela crée un "trou" dans la population et celle-ci est modifiée. Les variables $X_1, X_2, \dots, X_n$ perdent leur indépendance, elles ne sont plus des répliques de X : le résultat d'un processus influence le résultat du processus suivant, il y a dépendance. Sans remise, les échantillons seront choisis comme ce qui a été décrit pour la loi hypergéométrique. On remarque d'ailleurs que le facteur de correction $\frac{N-n}{N-1}$ qui apparaît dans la variance est le même facteur que celui qui était apparu dans cette loi.

3.3 Estimations pas intervalle de confiance

3.3.1 Préliminaires

Dans le cadre de l'estimation ponctuelle, on associe un nombre, une estimation à un paramètre dont la valeur est inconnue. La précision de cette estimation peut être déterminée en calculant un **intervalle de confiance** pour ce paramètre, c'est-à-dire un intervalle contenant la valeur inconnue du paramètre avec une grande probabilité donnée.

Définition 3.3.1 *Soit un modèle statistique $X \rightsquigarrow l(x,\theta)$ et soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire simple relatif à la variable X. On dit que $[C_1; C_2]$ est un **intervalle de confiance**, de niveau $1-\alpha$, du paramètre θ si on a*

$$p(\{C_1 \leq \theta \leq C_2\}) = 1 - \alpha.$$

Les bornes de l'intervalle C_1 et C_2 sont les statistiques basées sur l'échantillon aléatoire. À priori C_1 et C_2 sont des variables aléatoires, une fois les réalisations de l'échantillon obtenues, on dispose des valeurs numériques $x_1, x_2, \ldots, x_n$. On remplace C_1 et C_2 par leurs réalisations et on obtient les bornes de l'intervalle recherché. Cet intervalle est une réalisation de l'intervalle de confiance $[C_1; C_2]$.

Remarque 3.3.1 Généralement, on prend des intervalles à risque symétrique, c'est-à-dire tels que

$$p(\{\theta < C_1\}) = p(\{\theta > C_2\}) = \frac{\alpha}{2}$$

3.3.2 Intervalle de confiance pour une proportion

Dans une population donnée de grande taille, la proportion d'individus p ayant une caractéristique donnée $\mathcal{C}$ est inconnue. On désire déterminer, à partir d'un tirage d'un échantillon non exhaustif de taille n de la population, un intervalle de confiance au risque α de p.
Le tirage de cet échantillon peut être modélisé par un n-échantillon au hasard tiré d'une variable aléatoire F qui suit une loi de Bernoulli de paramètre p. Soient donc $X \rightsquigarrow \mathcal{B}(p)$ une loi de Bernoulli de paramètre p et $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire simple. La fréquence $F = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ est un bon estimateur (estimateur sans biais, convergent et efficace) du paramètre p, où chacune des variables aléatoires X_i suit une loi de Bernoulli. La fréquence est un estimateur asymptotiquement normal et on utilise l'approximation $F \rightsquigarrow \mathcal{N}\left(p, \sqrt{\dfrac{pq}{n}}\right)$ pour $n \geq 30$, $np \geq 5$ et $nq \geq 5$. Ces conditions seront appelées les **conditions de normalité**.

Les tables statistiques (cf Annexe A) fournissent les valeurs Z_α telles que $p(\{-Z_\alpha < Z < Z_\alpha\}) = 1 - \alpha$ avec $Z \rightsquigarrow \mathcal{N}(0,1)$. On applique cette relation à la variable $Z = \dfrac{F-p}{\sqrt{\frac{pq}{n}}}$ qui suit une loi normale $\mathcal{N}(0,1)$. On obtient

$$\boxed{p\left(\left\{-Z_\alpha < \frac{F-p}{\sqrt{\frac{pq}{n}}} < Z_\alpha\right\}\right) = 1 - \alpha}$$

Remarquons que $-Z_\alpha < \dfrac{F-p}{\sqrt{\frac{pq}{n}}} < Z_\alpha \Leftrightarrow -Z_\alpha < \dfrac{p-F}{\sqrt{\frac{pq}{n}}} < Z_\alpha \Leftrightarrow F - Z_\alpha\sqrt{\dfrac{pq}{n}} < p < F + Z_\alpha\sqrt{\dfrac{pq}{n}}$. On obtient un intervalle de confiance de p au niveau de confiance $1-\alpha$ soit $\left[F - Z_\alpha\sqrt{\dfrac{pq}{n}}; F + Z_\alpha\sqrt{\dfrac{pq}{n}}\right]$.
Pour un risque $\alpha = 5\%$, on trouve $Z_\alpha = 1,96$ et l'intervalle de confiance est :

$$\left[F - 1,96\sqrt{\frac{pq}{n}}; F + 1,96\sqrt{\frac{pq}{n}}\right],$$

pour un risque $\alpha = 1\%$, on trouve $Z_\alpha = 2,58$ et l'intervalle de confiance est :

$$\left[F - 2,58\sqrt{\frac{pq}{n}}; F + 2,58\sqrt{\frac{pq}{n}}\right],$$

Cet intervalle pose un problème pratique important, on peut affirmer que la proportion p appartient à cet intervalle avec une probabilité de $1 - \alpha$ mais les bornes de cet intervalle dépendent de p, la proportion inconnue. Deux possibilités sont utilisées :

1. on remplace p et q par leurs estimations ponctuelles f et $1 - f$. La réalisation de l'intervalle de confiance est alors $\left[f - Z_\alpha\sqrt{\frac{f(1-f)}{n}}; f + Z_\alpha\sqrt{\frac{f(1-f)}{n}}\right]$.

   Exemple 3.3.1 *En vue d'un contrôle de qualité on observe la fabrication d'un objet par une machine durant une période donnée. On décide de tirer un échantillon non exhaustif de taille* $n = 1000$ *dans la fabrication. On constate que 60 d'entre eux sont défectueux. Déterminer au risque de* 5% *un intervalle de confiance de la proportion d'objets défectueux durant la période donnée.*

   Les données sont $n = 1000$ et $f = \dfrac{60}{1000} = 0,06$. La taille de l'échantillon est grande ($n > 30$). Le risque de 5% conduit à $Z_\alpha = 1,96$. L'intervalle de confiance numérique est donc, au risque de 5% :

   $$\left[0,06 - 1,96\sqrt{\frac{0,06(1-0,06)}{1000}}; 0,06 + 1,96\sqrt{\frac{0,06(1-0,06)}{1000}}\right] = [0,045; 0,075].$$

   La proportion p d'objets défectueux fabriqués par la machine est, au risque de 5%, telle que $4,5\% \leq p \leq 7,5\%$.

2. Deuxième méthode : $pq = p(1-p) = -p^2 + p = f(p)$. Alors $f'(p) = -2p + 1$ et on en déduit que f est croissante sur $[0; \frac{1}{2}[$ et décroissante sur $]\frac{1}{2}; 1]$. Dans le cas où p est voisin de $\frac{1}{2}$, on remplace pq par sa valeur maximale $\frac{1}{4}$. La réalisation de l'intervalle de confiance est alors $\left[f - \dfrac{Z_\alpha}{2\sqrt{n}}; f + \dfrac{Z_\alpha}{2\sqrt{n}}\right]$.

   Cette méthode, qui permet un calcul rapide, donne un intervalle de confiance de grande amplitude car la valeur $\dfrac{1}{4}$ du produit $p(1-p)$ est surestimée.

   Exemple 3.3.2 *On a besoin d'estimer rapidement la proportion p d'accidents du travail dans une entreprise de construction. On a constaté sur un échantillon de* 200 *jours ouvrables qu'il y a eu 18 accidents. Déterminer, au risque de* 5%*, un intervalle de confiance de la proportion d'accidents.*

   Les données sont $n = 200$ et $f = \dfrac{18}{200} = 0,09$. Pour un calcul rapide, l'intervalle de confiance numérique est donc, au risque de 5% :

   $$\left[0,09 - 1,96\frac{1}{2\sqrt{200}}; 0,09 + 1,96\frac{1}{2\sqrt{200}}\right] = [0,02; 0,159].$$

   La proportion d'accidents est au risque de 5% telle que $2\% \leq p \leq 15,9\%$. On se rappellera que cette méthode augmente l'amplitude de l'intervalle de confiance. Le calcul fait avec la première méthode donnerait une proportion d'accident p telle que $5\% \leq p \leq 12,99\%$.

Remarque 3.3.2

- On a utilisé l'approximation normale déduite du théorème central limite pour établir l'intervalle de confiance. Il est donc nécessaire que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$. Dans la pratique, p est inconnue, on vérifie ces conditions sur f donc $n \geq 30$, $nf \geq 5$ et $n(1-f) \geq 5$.
- La longueur de l'intervalle de confiance est $L(\alpha, n) = 2Z_\alpha\sqrt{\dfrac{f(1-f)}{n}}$.

- La précision de l'estimation obtenue est $\frac{1}{2}L(\alpha,n) = Z_\alpha\sqrt{\frac{f(1-f)}{n}}$.
- Z_α étant une fonction décroissante de α (risque pris par le statisticien), lorsque $1-\alpha$ augmente, α diminue, Z_α augmente, la longueur de l'intervalle augmente.
- Lorsqu'on a choisi la valeur de α, on peut imaginer de déterminer la taille de l'échantillon nécessaire pour atteindre une précision donnée l soit $Z_\alpha\sqrt{\frac{f(1-f)}{n}} < l$. On obtient $n > Z_\alpha^2\frac{f(1-f)}{l^2}$.

Exemple 3.3.3

1. On réalise un sondage en vue de prévoir le résultat de l'élection présidentielle. On effectue un tirage aléatoire simple de 750 électeurs. Parmi eux, 324 déclarent qu'ils ont l'intention de voter pour le candidat A tandis que 426 électeurs affirment qu'ils vont voter pour le candidat B. Donner un intervalle de confiance au niveau 95% pour la proportion d'électeurs qui vont voter pour le candidat A.

   On définit la variable X de la manière suivante :
   — si un électeur quelconque vote pour le candidat A, $X = 1$ et $p(\{X = 1\}) = p$,
   — s'il vote pour le candidat B, $X = 0$ et $p(\{X = 0\}) = 1 - p$.
   X suit une loi de Bernoulli de paramètre p. $X_1, X_2, \ldots, X_{750}$ est un échantillon aléatoire simple. Une estimation ponctuelle de p est $f = \frac{324}{750} = 0,432$. L'intervalle de confiance est donné par

$$\left[f - Z_\alpha\sqrt{\frac{f(1-f)}{n}}; f + Z_\alpha\sqrt{\frac{f(1-f)}{n}}\right].$$

   Les conditions de normalité sont vérifiées car $n = 750 \geq 30$, $nf = 750 \times \frac{324}{750} = 324 \geq 5$, $n(1-f) = 426 \geq 5$. On obtient dans la table $Z_{0,05} = 1,960$. L'intervalle numérique est donné par :

$$\left[0,432 - 1,96\sqrt{\frac{0,432 \times 0,568}{750}}; 0,432 + 1,96\sqrt{\frac{0,432 \times 0,568}{750}}\right] = [0,39656; 0,47746].$$

   On peut affirmer que $p(\{0,39656 < p < 0,47746\}) = 0,95$ et $p(\{p < 0,39656\}) = p(\{p > 0,46746\}) = 0,025$.
   Pour $1-\alpha = 0,99$ donc un risque de 1% on a $Z_\alpha = 2,575829$, l'intervalle de confiance est $[0,3854; 0,4786]$.

   Remarque 3.3.3 Cet exemple fait apparaître le peu d'intérêt que présente souvent les sondages tels qu'on les donne dans la presse c'est-à-dire sans donner l'intervalle de confiance ni le niveau de confiance.

2. On souhaite estimer avec une précision de 2% au niveau de confiance $1 - \alpha = 90\%$ le pourcentage de sujets non immunisés après une vaccination. Sur combien de sujets l'observation doit-elle porter sachant que le pourcentage observé de personnes non immunisées est
   (a) $f = 0,20$
   (b) $0,2 < f < 0,3$

   (a) Supposons les conditions de normalité $n \geq 30$, $nf \geq 5$, $n(1-f) \geq 5$ vérifiées avec $\alpha = 0,10$, $f = 0,20$ et $Z_{0,10} = 1,645$. Il faut que

$$Z_\alpha\sqrt{\frac{f(1-f)}{n}} \leq 0,02 \Leftrightarrow n \geq \frac{Z_\alpha^2 f(1-f)}{(0,02)^2} = \frac{(1,645)^2 \times 0,2 \times 0,8}{0,0004} = 1083.$$

   (b) Étudions les variations de $f(1-f)$. Soit $g(x) = x(1-x) = -x^2 + x$, $g'(x) = -2x + 1$. On en déduit alors que g est croissante sur $[0,2; 0,3]$ (avec $g(0,2) = 0,2 \times 0,8$ et $g(0,3) = 0,3 \times 0,7$). Ainsi, $0,16 < f(1-f) < 0,21$ et par conséquent,

$$Z_\alpha\sqrt{\frac{f(1-f)}{n}} < Z_\alpha\sqrt{\frac{0,21}{n}} \leq 0,02 \Leftrightarrow n \geq \frac{(1,645)^2 \times 0,21}{0,0004} \simeq 1421.$$

3.3.3 Intervalle de confiance pour l'espérance

1. La variance σ^2 est supposée connue.
La variable aléatoire parente X suit une loi de probabilité de paramètre $m = E(X)$ inconnu et de variance σ^2 connue. Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire simple de X. On sait alors qu'un bon estimateur ponctuel de m est $\overline{X}$ (estimateur sans biais, convergent et efficace) et que

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \rightsquigarrow \mathcal{N}(m, \frac{\sigma}{\sqrt{n}}) \text{ et } Z = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow \mathcal{N}(0,1).$$

Les tables fournissent la valeur Z_α, pour α donné, telle que $p(\{-Z_\alpha < Z < Z_\alpha\}) = 1 - \alpha$. Or

$$-Z_\alpha < Z < Z_\alpha \Leftrightarrow -Z_\alpha < \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} < Z_\alpha$$
$$\Leftrightarrow -Z_\alpha \frac{\sigma}{\sqrt{n}} < \overline{X} - m < Z_\alpha \frac{\sigma}{\sqrt{n}} \Leftrightarrow -Z_\alpha \frac{\sigma}{\sqrt{n}} < m - \overline{X} < Z_\alpha \frac{\sigma}{\sqrt{n}}$$

On obtient ainsi $p(\{\overline{X} - Z_\alpha \frac{\sigma}{\sqrt{n}} < m < \overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\}) = 1 - \alpha$ c'est-à-dire un intervalle de confiance de m au niveau de confiance $1 - \alpha$ soit $\left[\overline{X} - Z_\alpha \frac{\sigma}{\sqrt{n}}; \overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right]$. Dans la pratique, on dispose d'un échantillon non exhaustif tiré au hasard de la population. Cet échantillon fournit une réalisation de $\overline{X}$ par le calcul de la moyenne $\overline{x}$. Ainsi l'échantillon donne une réalisation de l'intervalle de confiance au risque α qui est $\left[\overline{x} - Z_\alpha \frac{\sigma}{\sqrt{n}}; \overline{x} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right]$.

Exemple 3.3.4 Une machine M fabrique des engrenages en grande série. Des études antérieures permettent de dire que les mesures des diamètres forment une population normale d'écart-type $\sigma = 0,042$ cm. On extrait un échantillon non exhaustif de la fabrication journalière de taille $n = 200$ engrenages. La moyenne des diamètres sur cet échantillon est $\overline{x} = 0,824$ cm. Donner au seuil de confiance 95% un intervalle de confiance de la moyenne m des diamètres des engrenages.
Considérons D la variable aléatoire égale au diamètre des engrenages. L'énoncé dit que $D \rightsquigarrow \mathcal{N}(m, \sigma = 0,042)$. Soit $D_1, D_2, \ldots, D_{200}$ un 200-échantillon au hasard de D. Les $n = 200$ variables aléatoires D_i suivent la même loi $\mathcal{N}(m, \sigma = 0,042)$ que D. Soit m le diamètre moyen inconnu des engrenages. On considère alors l'estimateur sans biais et convergent $\overline{D} = \frac{1}{200}\sum_{i=1}^{200} D_i$ de m. Une réalisation de $\overline{D}$ est $\overline{d} = 0,824$. On sait que l'intervalle de confiance au rsique α est $\left[\overline{D} - Z_\alpha \frac{\sigma}{\sqrt{n}}; \overline{D} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right]$. Pour un risque de 5% on a $Z_\alpha = 1,96$. Ainsi, l'intervalle de confiance est $\left[\overline{D} - 1,96\frac{0,042}{\sqrt{200}}; \overline{D} + 1,96\frac{0,042}{\sqrt{200}}\right]$. L'échantillon fournit une réalisation de cet intervalle de confiance à savoir $\left[0,824 - 1,96\frac{0,042}{\sqrt{200}}; 0,824 + 1,96\frac{0,042}{\sqrt{200}}\right]$ soit $[0,818; 0,830]$.

2. La variance est inconnue.
Dans la pratique, si l'espérance $m = E(X)$ est inconnue, a fortiori, la variance $\sigma^2 = E[(X - m)^2]$ est également inconnue. Or nous venons de voir que l'intervalle de confiance de m tel qu'il vient d'être défini dépend de σ. Il est alors tentant de remplacer σ par son estimation ponctuelle s fournie par l'estimateur S^2. Ce nombre n'est autre que l'écart-type calculé sur l'échantillon de taille n avec $n - 1$ degrés de liberté (ddl). Dans ces conditions, on utlise le procédé dit de **Studentisation** qui consiste à remplacer la variable centrée réduite $Z = \frac{\overline{X} - E(\overline{X})}{\sigma(\overline{X})}$ par la variable $T = \frac{\overline{X} - E(\overline{X})}{\frac{S}{\sqrt{n}}}$ qui suit une loi de Student à $n - 1$ ddl. La table de Student nous permet de déterminer $t_{n-1,\alpha}$ tel que pour $n - 1$ ddl on ait

$$p(-t_{n-1,\alpha} \leq T \leq t_{n-1,\alpha}) = 1 - \alpha.$$

On obtiendra alors l'intervalle de confiance au risque α :

$$\left[\overline{X} - t_{n-1,\alpha}\frac{S}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}\right]$$

dont une réalisation sur l'échantillon est $\left[\overline{x} - t_{n-1,\alpha}\frac{s}{\sqrt{n}}; \overline{x} + t_{n-1,\alpha}\frac{s}{\sqrt{n}}\right]$.

Exemple 3.3.5 *Dans l'atmosphère, le taux de gaz nocif, pour un volume donné, suit une loi normale d'espérance et de variance inconnues. On effectue n prélèvements conduisant aux valeurs numériques* $x_1, x_2, \ldots, x_n$.

(a) Sur un échantillon de taille $n = 10$*, on observe* $\overline{x} = 50$ *et* $s^2 = 100$.
Quel est l'intervalle de confiance à 5% du taux moyen m de gaz dans l'atmosphère ?

(b) Quel serait cet intervalle si la variance σ^2 *du taux de gaz nocif était connue et valait exactement 100 ?*

(a) Considérons X la variable aléatoire égale au taux de gaz nocif dans l'atmosphère. L'énoncé dit que $X \rightsquigarrow \mathcal{N}(m, \sigma)$ avec m et σ^2 inconnues (m représente le taux moyen de gaz nocif dans l'atmosphère). Soit $X_1, X_2, \ldots, X_n$ un n-échantillon au hasard de X. Les n variables aléatoires X_i suivent la même loi $\mathcal{N}(m, \sigma)$ que X. Les valeurs observées $x_1, x_2, \ldots, x_n$ sont une réalisation du n-échantillon de X. On considère alors l'estimateur sans biais et convergent $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ de m.
Une réalisation de $\overline{X}$ sur l'échantillon est $\overline{x} = 50$. L'intervalle de confiance de m au risque α est l'intervalle aléatoire $\left[\overline{X} - Z_\alpha\frac{\sigma}{\sqrt{n}}; \overline{X} + Z_\alpha\frac{\sigma}{\sqrt{n}}\right]$ où Z_α est déterminé par $p(-Z\alpha < T < Z_\alpha) = 1 - \alpha$ avec $T \rightsquigarrow \mathcal{N}(0, 1)$.
Or σ est inconnu, on le remplace donc par son estimation s. L'intervalle de confiance sur l'échantillon devient : $\left[\overline{X} - t_{n-1,\alpha}\frac{S}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha}\frac{S}{\sqrt{n}}\right]$ où $t_{n-1,\alpha}$ est déterminé par $p(-t_{n-1,\alpha} < T < t_{n-1,\alpha}) = 1 - \alpha$ avec T qui suit une loi de Student à $n - 1$ ddl.
Pour $\alpha = 5\%$ et $n - 1 = 9$, on obtient dans la table $t_{9;0,05} = 2,262$. L'intervalle de confiance recherché est donc : $\left[\overline{X} - 2,262\frac{S}{\sqrt{n}}; \overline{X} + 2,262\frac{S}{\sqrt{n}}\right]$. Une réalisation de cet intervalle de confiance sur l'échantillon est : $\left[\overline{x} - 2,262\frac{s}{\sqrt{n}}; \overline{x} + 2,262\frac{s}{\sqrt{n}}\right]$ soit $\left[50 - 2,262\frac{10}{\sqrt{10}}; 50 + 2,262\frac{10}{\sqrt{10}}\right] = [42,84; 57,15]$ numériquement, qui est donc l'intervalle de confiance, au risque 5% du taux moyen du gaz nocif dans l'atmosphère.

(b) Si la variance est connue et égale à 100, on utilise la table de la loi normale pour déterminer $Z_{0,05} = 1,96$. L'intervalle de confiance est alors $\left[\overline{X} - 1,96\frac{\sigma}{\sqrt{n}}; \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right]$. Et une réalisation sur l'échantillon est : $\left[50 - 1,96\frac{10}{\sqrt{10}}; 50 + 1,96\frac{10}{\sqrt{10}}\right] = [43,80; 56,20]$.
L'utilisation de la loi normale donne un intervalle de confiance d'amplitude plus petit que celui obtenu à l'aide de la loi de Student. Ce résultat est cohérent car pour l'utilisation de la loi normale, on a supposé que l'écart-type σ était connu. On a donc une meilleure connaissance de la loi de X que dans le cas où σ est inconnu.

Remarque 3.3.4 Si la taille de l'échantillon est "grande" ($n > 30$), on peut utiliser la loi normale à la place de la loi de Student. C'est pour cette raison qu'on trouve dans la littérature l'expression : "la loi de Student est la loi des petits échantillons".

3.3.4 Intervalle de confiance pour la variance

Théorème 3.3.1 *Soit X une variable aléatoire telle que $X \rightsquigarrow \mathcal{N}(m,\sigma)$ et $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire de X. On utilise $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ comme estimateur sans biais et convergent de σ^2. Alors, la variable aléatoire $\dfrac{n-1}{\sigma^2}S^2$ suit une loi de χ^2 à $n-1$ ddl. On note*

$$\frac{n-1}{\sigma^2}S^2 \rightsquigarrow \chi^2.$$

Un intervalle de confiance au risque α est de la forme $[a;b]$ où a et b sont deux variables aléatoires construites à partir d'un n-échantillon au hasard de X telles que $p(a \leq \sigma^2 \leq b) = 1-\alpha$ or

$$a \leq \sigma^2 \leq b \Leftrightarrow \frac{1}{b} \leq \frac{1}{\sigma}^2 \leq \frac{1}{a} \Leftrightarrow \frac{(n-1)S^2}{b} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)S^2}{a}.$$

Posons $U = \dfrac{(n-1)S^2}{\sigma^2}$, $t_b = \dfrac{(n-1)S^2}{b}$ et $t_a = \dfrac{(n-1)S^2}{a}$. Alors,

$$p(a \leq \sigma^2 \leq b) = 1-\alpha \Leftrightarrow p(t_b \leq U \leq t_a) = 1-\alpha \Leftrightarrow \left\{p(U < t_b) = \frac{\alpha}{2} \text{ et } p(U > t_a) = \frac{\alpha}{2}\right\}.$$

Comme U suit une loi de χ^2, on détermine les valeurs de t_a et t_b à l'aide d'une table du χ^2 à $n-1$ degrés de liberté. Comme $b = \dfrac{(n-1)S^2}{t_b}$ et $a = \dfrac{(n-1)S^2}{t_a}$, l'intervalle de confiance cherché est alors : $\left[\dfrac{(n-1)S^2}{t_a}; \dfrac{(n-1)S^2}{t_b}\right]$. Une réalisation de cet intervalle de confiance sur l'échantillon est : $\left[\dfrac{(n-1)s^2}{t_a}; \dfrac{(n-1)s^2}{t_b}\right]$ où s est l'estimation ponctuelle de la variance de la population avec $n-1$ ddl.

Exemple 3.3.6 *Une variable aléatoire X est distribuée selon une loi normale de paramètres m et σ inconnus. On dispose d'un n-échantillon associé à X de taille $n = 16$. Sur cet échantillon on observe $\sum_{i=1}^{16}(x_i - \overline{x})^2 = 1500$. Déterminer un intervalle de confiance de la variance au seuil de confiance de 95%.*
La moyenne et la variance de la population sont inconnues. L'intervalle de confiance au seuil 95% de σ^2 est donné par $\left[\dfrac{(n-1)S^2}{t_a}; \dfrac{(n-1)S^2}{t_b}\right]$ et sa réalisation sur l'échantillon est donnée numériquement par $\left[\dfrac{(16-1)s^2}{t_a}; \dfrac{(16-1)s^2}{t_b}\right]$. On a $s^2 = \dfrac{1}{15} \times 1500 = 100$. Pour calculer t_a et t_b, on utilise la variable aléatoire U qui suit une loi du χ^2 à $16-1 = 15$ ddl. D'après les résultats précédents on a $p(U > t_b) = 0,975$ et $p(U < t_a) = 0,025$. La table donne $t_b = 6,26$ et $t_a = 27,5$. On obtient alors l'intervalle de confiance cherché $\left[\dfrac{1500}{27,5}; \dfrac{1500}{6,26}\right] = [54,54; 239,60]$.

3.4 Exercices

Exercice 18 Une entreprise fabrique des sacs en plastique biodégradables pour les enseignes de distribution. Elle s'intéresse au poids maximal que ces sacs peuvent supporter sans se déchirer. On suppose ici que le poids maximal que ces sacs peuvent supporter suit une loi normale d'espérance mathématique 58 (kg) et d'écart-type 3 (kg).

1. Sur 200 sacs reçus, une grande enseigne de distribution constate un poids moyen de 57,7 kg.
   (a) Donner un intervalle de confiance bilatéral de la moyenne des poids sur un échantillon de taille 200, au seuil de 1%.

(b) Quelle est votre conclusion sur le poids moyen constaté ?

2. Donner le poids moyen dépassé dans 97% des cas, sur un échantillon de taille 200.

Exercice 19 Les résultats d'un enquête effectuée sur une population de 1500 salariés d'une entreprise ont montré que dans 65% des cas, les individus avaient au moins un crédit en cours. Trouver la probabilité pour que 2 échantillons de 200 personnes chacun indiquent plus de 10 points d'écart entre les proportions de personnes ayant au moins un crédit en cours.

Exercice 20 Afin de mieux gérer les demandes de crédits de ses clients, un directeur d'agence bancaire réalise une étude relative à la durée de traitement des dossiers, supposée suivre une distribution normale. Un échantillon non exhaustif de 30 dossiers a donné :

Durée (mn)	0-10	10-20	20-30	30-40	40-50	50-60
Effectif	3	6	10	7	3	1

1. Calculer la moyenne et l'écart-type des durées de traitement des dossiers de cet échantillon.
2. En déduire les estimations ponctuelles de la moyenne m et de l'écart-type σ de la population des dossiers.
3. Donner une estimation de m par l'intervalle de confiance au seuil de risque 5%.

Exercice 21 La société G@E a mis au point un logiciel de gestion destiné essentiellement aux PME. Après une enquête dans la région Aquitaine-Limousin-Poitou-Charentes (ALPC), auprès de 100 entreprises déjà équipées d'un matériel informatique (PC) apte à recevoir ce logiciel, la société G@E décide de fixer le prix de vente à 200€. Elle espère diffuser son produit auprès de 68% des PME de la région (cette valeur constituera la proportion de ventes sur l'échantillon). On peut admettre que les 100 PME interrogées constituent un échantillon représentatif des 13250 PME formant le marché potentiel (en 2015).

1. Déterminer l'intervalle de confiance de la proportion p des entreprises intéressées par le logiciel, au seuil de risque 1%.
2. Quelle aurait dû être la taille de l'échantillon pour que l'amplmitude de l'intervalle de confiance soit de 20 points (erreur de $0,1$) ?

Exercice 22 Des observations sur une longue période de la fabrication d'un certain type de boulons ont montré que la résistance à la rupture suit une loi normale dont l'écart-type est $\sigma = 34,5$. Lors d'un contrôle de fabrication, on tire un échantillon de 8 éléments dans la population des boulons fabriqués qui est de très grand effectif. On trouve pour moyenne de résistance à la rupture dans cet échantillon $\overline{x}_e = 225$.

1. Pourquoi à votre avis le tirage des boulons se fait-il sans remise ? Pourquoi peut-on le considérer malgré tout comme un tirage avec remise ? (utiliser le facteur d'exhaustivité $\dfrac{N-n}{N-1}$)
2. Déterminer un estimation de la moyenne des résistances à la rupture des boulons de la fabrication. Prouver que la variable aléatoire $\overline{X}$ suit la loi normale $\mathcal{N}(225, 12.198)$ et donner l'intervalle de confiance, pour cette moyenne m, au seuil de 95%. Qu'en est-il au seuil de 99%.

Exercice 23 Lors d'un contrôle de qualité sur une population d'appareils électro-ménagers, au cours d'un mois de fabrication, on prélève d'une manière non exhaustive un échantillon de 1000 appareils. Après un test de conformité, on constate que 60 appareils ont un défaut.

1. Justifier que la variable « proportion d'appareils électro-ménagers défectueux » suit une loi binomiale de paramètres que l'on précisera. Les conditions de normalité sont-elles vérifiées et si oui, que permettent-elles d'affirmer ?

2. Donner un intervalle de confiance du pourcentage d'appareils défectueux, au risque de 5%.

Exercice 24 Soient les notes de mathématiques d'un étudiant :

4, 5, 8, 10, 12, 13

1. Calculer la moyenne et l'écart-type de la population des notes.
2. Former tous les échantillons exhaustifs possibles de taille 2
3. Calculer l'espérance et l'écart-type de la distribution d'échantillonnage des moyennes.

Exercice 25 On considère une population U de $N = 5$ individus pour lesquels on connaît les valeurs de la variable Y : $y_1 = 3$, $y_2 = 1$, $y_3 = 0$, $y_4 = 1$, $y_5 = 5$. On choisit un 3-échantillon aléatoire simple S dans cette population.

1. Donner les valeurs de la moyenne, de la médiane et de la variance de la variable Y dans la population. Lister tous les échantillons possibles de taille $n = 3$. Quelle est la probabilité de sélection de chaque échantillon ?
2. Pour un échantillon donné, on estime la moyenne (respectivement la médiane) de la population. Calculer les valeurs de ces estimateurs pour chaque échantillon et en déduire que l'estimateur de la moyenne est sans biais alors que l'estimateur de la médiane est biaisé.
3. Pour chaque échantillon, calculer l'estimateur S^2_{YS} de S^2_{YU} et en déduire que cet estimateur est sans biais (on rappelle que $S^2 = \dfrac{1}{n-1}\sum_i (X_i - \overline{X})^2$).

Exercice 26

1. En utilisant la table de nombres aléatoires données ci-après, engendrer un échantillon non exhaustif de taille $n = 10$, de la variable normale X de moyenne $\mu = 5$ et d'écart-type $\sigma = 2$.
2. Vérifier la normalité de l'échantillon obtenu à l'aide d'un graphique gausso-arithmétique.
3. Calculer une estimation sans biais de la moyenne de X donnée par cet échantillon.
4. Calculer une estimation sans biais de la variance de X, donnée par cet échantillon.
5. En déduire l'intervalle de confiance à 95% de la moyenne.

EXTRAITS D'UNE TABLE DE NOMBRES AU HASARD

(Kendall et Babington Smith, table tirée de Christian Labrousse, Statistique, Tome2, Dunod, Paris, 1962)

02 22 85 19 48 74 55 24 89 69 15 53 00 20 88 48 95 08
85 76 34 51 40 44 62 93 65 99 72 64 09 34 01 13 09 74
00 88 96 79 38 24 77 00 70 91 47 43 43 82 71 67 49 90
64 29 81 85 50 47 36 50 91 19 09 15 98 75 60 58 33 15
94 03 80 04 21 49 54 91 77 85 00 45 68 23 12 94 23 44
42 28 52 73 06 41 37 47 47 31 52 99 89 82 22 81 86 55
09 27 52 72 49 11 30 93 33 29 54 17 54 48 47 42 04 79
54 68 64 07 85 32 05 96 54 79 57 43 96 97 30 72 12 19
25 04 92 29 71 11 64 10 42 23 23 67 01 19 20 58 35 93
28 58 32 91 95 28 42 36 98 59 66 32 15 51 46 63 57 10
64 35 04 62 24 87 44 85 45 68 41 66 19 17 13 09 63 37
61 05 55 88 25 01 15 77 12 90 69 34 36 93 52 39 36 23

98 93 18 93 86 98 99 04 75 28 30 05 12 09 57 35 90 15
61 89 35 47 16 32 20 16 78 52 82 37 26 33 67 42 11 93
94 40 82 18 06 61 54 67 03 66 76 82 90 31 71 90 39 27
54 38 58 65 27 70 93 57 59 00 63 56 18 79 85 52 21 03
63 70 89 23 76 46 97 70 00 62 15 35 97 42 47 54 60 60
61 58 65 62 81 29 69 71 95 53 53 69 20 95 66 60 50 70
51 68 98 15 05 64 43 32 74 07 44 63 52 38 67 59 56 69
59 25 41 48 64 79 62 26 87 86 94 30 43 54 26 98 61 38
85 00 02 24 67 85 88 10 34 01 54 53 23 77 33 11 19 68
01 46 87 56 19 19 19 43 70 25 24 29 48 22 44 81 35 40
42 41 25 10 87 27 77 28 05 90 73 03 95 46 88 82 25 02
03 57 14 03 17 80 47 85 94 49 89 55 10 37 19 50 20 37
18 95 93 40 45 43 04 56 17 03 34 54 83 91 69 02 90 72

Chapitre 4

Le test du χ^2

4.1 Les données du problème

Certains tests statistiques ont pour objet de tirer des conclusions relatives à la valeur des paramètres (moyenne, fréquence, variance) d'une ou plusieurs populations, sur la base d'informations partielles fournies par un ou plusieurs échantillons.
La même démarche peut être appliquée pour porter un "jugement" sur les caractéristiques encore plus générales de la population : la forme même de distribution du caractère étudié, la validité de sa représentation à l'aide de telle ou telle loi de probabilité particulière, les relations éventuelles entre plusieurs variables.

Concrètement, on dispose d'une distribution statistique empirique se présentant sous la forme d'une table d'effectifs ou de fréquences du caractère étudié. On désire savoir si ces effectifs ou ces fréquences sont compatibles avec une distribution théorique déterminée telle que la loi binomiale, la loi de Poisson, la loi normale ou toute autre loi de probabilité. Il s'agit en d'autres termes d'apprécier l'adéquation d'une distribution théorique particulière, en tant que représentation d'un phénomène concret observé (série empirique).

La démarche consiste donc à tester l'hypothèse selon laquelle notre échantillon serait tiré d'une population régie par une certaine loi de probabilité.
Il est évident que, même si le phénomène étudié suit effectivement la loi de probabilité dont on teste l'adéquation, les fréquences expérimentales (ou empiriques) observées sur un échantillon particulier différeront nécessairement peu ou prou des probabilités (fréquences que l'on devrait théoriquement observer selon la loi en question).

La problématique du test revient en définitive à savoir si les différences constatées entre la distribution expérimentale et la distribution théorique supposée sont explicables par l'aléa lié à la constitution de l'échantillon ou si elles sont trop importantes pour être imputables au seul hasard. En ce cas, c'est l'hypothèse de travail avancée sur la nature de la distribution qui devrait être mise en cause.

4.2 Ajustement d'une distribution observée à une distribution théorique

4.2.1 Construction du test

1. Les hypothèses du test sont les suivantes :
   - H_0 : X suit la loi théorique L,
   - H_1 : X ne suit pas L.
2. La variable observée est :
   — soit discrète et prend k valeurs x_1, x_2,. . . ,x_k

— soit continue et classée en k classes $[a_0; a_1[$, $[a_1; a_2[$,. . . ,$[a_{k-1}; a_k[$ de centres respectifs x_1, x_2,. . . ,x_{k-1}, x_k.

3. Les N observations de l'échantillon sont réparties sur les k valeurs de X (si X est discrète) ou sur les k classes de X (si X est continue). On a les tableaux suivants :

x_i	n_i
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_k	n_k

Classes	Centres x_i	Effectifs n_i
$[a_0; a_1[$	x_1	n_1
$[a_1; a_2[$	x_2	n_2
$\vdots$	$\vdots$	$\vdots$
$[a_{k-1}; a_k[$	x_k	n_k

avec $N = \sum_{i=1}^{k} n_i = n_1 + n_2 + \ldots + n_k$.

4. Sous H_0 on note p_i la probabilité dite théorique définie par
   - $p_i = p(\{X = x_i / X \rightsquigarrow L\})$ si X est discrète,
   - $p_i = p(\{X \in [a_{i-1}; a_i[/ X \rightsquigarrow L\})$ si X est continue.

   $e_i = Np_i$ est l'effectif théorique de la i-ième classe de X.

5. L'indicateur d'écart entre les distributions observées et théoriques est

$$\boxed{\sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i}} \quad (1)$$

dit χ^2 **observé ou calculé**. Cet écart suit pour N suffisamment grand une loi du χ^2_ν d'où le nom du test.

Intuitivement, on comprend que cette grandeur statistique traduise l'écart entre l'échantillon et la loi conjecturée.
Si l'ajustement était parfait, cette expression du χ^2 serait nulle, les effectifs empiriques co'ıncidant exactement avec les effectifs théoriques.
En revanche, plus grands sont les écarts entre les effectifs observés et les effectifs théoriques $(n_i - e_i)$ et plus forte sera la valeur du χ^2.

En outre, comme la quantité (1) ne peut pas être négative, le test d'ajustement est nécessairement un test unilatéral droit.

Le paramètre ν indiçant χ^2_ν définit le **nombre de degrés de liberté**. C'est le nom donné au nombre d'observations linéairement indépendantes qui apparaissent dans une somme de carrés. Autrement dit, c'est le nombre d'observations aléatoires indépendantes à qui l'on soustrait le nombre de contraintes imposées à ces observations.

Le nombre ν de degrés de liberté est égal à

— si les paramètres de la loi d'ajustement L sont donnés,

$$\boxed{\nu = k - 1}$$

En effet, aucun paramètre n'est à estimer puisque la loi d'ajustement est totalement spécifiée. Le χ^2 est constitué de k écarts $(n_i - e_i)$. Les écarts sont reliés par la contrainte

$$\sum(n_i - e_i) = \sum(n_i - Np_i) = \sum n_i - N\sum p_i = N - N = 0$$

En d'autres termes, lorsqu'on connaît la valeur de $k-1$ écarts, on peut en déduire la valeur du dernier qui n'est donc pas "libre" de varier de manière aléatoire,

— si la loi d'ajustement L comporte r paramètres inconnus,

$$\boxed{\nu = k - r - 1}$$

On impose de ce fait autant de contraintes supplémentaires entre les observations, diminuant d'autant le nombre de degrés de liberté.

Remarque 4.2.1 *Le nombre d'observations par classes ne doit pas être faible, Np_i doit être* ***supérieur*** *à 5, $\forall i = 1, 2, \ldots, k$. Dans le cas contraire, on regroupe deux ou plusieurs classes adjacentes de façon à réaliser cette condition. On tient compte de ce regroupement pour le nombre de degrés de liberté.*

6. Pour un risque de première espèce α, la région critique est définie pour

$$\boxed{\sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \geq \chi^2_{\nu,1-\alpha}}$$

d'où la règle de décision :

- χ^2 observé $< \chi^2_{\nu,1-\alpha}$, on décide H_0 et $X \rightsquigarrow L$.
- χ^2 observé $\geq \chi^2_{\nu,1-\alpha}$, on décide H_1 et X ne suit pas la loi L.

Exemple 4.2.1 *Loi uniforme.*
Une statistique relative aux résultats du concours d'entrée à une grande école fait ressortir les répartitions des candidats et des admis selon la profession des parents.

Profession des candidats	Nombre de candidats	Nombre d'admis
① Fonctionnaires et assimilés	2244	180
② Commerce, industrie	988	89
③ Professions libérales	575	48
④ Propriétaires rentiers	423	37
⑤ Propriétaires agricoles	287	13
⑥ Artisans, petits commerçants	210	18
⑦ Banque, assurance	209	17
Total	4936	402

Question : Tester l'hypothèse (risque $\alpha = 0,05$) selon laquelle la profession des parents n'a pas d'influence sur l'accès à cette grande école.

Il s'agit du test d'ajustement d'une distribution théorique, on considère les hypothèses :

- H_0 : la profession des parents n'a pas d'influence sur l'accès à cette grande école, la proportion des admis est constante pour toutes les professions soit $p = \dfrac{402}{4936} \simeq 0,0814$.
- H_1 : la profession des parents influe sur l'accès à cette grande école.

Sous H_0, le nombre d'admis pour la i-ième profession est pN_i.

i	N_i	n_i effectif observé	N_ip effectif théorique	$\frac{(n_i - N_ip)^2}{N_ip}$
1	2244	180	$\frac{2244 \times 402}{4936} \simeq 182,76$	$0,0416$
2	988	89	$\frac{988 \times 402}{4936} \simeq 80,47$	$0,9042$
3	575	48	$\frac{575 \times 402}{4936} \simeq 46,83$	$0,0293$
4	423	37	$\frac{423 \times 402}{4936} \simeq 34,45$	$0,1887$
5	287	13	$\frac{287 \times 402}{4936} \simeq 23,37$	$4,6050$
6	210	18	$\frac{210 \times 402}{4936} \simeq 17,10$	$0,0471$
7	209	17	$\frac{209 \times 402}{4936} \simeq 17,02$	$\simeq 0$
Total	4936	402	402	$5,8181$

Le χ^2 observé vaut $5,8181$. Le nombre de degrés de liberté est $7 - 1 = 6$. La table de l'annexe B fournit $\chi^2_{6;0,95} = 12,59$ donc χ^2 observé $< \chi^2_{6/0,95}$. On choisit H_0, ce qui signifie que la profession des parents n'a pas d'influence sur l'accès à cette grande école.

Exemple 4.2.2 *Loi binomiale.*
Supposons qu'on ait recueilli 300 boîtes contenant chacune trois ampoules. Dans chaque boîte, on compte le nombre d'ampoules défectueuses. On obtient les résultats suivants :

Nombre d'ampoules défectueuses x_i	Nombre de boîtes observées n_i
0	190
1	95
2	10
3	5
Total	300

Pour chaque ampoule testée, on peut observer deux états différents : l'ampoule est défectueuse ou non. Le nombre X d'ampoules défectueuses par boîte suit une loi binomiale de paramètres $n = 3$ et p.
Dans la distribution observée, le nombre d'ampoules défectueuses est de

$$0 \times 190 + 1 \times 95 + 2 \times 10 + 3 \times 5 = 130$$

soit 130 ampoules défectueuses sur un total de 900 ampoules. La proportion d'ampoules défectueuses est alors de $\frac{130}{900} \simeq 0,144$. Prenons $p = 0,15$ et réalisons alors le test suivant : soit X le nombre d'ampoules défectueuses par boîte

- $H_0 : X \rightsquigarrow \mathcal{B}(3; 0,15)$.

- H_1 : X ne suit pas cette loi binomiale.

On détermine les probabilités théoriques :

- $p_0 = p(\{X = 0/X \rightsquigarrow \mathcal{B}\}) = (0,85)^3 \simeq 0,6141$
- $p_1 = p(\{X = 1/X \rightsquigarrow \mathcal{B}\}) = C_3^1(0,15)(0,85)^2 \simeq 0,3251$
- $p_2 = p(\{X = 2/X \rightsquigarrow \mathcal{B}\}) = C_3^2(0,15)^2(0,85) \simeq 0,0574$
- $p_3 = p(\{X = 3/X \rightsquigarrow \mathcal{B}\}) = (0,15)^3 \simeq 0,0034$

On a le tableau

x_i	effectif observé n_i	p_i	effectif théorique Np_i
0	190	0,6141	184,23
1	95	0,3251	97,53
2	10	0,0574	17,22
3	5	0,0034	1,02
Total	$N = 300$	1	300

L'effectif théorique de la quatrième classe est faible : $1,02 < 5$. On effectue un regroupement de classes, les classes "2" et "3".

x_i	n_i	Np_i	$\frac{(n_i - Np_i)^2}{Np_i}$
0	190	184,23	0,18071
1	95	97,53	0,06563
2 ou 3	15	18,24	0,57553
Total	300	300	0,82187

Après le regroupement, le nombre de classes est 3, le nombre de degrés de liberté est $3 - 1 = 2$.
Au risque $\alpha = 0,01$ on a $\chi^2_{2;0,99} = 9,21$. Donc

$$\chi^2 \text{ observé} = 0,82187 < \chi^2_{2;0,99}.$$

On ne rejette pas H_0 au profit de H_1. On considère que le nombre d'ampoules défectueuses par boîte suit une loi binomiale de paramètres $n = 3$ et $p = 0,15$ au risque $\alpha = 0,01$.

Exemple 4.2.3 *Loi normale.*
On suppose que le rendement X (quintaux par hectares d'une parcelle de blé) suit une loi normale $\mathcal{N}(m, \sigma)$. L'observation du rendement de 1000 parcelles a donné les résultats suivants :

Rendement	[0; 10[	[10; 20[	[20; 30[	[30; 40[	[40; 50[	[50; 60[	[60; 70[	[70; 80[	[80; 90[
Nombre de parcelles	5	6	40	168	288	277	165	49	2

1. Déterminer la moyenne arithmétique et l'écart-type de la distribution observée.

- $\overline{x} = \dfrac{\sum\limits_i n_i x_i}{N} = 49,76$

- $\sigma'^2 = \dfrac{\sum_i n_i x_i^2}{N} - \overline{x}^2 = 164,5424$ donc $\sigma' \simeq 12,827$.

2. Vérifier pour un test du χ^2 avec un risque de $0,05$ si l'ajustement de la distribution observée à une loi normale $\mathcal{N}(n = 50, \sigma = 13)$ est acceptable.

   Les hypothèses du test du χ^2 sont les suivantes :
   - $H_0 : X \rightsquigarrow \mathcal{N}(50, 13)$
   - $H_1 : X$ ne suit pas $\mathcal{N}(50, 13)$

   On désigne par $[a_0; a_1[$, $[a_1; a_2[$,...,$[a_8; a_9[$ les classes et par x_1, x_2,...,x_9 les centres de ces classes. Sous H_0, $X \rightsquigarrow \mathcal{N}(50, 13)$ et $Z = \dfrac{X - 50}{13} \rightsquigarrow \mathcal{N}(0, 1)$, donc $p_i = p(\{X \in [a_{i-1}; a_i[\}) = \Pi(z_i) - \Pi(z_{i-1})$ avec $z_i = \dfrac{a_i - 50}{13}$ et $z_{i-1} = \dfrac{a_{i-1} - 50}{13}$. L'effectif théorique de la i-ème classe est $1000p_i$ et

$$\sum_i \frac{(n_i - Np_i)^2}{Np_i} \rightsquigarrow \chi^2_\nu$$

Classe $[x_{i-1}; x_i[$	n_i	z_i	$\Pi(z_i)$	p_i	Np_i	Np_i corrigé	n_i corrigé	$\frac{(n_i - Np_i)^2}{Np_i}$
$[0; 10[$	5	$-3,0769$	$0,001$	$0,0009$	$0,9$	$10,4$	11	$0,0346$
$[10; 20[$	6	$-2,3077$	$0,0105$	$0,0095$	$9,5$			
$[20; 30[$	40	$-1,5385$	$0,0620$	$0,0515$	$51,5$	$51,5$	40	$2,568$
$[30; 40[$	168	$-0,7692$	$0,2209$	$0,1589$	$158,9$	$158,9$	168	$0,5211$
$[40; 50[$	288	0	$0,5$	$0,2791$	$279,1$	$279,1$	288	$0,283$
$[50; 60[$	277	$0,7692$	$0,7791$	$0,2791$	$279,1$	$279,1$	277	$0,0158$
$[60; 70[$	165	$1,5385$	$0,9380$	$0,1589$	$158,9$	$158,9$	165	$0,234$
$[70; 80[$	49	$2,3077$	$0,9895$	$0,0515$	$51,5$	$51,5$	49	$0,1214$
$[80; 90[$	2	$3,0769$	$0,9990$	$0,0095$	$9,5$	$9,5$	2	$5,9211$
Total	1000	–	–	1	1000	1000	1000	$9,7$

   On effectue le regroupement des deux premières classes car $Np_i < 5$. Le χ^2 observé vaut $9,7$. Après le regroupement, il reste 8 classes, les deux paramètres de la loi normale sont donnés, le nombre de degrés de liberté est $\nu = (8 - 1) - 0 = 7$. À l'aide de la table, on obtient $\chi^2_{7;0,95} = 14,07$. Ainsi,

$$\chi^2 \text{ observé} < \chi^2_{7;0,95}.$$

   On choisit H_0, l'ajustement de la distribution observée à une loi normale $\mathcal{N}(50, 13)$ est acceptable.

Exemple 4.2.4 *Loi de Poisson.*
Souvent, lorsqu'on envisage une modèle pour un phénomène qu'on étudie, on ne spécifie pas complètement la loi qu'on considère. Supposons qu'on s'intéresse au nombre de voitures se présentant par minute à un poste de péage sur une autoroute. On peut se demander si cette variable aléatoire peut être modélisée par une loi de Poisson. On souhaite donc tester l'hypothèse fondamentale

$$H_0 : X \rightsquigarrow \mathcal{P}(\lambda)$$

contre l'hypothèse alternative

$$H_1 : X \text{ ne suit pas } \mathcal{P}(\lambda).$$

On ne précise pas la valeur du paramètre λ. On peut toutefois l'estimer à partir des données disponibles mais dans ce cas, $r = 1$. Le nombre de degrés sera alors $\nu = k - r - 1 = k - 2$.

On effectue 200 comptages au péage.

x_i	0	1	2	3	4	5	6	7	8	≥ 9	Total
n_i	6	15	40	42	37	30	10	12	8	0	200
$n_i x_i$	0	15	80	126	148	150	60	84	64	0	727

où x_i est le nombre de voitures par minute lors de la i-ième l'observation et n_i est l'effectif correspondant. Par exemple, $x_1 = 0$ et $n_1 = 6$ c'est-à-dire que lors de 6 observations, il y a 0 voiture.
La moyenne arithmétique de cette distribution observée est

$$\overline{x} = \frac{\sum\limits_i n_i x_i}{\sum\limits_i n_i} = \frac{727}{200} = 3,635 \simeq 3,5$$

On peut tester l'hypothèse

$$H_0 : X \rightsquigarrow \mathcal{P}(\lambda = 3,5).$$

x_i	n_i	p_i	Np_i	Np_i corrigé	n_i corrigé	$\frac{(n_i - Np_i)^2}{Np_i}$
0	6	0,0302	6,04	6,04	6	0,00026
1	15	0,1057	21,14	21,14	15	1,78333
2	40	0,1850	37	37	40	0,24324
3	42	0,2158	43,16	43,16	42	0,03118
4	37	0,1888	37,76	37,76	37	0,01530
5	30	0,1322	26,44	26,44	30	0,47933
6	10	0.0771	15,42	15,42	10	1,90508
7	12	0,0385	7,7	7,7	12	2,40130
8	8	0,0169	3,38	5,34	8	1,32502
≥ 9	0	0,0098	1,96			
Total	200	1	200	200	200	8,18404

On a $p_i = p(\{X = x_i / X \rightsquigarrow \mathcal{P}(3,5)\})$ donc

- $p_0 = p(\{X = 0 / X \rightsquigarrow \mathcal{P}(3,5)\}) = e^{-3,5} \simeq 0,0302$ et
- $p_1 = p(\{X = 1 / X \rightsquigarrow \mathcal{P}(3,5)\}) = e^{-3,5} 3,5 \simeq 0,1057$.

On a effectué le regroupement des deux dernières classes car l'effectif théorique y est inférieur à 5. Après ce regroupement, le nombre de classes est de 9. Le nombre de degrés de liberté est $9 - 1 - 1 = 7$. Au risque $\alpha = 0,01$, $\chi^2_{7;0,99} = 18,48$ donc

$$\chi^2 \text{ observé} = 8,18404 < \chi^2_{7;0,99}.$$

On ne rejette pas l'hypothèse H_0 et $X \rightsquigarrow \mathcal{P}(\lambda = 3,5)$ au risque $\alpha = 0,01$.

4.3 Comparaison de distributions observées. Test d'indépendance. Test d'homogénéité

4.3.1 Présentation du test

Le test du χ^2 est également utilisé pour tester l'indépendance de deux variables aléatoires. Considérons l'exemple suivant :

Exemple 4.3.1 On a posé à des parents la question suivante : "Dans cette liste, quelle est la qualité que vous souhaitez transmettre prioritairement à votre (vos) fille(s) ?"

	Femmes		Total	Hommes	Total
	Actives	Au foyer			
l'honnêteté	47	50	97	108	205
le sens du devoir	16	21	37	46	83
la patience	7	7	14	8	22
la coquetterie	4	3	7	6	13
l'esprit de famille	18	24	42	38	80
l'indépendance	21	12	33	30	63
le sens créatif	9	4	13	16	29
le dévouement	7	8	15	6	21
les qualités ménagères	11	20	31	24	55
la volonté	20	15	35	36	71
les bonnes manières	10	8	18	26	44
la réussite des études	16	12	28	34	62
Total	186	184	370	378	748

On peut par exemple se demander si les pères et les mères attachent la même importance aux qualités proposées. Autrement dit, on peut regarder si la qualité retenue est indépendante du sexe du parent. On veut tester les hypothèses

. H_0 : la qualité préférée est indépendante du sexe du parent interrogé,

. H_1 : cette qualité dépend du sexe.

On peut aussi se demander si la qualité retenue par les mères est indépendante du fait qu'elles travaillent ou non.

. H_0 : la qualité préférée par la mère est indépendante de son activité,

. H_1 : la qualité préférée par la mère n'est pas indépendante de son activité.

4.3.2 Construction du test

1. On considère deux variables aléatoires X et Y (le plus souvent qualitatives), X prenant les modalités x_1, x_2,..., x_k et Y les modalités y_1, y_2,...,y_p. On considère la loi de probabilité du couple (X, Y),

$$p_{ij} = p(\{X = x_i\} \cap \{Y = y_j\})$$

et les lois marginales de X et de Y

$$p_{i.} = p\{X = x_i\} = \sum_{j=1}^{p} p_{ij} \text{ et } p_{.j} = p\{Y = y_j\} = \sum_{i=1}^{k} p_{ij}$$

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_p	$p_{i.}$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1p}	$p_{1.}$
x_2	p_{21}	p_{22}	$\cdots$				$p_{2.}$
$\vdots$				$\vdots$			$\vdots$
x_i			$\cdots$	p_{ij}	$\cdots$		$p_{i.}$
$\vdots$				$\vdots$			$\vdots$
x_k							$p_{k.}$
$p_{.j}$	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.j}$	$\cdots$	$p_{.p}$	1

2. On veut tester l'indépendance des variables X et Y.
   . H_0 : X et Y sont indépendantes,
   . H_1 : X et Y ne sont pas indépendantes.

3. Sous H_0, $p_{ij} = p_{i.} \times p_{.j}$ donc
$$p(\{X = x_i\} \cap \{Y = y_j\}) = p\{X = x_i\}p\{Y = y_j\}$$
$$\forall i = 1, \ldots, k, \forall j = 1, \ldots, p$$
   On peut reformuler les hypothèses H_0 et H_1
   . $H_0 : p_{ij} = p_{i.} \times p_{.j},\ \forall i,\ \forall j$
   . $H_1 : \exists i,\ \exists j,\ p_{ij} \neq p_{i.} \times p_{.j}$

4. Comme les valeurs p_{ij}, $p_{i.}$ et $p_{.j}$ sont inconnues, on les estime à l'aide des données relatives aux variables X et Y. Ces N données sont le plus souvent présentées sous la forme d'un tableau de contingence.

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_p	effectif de X : $n_{i.}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1.}$
x_2	n_{21}	n_{22}					$n_{2.}$
$\vdots$				$\vdots$			$\vdots$
x_i			$\cdots$	n_{ij}	$\cdots$		$n_{i.}$
$\vdots$				$\vdots$			$\vdots$
x_k							$n_{k.}$
effectif de Y : $n_{.j}$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.p}$	N

   Soient
   . n_{ij} l'effectif de $\{X = x_i\} \cap \{Y = y_j\}$,
   . $n_{i.} = \sum_{j=1}^{p} n_{ij}$ l'effectif de $\{X = x_i\}$,
   . $n_{.j} = \sum_{i=1}^{k} n_{ij}$ l'effectif de $\{Y = y_j\}$.

Parfois, on dispose du tableau des fréquences observées f_{ij} (fréquence de l'événement $\{X = x_i\} \cap \{Y = y_j\}$) où

$$f_{ij} = \frac{n_{ij}}{N},\ f_{i.} = \frac{n_{i.}}{N} \text{ et } f_{.j} = \frac{n_{.j}}{N}.$$

On estime les probabilités p_{ij}, $p_{i.}$ et $p_{.j}$ respectivement par

. $\hat{p}_{ij} = f_{ij} = \dfrac{n_{ij}}{N}$

. $\hat{p}_{i.} = f_{i.} = \dfrac{n_{i.}}{N}$

. $\hat{p}_{.j} = f_{.j} = \dfrac{n_{.j}}{N}$

5. Sous H_0, $p_{ij} = p_{i.} \times p_{.j}$ avec les estimations $\dfrac{n_{ij}}{N} \simeq \dfrac{n_{i.}}{N} \times \dfrac{n_{.j}}{N}$ soit $n_{ij} \simeq \dfrac{n_{i.}n_{.j}}{N}$, l'observation n_{ij} doit être proche d'une quantité notée $v_{ij} = \dfrac{n_{i.}n_{.j}}{N}$ appelée **effectif théorique**.

   Cet effectif théorique représente l'effectif qu'on doit approximativement observer si l'hypothèse d'indépendance est vraie.

   On construit une mesure de l'écart entre l'effectif observé et l'effectif théorique

$$\chi^2 = \sum_{i=1}^{k} \left[\sum_{j=1}^{p} \frac{(n_{ij} - v_{ij})^2}{v_{ij}} \right]$$

   Si l'hypothèse d'indépendance est vraie, cette quantité doit être petite.

   On peut montrer que si les variables X et Y sont indépendantes alors pour N grand, $X^2 \rightsquigarrow \chi^2_\nu$ avec $\nu = (k-1)(p-1)$ et on utilise la règle de décision usuelle.

- χ^2 observé $\geq \chi^2_{\nu,1-\alpha}$, H_0 est rejetée au profit de H_1, les variables X et Y ne sont pas indépendantes,
- χ^2 observé $< \chi^2_{\nu,1-\alpha}$, l'hypothèse H_0 est acceptée, c'est-à-dire que les variables X et Y sont indépendantes.

4.3.3 Exemples

1. Reprenons l'exemple 4.3.1 :

   On a les hypothèses

   . H_0 : la qualité préférée est indépendante du sexe du parent interrogé.

   . H_1 : la qualité préférée n'est pas indépendante du sexe du parent interrogé.

Sexe \ Qualité	1	2	3	4	5	6	7	8	9	10	11	12	$n_{i.}$
F	97	37	14	7	42	33	13	15	31	35	18	28	370
H	108	46	8	6	38	30	16	6	24	36	26	34	378
$n_{.j}$	205	83	22	13	80	63	29	21	55	71	44	62	$748 = N$

   On constitue le tableau des effectifs théoriques $v_{ij} = \dfrac{n_{i.}n_{j.}}{N}$, par exemple $v_{11} = \dfrac{205 \times 370}{748} \simeq 101,40$, $v_{21} = \dfrac{205 \times 378}{748} \simeq 103,60$

n_{ij}	v_{ij}	$\frac{(n_{ij}-v_{ij})^2}{v_{ij}}$	n_{ij}	v_{ij}	$\frac{(n_{ij}-v_{ij})^2}{v_{ij}}$
97	101,40	0,19093	16	14,66	0,12248
108	103,6	0,18687	15	10,39	2,04544
37	41,06	0,40145	6	10,61	2,00303
46	41,94	0,39303	31	27,21	0.52790
14	10,88	0,89471	24	27,79	0,51688
8	11,12	0,87540	35	35,12	0,00041
7	6,43	0,05053	36	35,88	0,00040
6	6,57	0,04945	18	21,76	0,64971
42	39,57	0,14923	26	22,24	0,63568
38	40,43	0,14605	28	30,67	0,28244
33	31,16	0,10865	34	31,33	0,22754
30	31,84	0,10633	Total = 748	748	10,63976
13	14,34	0,12522			

Le χ^2 observé est $10,63976$, le nombre de degrés de liberté est

$$(2-1)(12-1) = 11.$$

La table fournit

$$\chi^2_{11,0.95} = 19,68$$

Comme χ^2 observé $< \chi^2_{11;0,95}$, on ne rejette pas H_0 au risque de 5%. On admet l'indépendance entre la qualité préférée pour la fille et le sexe du parent interrogé et ceci au risque $\alpha = 0,05$.

2. Avec les mêmes données on peut aussi tester les hypothèses suivantes :
   . H_0 : la qualité préférée par la mère est indépendante de son activité,
   . H_1 : la qualité préférée par la mère n'est pas indépendante de son activité.

Z \ Y	1	2	3	4	5	6	7	8	9	10	11	12	$n_{i.}$
Femmes actives	47	16	7	4	18	21	9	7	11	20	10	16	186
Femmes au foyer	50	21	7	3	24	12	4	8	20	15	8	12	184
$n_{.j}$	97	37	14	7	42	33	13	15	31	35	18	28	$370 = N$

On détermine les effectifs théoriques v_{ij}, par exemple $v_{11} = \dfrac{97 \times 186}{370}$, puis l'indicateur d'écart, χ^2 observé et on conclut.

3. On peut aussi tester les hypothèses
   . H_0 : la qualité préférée est indépendante du sexe du parent actif interrogé,
   . H_1 : la qualité préférée n'est pas indépendante du sexe du parent actif interrogé.

T \ Y	1	2	3	4	5	6	7	8	9	10	11	12	$n_{.j}$
Femmes actives	47	16	7	4	18	21	9	7	11	20	10	16	186
Hommes	108	46	8	6	38	30	16	6	24	36	26	34	378
$n_{i.}$	155	62	15	10	56	51	25	13	35	56	36	50	$564 = N$

On détermine les effectifs théoriques v_{ij}, par exemple $v_{11} = \dfrac{155 \times 186}{564}$, puis l'indicateur d'écart χ^2 observé et on conclut.

Remarque 4.3.1 Il est nécessaire que les v_{ij} soient supérieurs à 5, si ce n'est pas le cas, on regroupe des lignes ou des colonnes adjacentes.

4.4 Comparaison de proportions

Les précédents tests du χ^2 peuvent également être utilisés pour
. comparer une proportion observée à une proportion théorique,
. comparer deux proportions.

4.4.1 Comparaison d'une proportion observée à une proportion théorique

On souhaite comparer la proportion p_0 d'individus possédant le caractère C, observée sur un échantillon de taille N, à la proportion théorique p.

La variable qualitative est à deux modalités
. les effectifs observés sont Np_0, $N(1-p_0)$
. les effectifs théoriques sont Np, $N(1-p)$

L'indicateur d'écart est alors :

$$\chi^2_{\text{observé}} = \frac{(Np_0 - Np)^2}{Np} + \frac{[N(1-p_0) - N(1-p)]^2}{N(1-p)}$$
$$\Leftrightarrow \chi^2_{\text{observé}} = \frac{N^2(p_0-p)^2}{Np} + \frac{N^2(p_0-p)^2}{N(1-p)}$$
$$\Leftrightarrow \chi^2_{\text{observé}} = N(p_0-p)^2\left[\frac{1}{p} + \frac{1}{1-p}\right] = \frac{N(p_0-p)^2}{p(1-p)}$$
$$\Leftrightarrow \chi^2_{\text{observé}} = \frac{N(p_0-p)^2}{p(1-p)}$$

On considère les hypothèses
. $H_0 : p = p_0$, le nombre de degrés de liberté est 1,
. $H_1 : p \neq p_0$.

Si $\chi^2_{obs} \geq \chi^2_{1;1-\alpha}$, H_0 est rejetée au profit de H_1.

Exemple 4.4.1 Pour l'année 1967, le pourcentage des candidats reçus au baccalauréat Mathématiques Élémentaires a été de $p = 64\%$. M$^{\text{r}}$ X a présenté 40 candidats, 31 furent déclarés reçus. Peut-on dire que M$^{\text{r}}$ X prépare mieux les candidats à l'examen ?

La proportion de reçus dans la classe de M$^{\text{r}}$ X est de $p_0 = \dfrac{31}{40} = 0,775$. Soient
. $H_0 : p = p_0$, les deux proportions sont identiques, M$^{\text{r}}$ X ne prépare pas mieux ses candidats à l'examen,

. $H_1 : p \neq p_0$.

L'indicateur d'écart est $\chi^2 = \dfrac{N(p_0 - p)^2}{p(1-p)}$. On a par conséquent

$$\chi^2_{obs} = 40\frac{(0,64 - 0,775)^2}{0,64 \times 0,36} = 3,164062$$

Le nombre de degrés de liberté est 1. La table nous donne $\chi^2_{1;0,95} = 3,841$ donc

$$\chi^2_{obs} < \chi^2_{1;0,95},$$

on ne rejette pas H_0.

Par conséquent, la différence observée entre les proportions n'est pas significative et M$^{\text{r}}$ X ne prépare pas mieux ses candidats à l'examen au risque $\alpha = 5\%$. On a également $\chi^2_{1;0,90} = 2,71$ donc

$$\chi^2 \text{ observé} \geq \chi^2_{1;0,90},$$

on rejette H_0 au risque $\alpha = 10\%$.

4.4.2 Comparaison de deux proportions

S'il s'agit de comparer deux proportions p_1 et p_2 observées sur deux échantillons de taille N_1 et N_2 prélevés dans les populations P_1 et P_2, on construit le tableau de contingence suivant :

variable \ population	P_1	P_2	Total
C	N_1p_1	N_2p_2	$N_1p_1 + N_2p_2$
$\overline{C}$	$N_1(1-p_1)$	$N_2(1-p_2)$	$N_1(1-p_1) + N_2(1-p_2)$
Total	N_1	N_2	$N_1 + N_2$

On procède ensuite comme dans les cas précédents.

Exemple 4.4.2 Une enquête sur la population active de la région parisienne révèle que sur 2400 hommes et 1600 femmes interrogées, les cadres moyens sont au nombre de 314 et 182 respectivement. Peut-on dire que l'accession à cette catégorie est égalitaire pour les deux sexes ?

. H_0 : homogénéité ou accession égalitaire,

. H_1 : accession inégalitaire.

La fréquence des cadres chez les hommes est $f_H = \dfrac{314}{2400}$,

celle des femmes est $f_F = \dfrac{182}{1600}$.

On compare les deux proportions f_H et f_F observées sur deux échantillons des populations masculine et féminine. On construit le tableau de contingence

sexe \ activité	C	$\overline{C}$	Total
H	314	2086	2400
F	182	1418	1600
Total	496	3504	4000

On a

$$v_{11} = \frac{2400 \times 496}{4000} = 297,6, \quad v_{12} = \frac{2400 \times 3504}{4000} = 2102,4$$
$$v_{21} = \frac{1600 \times 496}{4000} = 198,4, \quad v_{12} = \frac{2400 \times 3504}{4000} = 1401,6$$

L'indicateur d'écart est $\chi^2 = \sum \frac{(n_{ij} - v_{ij})^2}{v_{ij}}$ et χ^2 observé $= 2.5792$. Le nombre de degrés de liberté est $(2-1) \times (2-1) = 1$. Pour un risque α de 5%, $\chi^2_{1;0,95} = 3,84$ et

$$\chi^2 \text{ observé} < \chi^2_{1;0,95}.$$

On décide H_0, l'accession à la catégorie cadre est égalitaire pour les deux sexes.

4.5 Exercices

Exercice 27 En lançant successivement 60 fois un dé, un joueur obtient les résultats suivants :

Faces x_i	1	2	3	4	5	6
Effectifs n_i	15	7	4	11	6	17

Le dé est-il truqué ?

Exercice 28 On a enregistré le nombre X de clients entrant dans un magasin en 1 minute. On a obtenu le tableau suivant :

Nombre de clients x_i	Nombre de minutes où $X = x_i$ (où il est entré x_i clients)
0	23
1	75
2	68
3	51
4	30
5	10
plus de 5	7

Peut-on admettre que les arrivées sont régies par une loi de Poisson de paramètre $m = 2$ (au seuil $\alpha = 0,05$) ?

Exercice 29 Une enquête sur les chiffres d'affaires mensuels de 103 magasins de détail a donné les résultats suivants (en milliers d'euros) :

Classes de chiffres d'affaires	Centres de classes	Nombre de magasins
$5,5$ à moins de $6,5$	6	2
$6,5$ à moins de $7,5$	7	3
$7,5$ à moins de $8,5$	8	12
$8,5$ à moins de $9,5$	9	27
$9,5$ à moins de $10,5$	10	23
$10,5$ à moins de $11,5$	11	15
$11,5$ à moins de $12,5$	12	12
$12,5$ à moins de $13,5$	13	5
$13,5$ à moins de $14,5$	14	2
$14,5$ à moins de $15,5$	15	2

Peut-on considérer que l'échantillon est tiré d'une loi normale ?

Exercice 30 On a étudié le nombre de garçons dans 1883 familles de 7 enfants. Les résultats sont présentés en fonction du nombre x_i de garçons, rangés de 0 à 7 :

Nombre de garçons x_i	Effectif des familles n_i
0	27
1	111
2	287
3	480
4	529
5	304
6	126
7	19
Total	1883

Peut-on admettre au seuil de 5% que le nombre x_i de garçons par famille obéisse à une loi binomiale ? Laquelle ?

Exercice 31 Une entreprise achète une machine dont le fabricant assure que 95% des pièces qu'elle permet d'usiner satisfont aux normes exigées. Sur un échantillon de 100 pièces, 9 ne satisfont pas aux normes.
Peut-on admettre, au seuil de 5% que les caractéristiques réelles de la machine ne correspondent pas aux garanties du fournisseur ? (Résoudre ce test de comparaison d'une fréquence à une valeur standard à l'aide du test du χ^2).

Exercice 32 On fait passer une épreuve à deux groupes, l'un de 50 personnes et l'autre de 30 personnes.
Dans le premier groupe, il y a 42 succès à l'épreuve.
Dans le second groupe, il y a 18 succès à l'épreuve.
L'hypothèse à tester est que les résultats sont les mêmes pour les deux populations d'où l'on a extrait les deux échantillons.

n_i	Succès	Échecs	Total
1^{er} groupe	42	8	$N_1 = 50$
$2^{\text{ème}}$ groupe	18	12	$N_2 = 30$
Total	60	20	$N = 80$

Exercice 33 Il s'agit de rechercher chez 160 adolescents asthmatiques de la région Aquitaine un lien entre l'intensité de l'asthme et la présence ou l'absence d'éczéma durant l'année d'étude ou antérieurement. On a établi le tableau suivant :

Asthme \ Eczéma	Présent	Passé	Jamais	Total
Fort	22	28	28	78
Faible	12	33	37	82
Total	34	61	65	160

On souhaite examiner l'hypothèse

H_0 : "l'intensité de l'asthme et la présence d'eczéma sont indépendantes"

1. Déterminer dans un tableau les effectifs théoriques à $0,1$ au plus près.
2. Préciser le nombre de degrés ν de liberté.
3. Calculer le χ^2.
4. Quelle conclusion peut-on déduire au seuil de signification $0,95$ (avec pour risque d'erreur 5%) ?

Exercice 34 On cherche à savoir si les résultats électoraux d'un village du Sud-Est de la France correspondent aux résultats nationaux pour une élection où les cinq candidats A,B,C,D et E se présentent :

	A	B	C	D	E
Nombre de votants du village	475	364	1968	1633	560
Résultats nationaux en %	$10,3$	$7,1$	$40,8$	$32,1$	$9,7$

1. Faire un tableau comportant les résultats des 5000 votants du village et les résultats théoriques si l'hypothèse de ressemblance est vérifiée.
2. Calculer le χ^2 associé.
3. Interpréter en utilisant la table du χ^2 au risque de 1%.

Exercice 35 Au cours d'une élection, il y a 5 candidats en présence.
Une enquête d'opinion est faite sur un échantillon de 875 sujets, 410 hommes et 465 femmes.

Candidats	A	B	C	D	E	Totaux
Hommes	25	75	105	130	75	410
Femmes	147	116	17	80	105	465
Total	172	191	122	210	180	875

1. Déterminer à l'aide d'un tableau les effectifs théoriques à $0,01$ au plus près.
2. Préciser le nombre de degrés de liberté ν.
3. Calculer le χ^2.

4. Quelle conclusion peut-on déduire au risque de 1% ?

Exercice 36 On veut savoir si le rythme cardiaque est différent entre la population adulte urbaine de l'agglomération bordelaise et la population rurale de l'Aquitaine.
Pour cela, on choisit un échantillon aléatoire de 150 personnes adultes de l'agglomération bordelaise et un échantillon aléatoire de 400 personnes adultes vivant dans des communes rurales de moins de 5000 habitants de la région Aquitaine. Ces observations conduisent au tableau suivant :

Population \ Rythme R	$R < 65$	$65 \leq R < 70$	$70 \leq R < 75$	$75 \leq R < 80$	$80 \leq R$
Urbaine	6	21	53	60	10
Rurale	32	80	163	112	13

On considère l'hypothèse :

H_0 : "Les deux caractères *rythme cardiaque* et *vie urbaine ou rurale* sont indépendants".

1. Déterminer dans un tableau les effectifs théoriques à $0,01$ au plus près.
2. Préciser le nombre de degrés ν de liberté.
3. Calculer le χ^2.
4. Quelle conclusion peut-on déduire au seuil de signification $0,95$ (avec pour risque d'erreur 5%) ?

Exercice 37 Trois groupes d'étudiants de même niveau, constitués indépendamment les uns des autres, sont soumis à des méthodes pédagogiques différentes :
- Le premier groupe reçoit un enseignement traditionnel.
- Le second groupe bénéficie d'un renforcement pédagogique dans le cadre des méthodes traditionnelles.
- Le troisième groupe expérimente une nouvelle méthode pédagogique.

Les étudiants des trois groupes passent en fin d'année un même examen. On enregistre les résultats suivants :
- Sur le premier groupe d'effectif égal à $N_1 = 115$, $K_1 = 82$ étudiants sont admis.
- Sur le deuxième groupe d'effectif égal à $N_2 = 90$, $K_1 = 71$ étudiants sont admis.
- Sur le troisième groupe d'effectif égal à $N_3 = 35$, $K_3 = 26$ étudiants sont admis.

Peut-on admettre que l'une des trois méthodes est plus efficace que les autres (au seuil $\alpha = 0,05$) ?

Exercice 38 Sur un échantillon de 200 ménages choisis au hasard, on a étudié la propension moyenne à épargner (variable Y) en fonction du revenu disponible (variable X).
Pour la variable X, on a distingué 3 classes :
- x_1 : faibles revenus,
- x_2 : revenus intermédiaires,
- x_3 : revenus élevés.

De même, les taux d'épargne ont été classés en 3 niveaux :
- y_1 : faibles taux,
- y_2 : taux intermédiaires,
- y_3 : taux élevés.

Les résultats sont présentés dans la table de contingence :

Revenus \ Épargne	y_1	y_2	y_3	Total $N_{i.}$
x_1	53	14	6	73
x_2	15	58	8	81
x_3	7	10	29	46
Total $N_{.j}$	75	82	43	200

Existe t-il une relation entre le taux d'épargne et le niveau de revenu disponible ?

Exercice 39 Une étude est menée dans une petite université sur l'absentéisme des étudiants. On aimerait savoir si certaines plages horaires sont plus propices à une absence aux cours qu'une autre. Pour cela, on a relevé, au cours d'un mois, le nombre d'absences d'étudiants aux cours d'une petite composante à différents moments de la journée.

Heures de la journée	Nombre d'étudiants absents
8-10	25
10-12	15
13-15	18
15-17	32

En considérant cet échantillon tiré au hasard, peut-on dire, au seuil de 5%, que les absences des étudiants aux cours se répartissent uniformément tout au long de la journée ?

Exercice 40 Une entreprise fabriquant des produits alimentaires sucrés veut élargir sa gamme de barres de céréales en en lançant une nouvelle sur le marché. Le directeur du marketing décide de faire une enquête de goût en faisant tester ce nouveau produit à 200 personnes. Le test a lieu en aveugle, et les personnes ont donc à se prononcer sur leur préférence concernant la nouvelle barre et quatre autres barres de céréales concurrentes. Les produits étant appelés A (la nouvelle barre), B, C, D et E, les résultats du test dont les suivants :

Barres	A	B	C	D	E
Nombre de préférences	40	35	55	40	30

Au seuil 5%, peut-on dire au vu des résultats d'échantillon que la nouvelle barre a meilleur goût que les autres ?

Exercice 41 Un responsable qualité d'une entreprise fabriquant de l'appareillage électronique a mesuré la durée de vie de 60 dispositifs électroniques d'un même type. Il a obtenu les résultats suivants :

Durée de vie (en heures)	Nombre de dispositifs
$[250; 270[$	3
$[270; 290[$	5
$[290; 310[$	15
$[310; 330[$	22
$[330; 350[$	13
$[350; 370[$	2

Les donnée permettent-elles, au seuil de 5%, de penser que la durée de vie d'un dispositif électronique de ce type est distribué selon une loi normale ?

Exercice 42 On a observé le nombre de défauts de pièces de tissu traitées par un teinturier. Les résultats de 50 observations sont reproduits dans le tableau ci-dessous (par exemple, 8 des 50 pièces présentaient 3 défauts). Des études antérieures avaient permis de faire l'hypothèse que le nombre de défauts par pièce pouvait être considéré comme une variable aléatoire X obéissant à une loi de Poisson. Les observations permettent-elles de confirmer cette hypothèse, au seuil de 5% ?

x_i	0	1	2	3	4	5
n_i	6	14	16	8	4	2