Université Claude Bernard Lyon 1
IREM de Lyon - Département de mathématiques
Stage ATSM - Août 2010

Cours de probabilités et statistiques

A. Perrut

contact : Anne.Perrut@univ-lyon1.fr

Table des matières

Chapitre 1

Le modèle probabiliste

1.1 Introduction

Les probabilités vont nous servir à modéliser une **expérience aléatoire**, c'est-à-dire un phénomène dont on ne peut pas prédire l'issue avec certitude, et pour lequel on décide que le dénouement sera le fait du hasard.
Exemples :
- l'enfant à naître sera une fille,
- l'équipe de l'OL va battre l'OM lors du prochain match qui les opposera,
- le dé va faire un nombre pair.

La première tâche qui vous attend est de décrire les différentes issues possibles de cette expérience aléatoire. Puis on cherche à associer à chacune de ces **éventualités** un nombre compris entre 0 et 1 qui mesure la chance qu'elles ont de se réaliser. Comment interpréter/fixer ce nombre, appelé probabilité ? Il existe plusieurs manières de voir.

- Proportion :

On lance un dé. Quelle est la probabilité de A="obtenir un chiffre pair" ? Chaque face du dé a la même chance, et il y en a 6. Quant aux chiffres pairs, ils sont 3. D'où, intuitivement, $P(A) = \frac{3}{6} = 1/2$.

- Fréquence :

Un enfant est attendu. Quelle est la probabilité que ce soit une fille ? On a observé un grand nombre de naissances. Notons k_n le nombre de filles nées en observant n naissances. Alors

$$P(\text{fille}) = \lim_{n \to +\infty} \frac{k_n}{n}$$

mais cette limite a-t-elle un sens ?

- Opinion :

Quelle est la probabilité pour que l'équipe de Tunisie gagne la coupe d'Afrique des nations ? pour que l'OL soit championne de France ? Dans ce cas, on ne peut pas rejouer le même match dans les mêmes conditions plusieurs fois. On peut considérer les qualités des joueurs, des entraîneurs, les résultats de la saison... Mais le choix de la probabilité est forcément subjectif.

Attention aux valeurs des probabilités ! Elles sont choisies de manière arbitraire par le modélisateur et il faut les manipuler avec soin.

1.2 Espace des possibles, événements

On étudie une expérience aléatoire. L'**espace des possibles** ou **univers** décrit tous les résultats possibles de l'expérience. Chacun de ces résultats est appelé **événement élémentaire**. On note souvent l'espace des possibles Ω et un résultat élémentaire ω. Un **événement** est un sous-ensemble de Ω, ou une réunion d'événements élémentaires. On dit qu'un événement est réalisé si un des événements élémentaires qui le constitue est réalisé. Les événements sont des ensembles, représentés souvent par des lettres capitales.

Exemples :
- Match OL-OM : $\Omega = \{$OL gagne, OM gagne, match nul$\}$. Donc Ω est composé de trois événements élémentaires. On peut considérer par exemple l'événement qui correspond à "Lyon ne gagne pas".
- On lance un dé : $\Omega = \{1, 2, ..., 6\}$. On peut s'intéresser à l'événement A="on obtient un chiffre pair", ie $A = \{2, 4, 6\}$.
- On lance deux dés : $\Omega = \{1, ..., 6\} \times \{1, ..., 6\} = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$. Ici, un événement élémentaire ω est un couple (i, j), où i représente le résultat du premier dé et j celui du second.
- On lance trois fois une pièce de monnaie. Les événements élémentaires vont décrire le plus précisément possible le résultat de cette expérience. Donc un événement élémentaire ω est un triplet (r_1, r_2, r_3) qui donne les résultats des trois lancers (dans l'ordre). L'événement B : "on obtient pile au deuxième lancer" est

$$B = \{(f, p, f), (f, p, p), (p, p, f), (p, p, p)\}$$

L'événement B est réalisé si on obtient l'un des événements élémentaires listés ci-avant. Il n'est parfois pas nécessaire de connaître tous ces détails. On pourra choisir : ω représente le nombre de "face" obtenus. Alors, $\Omega = \{0, 1, 2, 3\}$. Le modèle est beaucoup plus simple, mais ne permet pas de décrire des événements tels que B.

Il existe un vocabulaire propre aux événements, différent du vocabulaire ensembliste.

notations	vocabulaire ensembliste	vocabulaire probabiliste
Ω	ensemble plein	événement certain
$\emptyset$	ensemble vide	événement impossible
ω	élément de Ω	événement élémentaire
A	sous-ensemble de Ω	événement
$\omega \in A$	ω appartient à A	ω réalise A
$A \subset B$	A inclus dans B	A implique B
$A \cup B$	réunion de A et B	A ou B
$A \cap B$	intersection de A et B	A et B
A^c ou $\overline{A}$	complémentaire de A	événement contraire de A
$A \cap B = \emptyset$	A et B disjoints	A et B incompatibles

1.3 Probabilité

On se limite dans ce cours à étudier les univers dénombrables. La **probabilité** d'un événement est une valeur numérique qui représente la proportion de fois où l'événement va se réaliser, quand on répète l'expérience dans des conditions identiques. On peut déduire de cette définition qu'une probabilité doit être entre 0 et 1 et que la probabilité d'un événement est la somme des probabilités de chacun des événements élémentaires qui le constituent. Enfin, la somme des probabilités de tous les éléments de Ω est 1.
Important : rappelons qu'un événement n'est rien d'autre qu'une partie de Ω. Une probabilité associe à chaque événement un nombre entre 0 et 1. Il s'agit donc d'une application de l'ensemble des parties de Ω, noté $\mathcal{P}(\Omega)$, dans $[0, 1]$.
Exemple : soit $\Omega = \{0, 1, 2\}$. Construisons $\mathcal{P}(\Omega)$.

$$\mathcal{P}(\Omega) = \Big\{\emptyset, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \Omega\Big\}$$

Définition 1 *Une probabilité est une application sur $\mathcal{P}(\Omega)$, l'ensemble des parties de Ω, telle que :*
- $0 \leq P(A) \leq 1$, *pour tout événement* $A \subset \Omega$
- $P(A) = \sum_{\omega \in A} P(\omega)$, *pour tout événement* A
- $P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = 1$

Que signifie "un événement A a pour probabilité..." ?
0.95 : A va très probablement se produire.
0.03 : A a très peu de chance d'être réalisé.
4.0 : incorrect.
-2 : incorrect.
0.4 : A va se produire dans un peu moins de la moitié des essais.
0.5 : une chance sur deux.
0 : aucune chance que A soit réalisé.

De la définition, on peut facilement déduire la proposition suivante, fort utile pour faire quelques calculs :

Proposition 2 *Soient A et B deux événements.*
1) Si A et B sont incompatibles, $P(A \cup B) = P(A) + P(B)$.
2) $P(A^c) = 1 - P(A)$.
3) $P(\emptyset) = 0$.
5) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

preuve : 1) immédiat d'après le second point de la définition d'une probabilité.
2) Comme A et A^c sont incompatibles, $1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$.
3) $P(\emptyset) = 1 - P(\emptyset^c) = 1 - P(\Omega) = 0$.
4) La technique est très souvent la même pour calculer la probabilité d'une réunion d'ensembles : on écrit cette réunion comme une union d'ensembles incompatibles, puis on utilise le 1). Ici, on écrit $A \cup B = A \cup (B \cap A^c)$ et on obtient : $P(A \cup B) = P(A) + P(A \cup (B \cap A^c))$. Puis on écrit $B = (B \cap A) \cup (B \cap A^c)$ pour déduire $P(B) = P(B \cap A) + P(B \cap A^c)$. En rassemblant ces deux égalités, on obtient la proposition.

Signalons une définition plus générale de probabilité, valable pour des espaces des possibles non dénombrables.

Définition 3 *Soit une expérience aléatoire et Ω l'espace des possibles associé. Une probabilité sur Ω est une application, définie sur l'ensemble des événements, qui vérifie :*
- axiome 1 : $0 \leq P(A) \leq 1$, pour tout événement A
- axiome 2 : pour toute suite d'événements $(A_i)_{i \in \mathbb{N}}$, deux à deux incompatibles,

$$P\Big(\bigcup_{i \in \mathbb{N}} A_i\Big) = \sum_{i \in \mathbb{N}} P(A_i)$$

- axiome 3 : $P(\Omega) = 1$

NB : les événements $(A_i)_{i \in \mathbb{N}}$ sont deux à deux incompatibles, si pour tous $i \neq j$, $A_i \cap A_j = \emptyset$.

Exemple important : probabilité uniforme

Soit Ω un ensemble fini. Il arrive, comme quand on lance un dé équilibré, que les événements élémentaires ont tous la même probabilité. On parle alors d'événements élémentaires équiprobables. Notons p la probabilité de chaque événement élémentaire. Alors

$$1 = P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = \sum_{\omega \in \Omega} p = p \times \mathrm{card}(\Omega)$$

D'où $p = P(\omega) = \dfrac{1}{\mathrm{card}(\Omega)}$, pour tout ω. La probabilité ainsi définie sur l'ensemble Ω s'appelle probabilité uniforme. La probabilité d'un événement A se calcule facilement :

$$P(A) = \sum_{\omega \in A} P(\omega) = \frac{\mathrm{card}(A)}{\mathrm{card}(\Omega)}$$

Attention ! Cette formule n'est valable que lorsque les événements élémentaires sont bien équiprobables. Dans ce cas, il suffit de savoir calculer le cardinal des ensembles considérés pour calculer les probabilités.

Un rappel des techniques de dénombrement est disponible à l'annexe A.

On est maintenant en mesure de **modéliser** des expériences aléatoires simples, c'est-à-dire :
- choisir Ω,
- choisir une probabilité sur Ω en justifiant ce choix.

Attention, pour décrire une probabilité, il faut donner $P(A)$ pour tout $A \subset \Omega$. Ou alors, on peut plus simplement donner $P(\omega)$ pour tout $\omega \in \Omega$. Le lecteur déduira $P(A)$ pour tout A d'après la définition d'une probabilité.

1.4 Indépendance et conditionnement

Exemple 4 *Quelle est la probabilité d'avoir un cancer du poumon ?*
Information supplémentaire : vous fumez une vingtaine de cigarettes par jour. Cette information va changer la probabilité.

L'outil qui permet cette mise à jour est la probabilité conditionnelle.

Définition 5 *Étant donnés deux événements A et B, avec $P(A) > 0$, on appelle probabilité de B conditionnellement à A, ou sachant A, la probabilité notée $P(B|A)$ définie par*

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

On peut écrire aussi $P(A \cap B) = P(B|A)P(A)$.

Utilisation 1 : quand $P(A)$ et $P(A \cap B)$ sont faciles à calculer, on peut en déduire $P(B|A)$.
Utilisation 2 : Quand $P(B|A)$ et $P(A)$ sont faciles à trouver, on peut obtenir $P(A \cap B)$.

De plus, la probabilité conditionnelle sachant A, $P(.|A)$, est une nouvelle probabilité et possède donc toutes les propriétés d'une probabilité.

Exemple 6 *Une urne contient r boules rouges et v boules vertes. On en tire deux, l'une après l'autre (sans remise). Quelle est la probabilité d'avoir deux boules rouges ?*
Choisissons Ω qui décrit les résultats de l'expérience précisément.

$$\Omega = \{rouge, verte\} \times \{rouge, verte\}$$

Un événement élémentaire est un couple (x, y) où x est la couleur de la première boule tirée et y la couleur de la seconde.
Soit A l'événement "la première boule est rouge" et B l'événement "la seconde boule est rouge".

$$P(A \cap B) = P(B|A)P(A) = \frac{r-1}{r+v-1} \cdot \frac{r}{r+v}$$

Proposition 7 (Formule des probabilités totales) *Soit A un événement tel que $0 < P(A) < 1$. Pour tout événement B, on a*

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

preuve : Comme $A \cup A^c = \Omega$, $P(B) = P(B \cap (A \cup A^c)) = P((B \cap A) \cup (B \cap A^c))$. Or $B \cap A$ et $B \cap A^c$ sont incompatibles. On en déduit

$$P(B) = P(B \cap A) + P(B \cap A^c)$$

La définition de la probabilité conditionnelle permet de conclure. □

Exemple 6 (suite) : quelle est la probabilité pour que la seconde boule tirée soit rouge ? On garde le même formalisme.

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|A^c)P(A^c) \\ &= \frac{r-1}{r+v-1} \cdot \frac{r}{r+v} + \frac{r}{r+v-1} \cdot \frac{v}{r+v} \\ &= \frac{r}{r+v} \end{aligned}$$

Définition 8 *Soit $(A_i)_{i \in I}$ une famille d'événements. On l'appelle partition de Ω si elle vérifie les deux conditions :*
(i) $\cup_{i \in I} A_i = \Omega$
(ii) les A_i sont deux à deux incompatibles : pour tous $i \neq j$, $A_i \cap A_j = \emptyset$.

Proposition 9 (Formule des probabilités totales généralisée) *Soit $(A_i)_{i \in I}$ une partition de Ω, telle que $P(A_i) > 0$, pour tout $i \in I$. Alors, pour tout événement B,*

$$P(B) = \sum_{i \in I} P(B|A_i)P(A_i)$$

La formule des probabilités totales permet de suivre les étapes de l'expérience aléatoire dans l'ordre chronologique. Nous allons maintenant voir une formule à remonter le temps...

Proposition 10 (Formule de Bayes) *Soit A et B deux événements tels que $0 < P(A) < 1$ et $P(B) > 0$. Alors,*

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

preuve :

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

et on conclut en remplaçant $P(B)$ par son expression donnée par la formule des probabilités totales. □

Proposition 11 (Formule de Bayes généralisée) *Soit $(A_i)_{i \in I}$ une partition de Ω, telle que $P(A_i) > 0$, pour tout $i \in I$. Soit un événement B, tel que $P(B) > 0$. Alors, pour tout $i \in I$,*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j \in I} P(B|A_j)P(A_j)}$$

Exemple 12 *Deux opérateurs de saisie, A et B, entrent respectivement 100 et 200 tableaux sur informatique. Les tableaux de A comportent des fautes dans 5,2% des cas et ceux de B dans 6,7% des cas. On prend un tableau au hasard. Il comporte des fautes. Quelle est la probabilité pour que A se soit occupé de ce tableau ?*
Soient les événements :
$T_A =$*" le tableau est entré par A",*
$T_B = (T_A)^c$*" le tableau est entré par B",*
$F =$*" le tableau comporte des fautes".*
D'après le théorème de Bayes,

$$\begin{aligned} P(T_A|F) &= \frac{P(F|T_A)P(T_A)}{P(F|T_A)P(T_A) + P(F|T_B)P(T_B)} \\ &= \frac{0.052 * 1/3}{0.052 * 1/3 + 0.067 * 2/3} = 0.279 \end{aligned}$$

Définition 13 *Deux événements A et B sont dits* **indépendants** *si*

$$P(A \cap B) = P(A)P(B)$$

S'il sont de probabilité non nulle, alors

$$P(B|A) = P(B) \Longleftrightarrow P(A|B) = P(A) \Longleftrightarrow P(A \cap B) = P(A)P(B)$$

remarque 1 : A et B sont donc indépendants si la connaissance de la réalisation de l'un n'influence pas la probabilité de l'autre.
remarque 2 : deux événements incompatibles A et B, avec $P(A) > 0$ et $P(B) > 0$, ne sont jamais indépendants. En effet, $A \cap B = \emptyset$ entraîne $P(A \cap B) = 0 \neq P(A)P(B)$.

1.5 Répétitions indépendantes

Quand on étudie une expérience aléatoire qui peut se décomposer en plusieurs petites expériences aléatoires indépendantes, les calculs sont aisés. Et quand on a la probabilité uniforme pour chacune de ces petites expériences aléatoires, on a encore la probabilité uniforme sur l'expérience aléatoire totale.

Proposition 14 *Soit $\Omega = E \times F$ où E est de cardinal n et F de cardinal p. Supposons que l'on choisisse avec la probabilité uniforme un élément de E, et, de manière indépendante, un élément de F toujours avec la probabilité uniforme. Alors chaque élément $\omega = (x, y)$ de Ω a la même probabilité, qui vaut*

$$P(\omega) = P((x, y)) = \frac{1}{\text{card}(\Omega)} = \frac{1}{np} = P_E(\{x\})P_F(\{y\})$$

Exemple 15 *On lance une pièce de monnaie équilibrée et un dé équilibré.*

$$\Omega = \{P, F\} \times \{1, ..., 6\}$$

Comme on a la probabilité uniforme sur $\{P, F\}$ *et sur* $\{1, ..., 6\}$*, on a finalement la probabilité uniforme sur* Ω *et*

$$\forall \omega \in \Omega, \qquad P(\omega) = \frac{1}{\text{card}(\Omega)} = 1/12$$

Proposition 16 *On répète* N *fois, de manière indépendante, la même expérience aléatoire modélisée par un univers* Ω *et par une probabilité* P*. Alors le nouvel univers est* $\Omega^N = \Omega \times \cdots \Omega$*, et la probabilité associée est*

$$P^N\Big((\omega_1, ..., \omega_N)\Big) = P(\omega_1) \cdots P(\omega_N)$$

En particulier, si P *est la probabilité uniforme sur* Ω*, alors* P^N *est la probabilité uniforme sur* Ω^N*.*

Le chevalier de Méré :
le Chevalier de Méré avait constaté qu'il obtenait plus souvent 11 que 12 avec trois dés. Pourtant, le nombre de combinaisons dont la somme fait 12 est le même que le nombre de combinaisons dont la somme fait 11. Alors ?

1.6 Exercices

Exercice 1 – Proposer un univers Ω pour les expériences aléatoires suivantes et dénombrer les résultats possibles :
1) On lance un dé.
2) On lance 2 dés.
3) On tire trois cartes dans un jeu .
4) On place les 5 lettres qui forment "proba" au hasard sur une réglette de Scrabble.
5) On place les 6 lettres qui forment "erreur" au hasard sur une réglette de Scrabble.

Exercice 2 – Soit P une probabilité sur un ensemble Ω et deux événements A et B. On suppose que

$$P(A \cup B) = 7/8, \qquad P(A \cap B) = 1/4, \qquad P(A) = 3/8.$$

Calculer $P(B)$, $P(A \cap B^c)$, $P(B \cap A^c)$.

Exercice 3 – Supposons que les faces d'un dé sont truquées de telle manière que les numéros impairs ont chacun la même chance d'apparaître, chance qui est deux fois plus grande que pour chacun des numéros pairs. On jette le dé. Quelle est la probabilité d'obtenir un nombre supérieur ou égal à 4 ?

Exercice 4 – Un parking contient douze places alignées. Huit voitures s'y sont garées au hasard, et l'on observe que les quatre places libres se suivent. Est-ce surprenant ?

Exercice 5 — La probabilité qu'un objet fabriqué à la chaîne ait un défaut est de $0,01$. Trouver la probabilité que, dans un lot de 100 objets, il y ait au moins un objet défectueux. Quelle est la probabilité qu'il y ait, dans un tel lot, exactement un objet défectueux ?

Exercice 6 – Soient M_1, M_2, M_3 trois personnes. La première M_1 dispose d'une information codée sous forme + ou −. Elle la transmet à la deuxième personne M_2. Puis M_2 la transmet à M_3. Malheureusement, à chaque fois que l'information est transmise, il y a une probabilité p que l'information soit changée en son contraire. En tenant compte du fait que deux changements rétablissent la vérité, quelle est la probabilité pour que M_3 ait le bon message ?
Et si M_3 transmet l'information dont il dispose à une quatrième personne M_4, quelle est la probabilité pour que M_4 ait la bonne information ?
Lorsque $p = 0.2$, quelle est la valeur numérique de cette probabilité ?

Exercice 7 — Un nouveau vaccin a été testé sur 12500 personnes ; 75 d'entre elles, dont 35 femmes enceintes, ont eu des réactions secondaires nécessitant une hospitalisation. Parmi les 12500 personnes testées, 680 personnes sont des femmes enceintes.
1. Quelle est la probabilité pour une femme enceinte, d'avoir une réaction secondaire si elle reçoit un vaccin ?
2. Quelle est la probabilité pour une personne non enceinte, d'avoir une réaction secondaire ?

Exercice 8 — Dans une usine, la machine A fabrique 60% des pièces, dont 2% sont défectueuses.
La machine B fabrique 30% des pièces, dont 3% sont défectueuses.
La machine C fabrique 10% des pièces, dont 4% sont défectueuses.
1. On tire une pièce au hasard dans la fabrication. Quelle est la probabilité qu'elle soit défectueuse ?
2. On tire une pièce au hasard dans la fabrication. Elle est défectueuse. Quelle est la probabilité qu'elle ait été fabriquée par la machine A ? par la machine B ? par la machine C ?

Exercice 9 — Dans une jardinerie : 25% des plantes ont moins d'un an, 60% ont de 1 à 2 ans, 25% ont des fleurs jaunes, 60% ont des fleurs roses, 15% ont des fleurs jaunes et moins d'un an, 3% ont plus de 2 ans et n'ont ni fleurs jaunes, ni fleurs roses. 15% de celles qui ont de 1 à 2 ans, ont des fleurs jaunes, 15% de celles qui ont de 1 à 2 ans, n'ont ni fleurs jaunes ni fleurs roses. On suppose que les fleurs ne peuvent pas être à la fois jaunes et roses.
On choisit une plante au hasard dans cette jardinerie.
1. Quelle est la probabilité qu'elle ait moins d'un an et des fleurs roses ?
2. Quelle est la probabilité qu'elle ait des fleurs roses, sachant qu'elle a plus de 2 ans ?
3. Quelle est la probabilité qu'elle ait plus de deux ans et des fleurs jaunes ?

Exercice 10 — Deux chauffeurs de bus se relaient sur la même ligne. Lors d'une grève, le premier a 60% de chances de faire grève et le second 80%. Pendant la prochaine grève, quelle est la probabilité pour qu'un seul des deux chauffeurs fasse grève ?

Exercice 11 — Une loterie comporte 500 billets dont deux seulement sont gagnants. Combien doit-on acheter de billets pour que la probabilité d'avoir au moins un billet gagnant soit supérieure ou égale à 0.5 ?

Exercice 12 — 1. Dans une classe de 36 élèves, quelle est la probabilité pour que deux élèves au moins soient nés le même jour ? (on considèrera que l'année compte 365 jours, et que toutes les dates d'anniversaires sont indépendantes et équiprobables).
2. Généraliser ce résultat pour une classe de n élèves. Tracer le résultat obtenu.

Chapitre 2

Variables aléatoires discrètes

Le travail sur les événements devient vite fastidieux, ainsi nous allons maintenant nous restreindre à étudier des grandeurs numériques obtenues pendant l'expérience aléatoire.

2.1 Définitions

Définition 17 *Une variable aléatoire (v.a.) X est une fonction définie sur l'espace fondamental Ω, qui associe une valeur numérique à chaque résultat de l'expérience aléatoire étudiée. Ainsi, à chaque événement élémentaire ω, on associe un nombre $X(\omega)$.*

Exemple 18 *On lance trois fois une pièce et on s'intéresse au nombre X de fois ou PILE apparaît. Il y a deux manières de formaliser cette phrase. Tout d'abord, à chaque événement élémentaire ω, on associe $X(\omega)$. Ainsi,*

ω	*PPP*	*PPF*	*PFP*	*FPP*	*FFP*	*FPF*	*PFF*	*FFF*
valeur de X	*3*	*2*	*2*	*2*	*1*	*1*	*1*	*0*

Ensuite, comme on observe que plusieurs événements élémentaires donnent la même valeur, on peut les regrouper et obtenir des événements (événement = réunion d'événements élémentaires) qui correspondent à des valeurs distinctes de X :

k (valeur prise par X)	3	2	1	0
événement $[X = k]$	$\{PPP\}$	$\{PPF,PFP,FPP\}$	$\{PFF,FPF,FFP\}$	$\{FFF\}$

On peut d'emblée observer que les événements $(X = 0)$, $(X = 1)$, $(X = 2)$ et $(X = 3)$ sont deux à deux disjoints. De plus, la réunion de ces événements est Ω.

Il est aisé de voir que pour toute v.a., **les événements correspondant à des valeurs distinctes de X sont incompatibles**. Autrement dit, dès que $i \neq j$, les événements $(X = i)$ et $(X = j)$ sont incompatibles : $(X = i) \cap (X = j) = \emptyset$. De plus, **la réunion de ces événements forme l'espace Ω tout entier** :

$$\bigcup_k (X = k) = \Omega$$

Une variable qui ne prend qu'un nombre dénombrable de valeurs est dite **discrète**, sinon, elle est dite **continue** (exemples : hauteur d'un arbre, distance de freinage d'une voiture roulant à 100 km/h). Les v.a. continues seront étudiées plus tard.

Définition 19 *La loi d'une variable aléatoire discrète X est la liste de toutes les valeurs différentes que peut prendre X avec les probabilités qui leur sont associées. On utilisera souvent une formule, plutôt qu'une liste.*

Exemple 18 : nous avons déjà la liste de tous les événements élémentaires et ils sont équiprobables, de probabilité 1/8. D'après la composition des événements $[X = k]$, pour $k = 0, ..., 3$, on peut déduire facilement la loi de X.

valeur de X (événement)	$[X = 3]$	$[X = 2]$	$[X = 1]$	$[X = 0]$
composition de l'événement	{PPP}	{PPF,PFP,FPP}	{PFF,FPF,FFP}	{FFF}
probabilité	1/8	3/8	3/8	1/8

Un autre outil permet de caractériser la loi d'une v.a. : il s'agit de la fonction de répartition empirique.

Définition 20 *Soit X une v.a.. On appelle fonction de répartition de X la fonction de $\mathbb{R}$ dans $[0, 1]$, définie pour tout $x \in \mathbb{R}$ par*

$$F(x) = P[X \leq x]$$

Exemple : X est le nombre de Face quand on lance trois fois une pièce. On a vu que la loi de X est

$$P[X = 0] = 1/8, \qquad P[X = 1] = P[X = 2] = 3/8, \qquad P[X = 3] = 1/8$$

D'où,

$$F(x) = \begin{cases} 0 & \text{si } x < 0, \\ 1/8 & \text{si } 0 \leq x < 1, \\ 4/8 & \text{si } 1 \leq x < 2, \\ 7/8 & \text{si } 2 \leq x < 3, \\ 1 & \text{si } x \geq 3 \end{cases}$$

Remarque : deux v.a. ayant même loi ont même fonction de répartition.

Proposition 21 *Soit F une fonction de répartition. Alors*
1) F est croissante,
2) F est continue à droite et admet une limite à gauche en tout point x égale à $P[X < x]$,
3)

$$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to +\infty} F(x) = 1$$

Pour une v.a. discrète, la fonction de répartition est une fonction en escalier, avec un saut en chaque valeur k de $X(\Omega)$ et la hauteur de ces sauts est la probabilité $P(X = k)$.

Une fois la loi d'une v.a. établie, on peut calculer, comme pour une série statistique, un indicateur de position (l'espérance) et un indicateur de dispersion (la variance).

Définition 22 *L'espérance ou moyenne d'une v.a. discrète X est le réel*

$$E[X] = \sum_k kP[X = k]$$

où on somme sur toutes les valeurs k que peut prendre X.

Un résultat remarquable permet de calculer facilement l'espérance d'une fonction de X connaissant la loi de X : c'est le **théorème du transfert**.

Théorème 23 *Pour toute fonction g,*

$$E[g(X)] = \sum_k g(k)P[X = k]$$

preuve : observons que $g(X) = y$ ssi $X = x$ avec $g(x) = y$. Ainsi,

$$P(g(X) = y) = \sum_{x:g(x)=y} P(X = x)$$

Multiplions cette égalité par y puis sommons :

$$E(Y) = \sum_y yP(Y = y) = \sum_y \sum_{x:g(x)=y} g(x)P(X = x) = \sum_x g(x)P(X = x)$$

□

Définition 24 *La variance d'une v.a. discrète X est le réel positif*

$$\mathrm{Var}(X) = E\Big[(X - E[X])^2\Big] = \sum_k (k - E[X])^2 P[X = k] = E[X^2] - E[X]^2$$

et l'écart-type de X est la racine carrée de sa variance.

Remarque : l'espérance n'a pas toujours un sens quand $X(\Omega)$ est infini. Dans ce cas, X a une espérance si

$$\sum_{k \in X(\Omega)} |k| P(X = k) < \infty$$

L'espérance d'une v.a. est la moyenne des valeurs que peut prendre X, pondérée par les probabilités de ces valeurs. On appelle souvent l'espérance tout simplement moyenne de X : elle correspond à une valeur moyenne autour de laquelle sont réparties les valeurs que peut prendre X. L'écart-type (ou la variance) mesure la dispersion de la v.a. X autour de sa valeur moyenne $E[X]$.
L'espérance et sa variance ne dépendent de X qu'à travers sa loi : deux variables qui ont même loi ont même espérance, même variance.
Exemple 18 : nous avons la loi du nombre X de PILE quand on lance trois fois une pièce.

$$\begin{aligned}
E[X] &= \sum_{k=0}^{3} kP[X = k] = 3 \cdot \frac{1}{8} + 2 \cdot \frac{3}{8} + 1 \cdot \frac{3}{8} + 0 \cdot \frac{1}{8} = \frac{12}{8} = \frac{3}{2} \\
\mathrm{Var}(X) &= E[X^2] - E[X]^2 = \sum_{k=0}^{3} k^2 P[X = k] - E[X]^2 \\
&= 3^2 \cdot \frac{1}{8} + 2^2 \cdot \frac{3}{8} + 1^2 \cdot \frac{3}{8} + 0^2 \cdot \frac{1}{8} - \left(\frac{3}{2}\right)^2 \\
&= \frac{3}{4}
\end{aligned}$$

Fréquences empiriques VERSUS loi de probabilité

Il ne faut pas confondre les fréquences observées sur un échantillon et la loi de probabilité. Reprenons l'exemple 18. J'ai fait l'expérience de lancer 10 fois trois pièces de monnaie en relevant à chaque fois le nombre de PILE obtenus. Voici les fréquences observées :

nbr de PILE	$[X=3]$	$[X=2]$	$[X=1]$	$[X=0]$
probabilité	0.125	0.375	0.375	0.125
fréquence observée	0.2	0.6	0.1	0.1

Quand le nombre d'essais augmente, les fréquences observées sont de plus en plus proches des valeurs théoriques données par la loi de probabilité (preuve plus tard). C'est pourquoi, quand on ne peut pas déterminer la loi d'une v.a. aussi facilement que dans cet exemple, on considère que les fréquences empiriques (c'est-à-dire mesurées sur un échantillon) sont des valeurs approchées des probabilités. Il n'en reste pas moins qu'il ne faut pas assimiler ces deux objets, et bien comprendre la différence entre la moyenne théorique, calculée à partir de la loi de probabilité, et la moyenne empirique, calculée à partir de quelques observations.

2.2 Indépendance et conditionnement

Définition 25 *Deux v.a. X et Y sont dites indépendantes si, pour tous i et j, les événements $\{X=i\}$ et $\{Y=j\}$ sont indépendants, ie*

$$P[X=i, Y=j] = P[X=i]P[Y=j]$$

Attention ! Si X et Y ne sont pas indépendantes, connaître la loi de X et celle de Y ne suffit pas pour connaître la loi de (X,Y), qui est la donnée, pour tous i et j, de $P[(X,Y)=(i,j)] = P[X=i, Y=j]$.

Proposition 26 (La formule des probabilités totales) *Soient X et Y deux v.a.. Pour tout $i \in X(\Omega)$,*

$$P[X=i] = \sum_{j \in Y(\Omega)} P[X=i|Y=j]P[Y=j]$$

preuve : on peut appliquer la formule des probabilités totales 9 car les événements $Y=j$ forment une partition de Ω. □

Exemple 27 *On lance deux dés équilibrés et on note X et Y les deux chiffres obtenus. Soit $Z = X+Y$. Quelle est la loi de Z ?*

Méthode 1 : bien sûr, dans ce cas fini, on peut faire un tableau à double entrée avec la valeur que prend X, la valeur que prend Y et la valeur de Z.

$X\backslash Y$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Les deux dés ne se téléphonent pas avant de décider sur quelle face ils vont tomber : leurs résultats sont indépendants. Donc

$$\text{pour tous } 1 \leq i, j \leq 6, \qquad P[X = i, Y = j] = P[X = i]P[Y = j] = 1/36$$

Autrement dit, chaque case du tableau a la même probabilité $1/36$. On en déduit : $P[Z = 2] = P[Z = 12] = 1/36$, $P[Z = 3] = P[Z = 11] = 2/36$, $P[Z = 4] = P[Z = 10] = 3/36$, $P[Z = 5] = P[Z = 9] = 4/36$, $P[Z = 6] = P[Z = 8] = 5/36$, $P[Z = 7] = 6/36$. On vérifie immédiatement que la somme de ces probabilité fait 1.
Méthode 2 : on utilise la formule des probabilités totales. Précisément, on conditionne par les événements $(X = i)$ $(1 \leq i \leq 6)$, qui forment une partition de Ω, car quand on connaît la valeur que prend X, on peut facilement calculer la probabilité pour avoir $Z = j$. Soit $1 \leq j \leq 12$.

$$\begin{aligned} P[Z = j] &= \sum_{i=1}^{6} P[Z = j|X = i]P[X = i] \\ &= \frac{1}{6}\sum_{i=1}^{6} P[X + Y = j|X = i] \\ &= \frac{1}{6}\sum_{i=1}^{6} P[Y = j - i|X = i] \\ &= \frac{1}{6}\sum_{i=1}^{6} P[Y = j - i] \end{aligned}$$

La dernière égalité vient du fait que X et Y sont indépendants. Il suffit maintenant de se rappeler que $P[Y = k] = 1/6$ seulement si k est dans $\{1, ..., 6\}$.

Théorème 28 *1) Pour toutes v.a. X et Y, $E[X + Y] = E[X] + E[Y]$.*
2) Si X et Y sont indépendantes, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

preuve : pour le premier point, il faut observer que

$$\sum_y P(X = x, Y = y) = P\Big((X = x) \cap (\cup_y (Y = y))\Big) = P\Big((X = x) \cap \Omega\Big) = P(X = x)$$

et il vient

$$\begin{aligned} E[X+Y] &= \sum_{x,y}(x+y)P(X=x,Y=y) \\ &= \sum_{x,y} xP(X=x,Y=y) + \sum_{x,y} yP(X=x,Y=y) \\ &= \sum_{x} xP(X=x) + \sum_{y} yP(Y=y) = E[X]+E[Y] \end{aligned}$$

Pour le second point, on montre tout d'abord que $E(XY) = E(X)E(Y)$, la suite venant facilement. Ainsi,

$$\begin{aligned} E[XY] &= \sum_{x,y} xyP(X=x,Y=y) \\ &= \sum_{x,y} xyP(X=x)P(Y=y) \\ &= \Big(\sum_{x} xP(X=x)\Big)\Big(\sum_{y} yP(Y=y)\Big) \\ &= E(X)E(Y) \end{aligned}$$

2.3 Schéma de Bernoulli et loi binomiale

On s'intéresse ici à la réalisation ou non d'un événement. Autrement dit, on n'étudie que les expériences aléatoires qui n'ont que **deux issues possibles** (ex : un patient à l'hôpital survit ou non, un client signe le contrat ou non, un électeur vote démocrate ou républicain...). Considérons une expérience aléatoire de ce type. On l'appelle une **épreuve de Bernoulli**. Elle se conclut par un succès si l'événement auquel on s'intéresse est réalisé ou un échec sinon. On associe à cette épreuve une variable aléatoire Y qui prend la valeur 1 si l'événement est réalisé et la valeur 0 sinon. Cette v.a. ne prend donc que deux valeurs (0 et 1) et sa loi est donnée par :

$$P[Y=1] = p, \qquad P[Y=0] = q = 1-p$$

On dit alors que Y suit une **loi de Bernoulli de paramètre** p, notée $\mathcal{B}(p)$. La v.a. Y a pour espérance p et pour variance $p(1-p)$. En effet, $E[Y] = 0 \times (1-p) + 1 \times p = p$ et $\mathrm{Var}(Y) = E[Y^2] - E[Y]^2 = E[Y] - E[Y]^2 = p(1-p)$.

Un schéma de Bernoulli est la répétition n fois de la même épreuve dans les mêmes conditions.

Schéma de Bernoulli :
1) Chaque épreuve a deux issues : succès [S] ou échec [E].
2) Pour chaque épreuve, la probabilité d'un succès est la même, notons $P(S) = p$ et $P(E) = q = 1-p$.
3) Les n épreuves sont indépendantes : la probabilité d'un succès ne varie pas, elle ne dépend pas des informations sur les résultats des autres épreuves.

Soit X la v.a. qui représente le nombre de succès obtenus lors des n épreuves d'un schéma de Bernoulli. Alors on dit que X suit une **loi binomiale de paramètres** (n, p), notée $\mathcal{B}(n, p)$. Cette loi est donnée par :

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{pour tout } 0 \leq k \leq n$$

où $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.
En effet, on peut tout d'abord observer que le nombre de succès est un entier nécessairement compris entre 0 et le nombre d'épreuves n. D'autre part, l'événement $\{X = k\}$ est une réunion d'événements élémentaires et chacun de ces événements élémentaires est une suite de longueur n de la forme ω =[SSES...E] avec k S et $n - k$ E. Un tel événement élémentaire a la probabilité

$$P(\omega) = p^k (1-p)^{n-k}$$

Combien existe-t-il de suites à n éléments avec k S et $n - k$ E ?
Il en existe $\binom{n}{k}$, le nombre de combinaisons de k S parmi n éléments. Finalement,

$$\begin{aligned} P(X = k) &= \sum_{\omega : X(\omega) = k} P(\omega) \\ &= \text{card}(\{\omega : X(\omega) = k\})\ p^k (1-p)^{n-k} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

Proposition 29 *Si X est une v.a. de loi $\mathcal{B}(n, p)$, l'espérance de X vaut np et sa variance $np(1-p)$.*

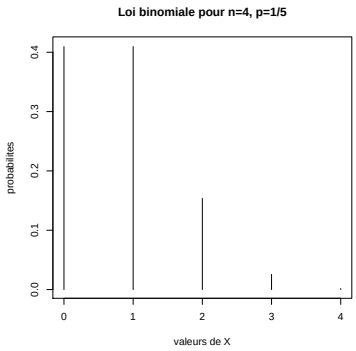
(preuve)

Exemple 30 (échantillon avec remise) *On considère une population avec deux catégories d'individus, A et B. On choisit de manière équiprobable un individu dans la population. On note le résultat de l'épreuve, c'est-à-dire si l'individu appartient à la catégorie A ou B. Puis on le remet dans le lot et on recommence : on choisit à nouveau un individu dans la population... Cela constitue notre schéma de Bernoulli.*
Supposons que les individus de la catégorie A sont en nombre N_A dans la population qui contient N individus. Alors pour chaque épreuve de Bernoulli, la probabilité d'avoir un individu de la catégorie A (ce que nous appellerons un succès) est $p = N_A/N$. Le nombre d'individus, présents dans l'échantillon, qui appartiennent à la catégorie A est une variable aléatoire, car sa valeur dépend du choix de l'échantillon, c'est-à-dire de l'expérience aléatoire.

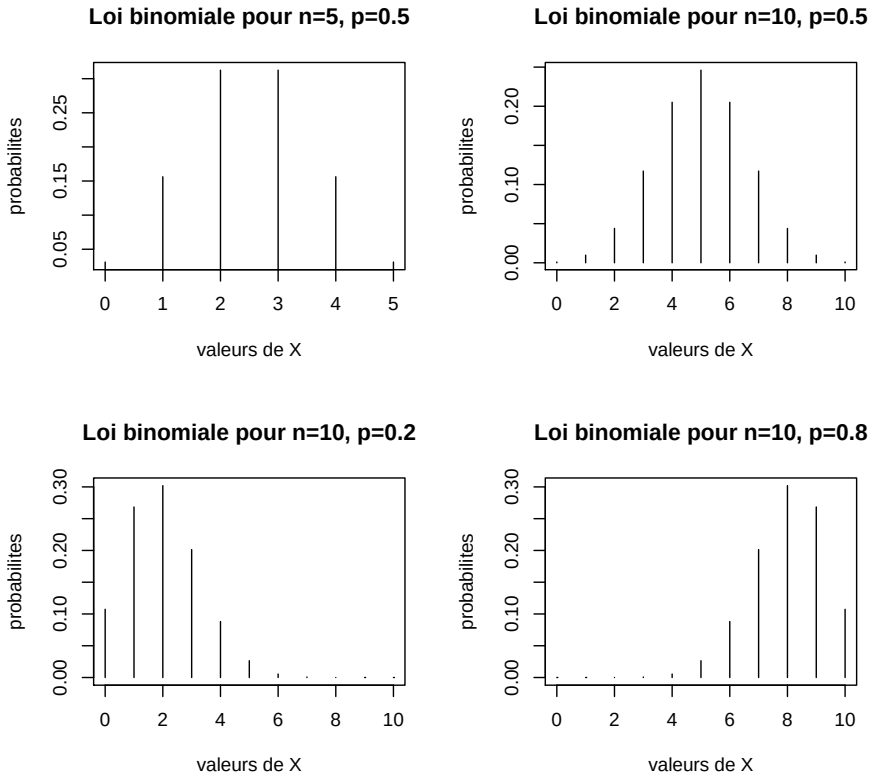
Quel est le nombre X d'individus de la catégorie A dans l'échantillon ? D'après ce qu'on vient de voir, on est en présence d'un schéma de Bernoulli où N_A/N est la probabilité de succès et n est le nombre d'épreuves : une épreuve est le tirage d'un individu et un succès correspond à l'événement "l'individu appartient à la catégorie A". Donc X suit une loi binomiale $\mathcal{B}(n, N_A/N)$.

Exemple 31 *On s'intéresse à l'infection des arbres d'une forêt par un parasite. Soit p la proportion d'arbres infectés. On étudie 4 arbres. Si un arbre est infecté, on dit qu'on a un succès, sinon un échec. Soit X le nombre d'arbres infectés, parmi les 4. Alors X suit la loi binomiale $\mathcal{B}(4,p)$.*

$P(X=0) = \binom{4}{0}\ q^4 = q^4$,
$P(X=1) = \binom{4}{1}\ p^1q^3 = 4pq^3$,
$P(X=2) = \binom{4}{2}\ p^2q^2 = 6p^2q^2$,
$P(X=3) = \binom{4}{3}\ p^3q^1 = 4p^3q$,
$P(X=4) = \binom{4}{4}\ p^4 = p^4$.
Pour $p = 1/5$, on obtient les valeurs :



Voici d'autres exemples.



Remarque : associons à chaque épreuve de Bernoulli une v.a. Y_i ($1 \leq i \leq n$) qui vaut 1 si on observe un succès au i-ème essai et 0 sinon. Alors le nombre de succès, noté X, vérifie

$$X = \sum_{i=1}^{n} Y_i$$

Autrement dit, une v.a de loi $\mathcal{B}(n,p)$ est une somme de n v.a indépendantes de loi $\mathcal{B}(p)$.

Exemple 32 *Un nouveau traitement résout le problème de fragilité d'un matériel dans 50% des cas. Si on essaye sur 15 objets, quelle est la probabilité pour qu'au plus 6 de ces objets soient résistants, pour que le nombre d'objets résistants soit compris entre 6 et 10, pour que deux au plus restent fragiles. Quel est le nombre moyen d'objets rendus résistants par le traitement ?*

On suppose que les résultats concernant les 15 objets sont indépendants. Soit X le nombre d'objets résistants à la suite du nouveau traitement. Alors X suit une loi $\mathcal{B}(15, 0.5)$ et

$$\begin{aligned} P[X \leq 6] &= P[X=0] + P[X=1] + \cdots + P[X=6] \\ &= \frac{1}{2^{15}}\left(\binom{15}{0} + \binom{15}{1} + \binom{15}{2} + \binom{15}{3} + \binom{15}{4} + \binom{15}{5} + \binom{15}{6}\right) \\ &= \frac{1}{2^{15}}(1 + 15 + 105 + 455 + 1365 + 3003 + 5005) \\ &= 0.304 \end{aligned}$$

$$P[6 \leq X \leq 10] = P[X=6] + P[X=7] + P[X=8] + P[X=9] + P[X=10] = 0.790$$

$$\begin{aligned} P[X \geq 12] &= P[X=12] + P[X=13] + P[X=14] + P[X=15] \\ &= (455 + 105 + 15 + 1)/2^{15} \\ &= 0.018 \end{aligned}$$

Enfin, $E[X] = 15/2 = 7,5$.

2.4 Trois autres lois discrètes

2.4.1 Loi géométrique

Au lieu de réaliser un nombre fixé d'essais lors d'un schéma de Bernoulli, l'expérimentateur s'arrête au premier succès. La valeur qui nous intéresse est le nombre d'essais effectués jusqu'au premier succès inclus. Le nombre de succès est donc fixé à 1, mais le nombre d'essais total Y est aléatoire et peut prendre n'importe quelle valeur entière supérieure ou égale à 1. De plus,

$$\forall k = 1, 2, ... \qquad P[Y=k] = p(1-p)^{k-1}$$

où p est toujours la probabilité de succès des épreuves de Bernoulli. On dit alors que Y suit la **loi géométrique de paramètre** p, notée $\mathcal{G}(p)$.

Proposition 33 *L'espérance de Y vaut $1/p$ et sa variance $(1-p)/p^2$.*

preuve : admettons tout d'abord que, sur $[0, 1[$,

$$\left(\sum_{k=0}^{\infty} x^k\right)' = \sum_{k=0}^{\infty} (x^k)' = \sum_{k=1}^{\infty} k x^{k-1}$$

et

$$\left(\sum_{k=0}^{\infty} x^k\right)' = \left(\frac{1}{1-x}\right)' = \frac{1}{(1-x)^2}$$

D'où, pour $x = 1-p$,

$$E[Y] = \sum_{k=1}^{\infty} kP[X=k] = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = p/p^2 = 1/p$$

Un calcul analogue permet de calculer la variance (exercice).

2.4.2 Loi de Poisson

Cette loi est une approximation de la loi binomiale quand np est petit et n grand (en pratique, $n \geq 50$ et $np \leq 10$). Une v.a. X de loi de Poisson de paramètre λ, notée $\mathcal{P}(\lambda)$, vérifie

$$\forall k \in \mathbb{N}, \qquad P[X=k] = \exp(-\lambda)\frac{\lambda^k}{k!}$$

L'espérance et la variance de X sont égales à λ.

On utilise la loi de Poisson pour modéliser le nombre de tâches qui arrivent à un serveur informatique pendant une minute, le nombre de globules rouges dans un ml de sang, le nombre d'accidents du travail dans une entreprise pendant un an...
Dans le cas de l'approximation de la loi binomiale par la loi de Poisson, le paramètre de la loi de Poisson est $\lambda = np$.

2.4.3 Loi uniforme

Mis à part le prestige dû à son nom, la loi uniforme est la loi de l'absence d'information. Supposons qu'une v.a. X prenne les valeurs 1,2,...,n, mais que nous n'ayons aucune idée de la loi de probabilité de X ; dans ce cas, après justification, on peut affecter à chaque valeur le même poids : $1/n$. Et

$$\forall k = 1, ..., n, \quad P[X=k] = \frac{1}{n}$$

On montre facilement que

$$E[X] = \frac{n+1}{2} \qquad \text{et} \qquad \text{Var}(X) = \frac{(n+1)(n-1)}{12}$$

Exercice 2 — Soit X une v.a. dont la loi est donnée par

$$P[X=-1] = 0.2, \qquad P[X=0] = 0.1, \qquad P[X=4] = 0.3, \qquad P[X=5] = 0.4$$

Calculer $P[X \leq 3]$, $P[X > 2]$, l'espérance et la variance de X.

2.5 Exercices

Exercice 1 — Soit X une v.a. dont la loi est donnée par

$$P[X=-1]=0.2, \qquad P[X=0]=0.1, \qquad P[X=4]=0.3, \qquad P[X=5]=0.4$$

Calculer $P[X \leq 3]$, $P[X > 2]$, l'espérance et la variance de X.

Exercice 2 — On lance deux dés. Modéliser l'expérience. Quelle est la probabilité pour obtenir au moins un 4 ? Quelle est la probabilité pour que le plus petit donne un nombre inférieur ou égal à 4 ? Quelle est l'espérance de la somme des deux dés ?

Exercice 3 — On admet que le nombre d'accidents survenant sur une autoroute quotidiennement est une va qui suit la loi de Poisson de paramètre $\lambda = 3$. Calculer $P[X = k]$ pour $k = 0, ..., 6$. Faire une représentation graphique. Quelle est la probabilité qu'il y ait au moins 2 accidents lors d'un jour donné ?

Exercice 4 — Pour ces expériences aléatoires, donner les espaces fondamentaux Ω et les probabilités qui les décrivent. Puis, pour chacune des variables aléatoires que l'on étudie, préciser le nom de la loi, ses paramètres et donner l'expression de $P(X = k)$.
a) On lance un dé 20 fois. Quelle est la loi du nombre de 5 obtenus ? Quelle est la probabilité d'obtenir moins de 3 fois un 5 ?
b) Une urne contient une boule blanche et une boule noire. On prend dans cette urne une boule au hasard, on la remet et on ajoute une boule de la même couleur. Quelle est la loi du nombre de boules blanches dans l'urne ?
c) On recommence ce petit jeu. Quelle est la nouvelle loi du nombre de boules blanches dans l'urne ? Donner aussi la loi du nombre de boules noires et son espérance.
d) Au bord de l'A7, un étudiant fait du stop. En cette saison, un vingtième des automobilistes s'arrête pour prendre un stoppeur. Quelle est la loi du nombre de véhicules que l'étudiant verra passer avant qu'il ne trouve un chauffeur ? Quelle est la probabilité qu'il monte dans la quatrième voiture qui passe ? Quelle est la probabilité qu'il voit passer au moins 6 voitures qui ne s'arrêtent pas ?

Exercice 5 — Soient X et Y deux variables aléatoires indépendantes de lois respectives $\mathcal{B}(n, p)$ et $\mathcal{B}(n', p)$. Quelle est la loi de la somme $X + Y$?

Exercice 6 — Un ingénieur d'une entreprise de télécommunication recherche en milieu rural des sites pour implanter des relais pour téléphones portables. Ces sites sont des clochers d'église, châteaux d'eau... Sur une carte IGN, l'ingénieur relève 250 sites potentiels dans sa région. De plus, il utilise le quadrillage de sa carte : sa région est divisée en 100 carreaux. On suppose que les sites potentiels sont répartis uniformément et indépendamment sur le territoire. Étant donné un carreau, quelle est la loi du nombre de sites situés

dans ce carreau ? Quelle est la probabilité pour qu'on trouve 5 sites dans ce carreau ? Plus de 5 sites ?

Exercice 7 — Déterminer la loi de la somme de deux v.a. indépendantes de loi de Poisson.

Exercice 8 — Un filtre à particules reçoit un flux de particules, dont le nombre par unité de temps suit une loi de Poisson de paramètre λ. Il filtre ce flux de telle sorte que les particules non toxiques sont rejetées dans l'air. Ces particules non toxiques sont présentes en proportion p dans le gaz originel. Quelle est la loi du nombre de particules rejetées dans l'air par unité de temps ?

Exercice 9 — Un géologue cherche 5 sites potentiels pour exploiter un gisement d'uranium. Après les études de faisabilité, le conseil d'administration choisira le site définitif. Notre géologue charge donc 5 subordonnés de trouver 5 sites vérifiant certaines caractéristiques. Chaque lundi, les employés présentent, lors d'une réunion de travail, un site chacun. On met de côté les sites qui pourraient convenir à première vue. Ces sites seront alors soumis aux études de faisabilité par les employés qui les ont dénichés et on renvoie les autres sur le terrain, jusqu'à ce qu'ils trouvent un site correct.
Soient X la v.a. égale au nombre de lundis de présentation et Y la v.a. égale au nombre de sites présentés.
Calculer $P[X \leq k]$ puis $P[X = k]$ pour $k \in \mathbb{N}$. Puis déterminer la loi de Y.

Chapitre 3

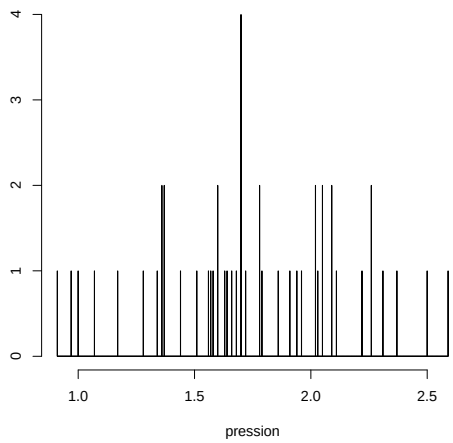
Variables aléatoires continues

On utilise des v.a. discrètes pour compter des événements qui se produisent de manière aléatoire, et des v.a. continues quand on veut mesurer des grandeurs "continues" (distance, masse, pression...).

3.1 Loi d'une v.a. continue

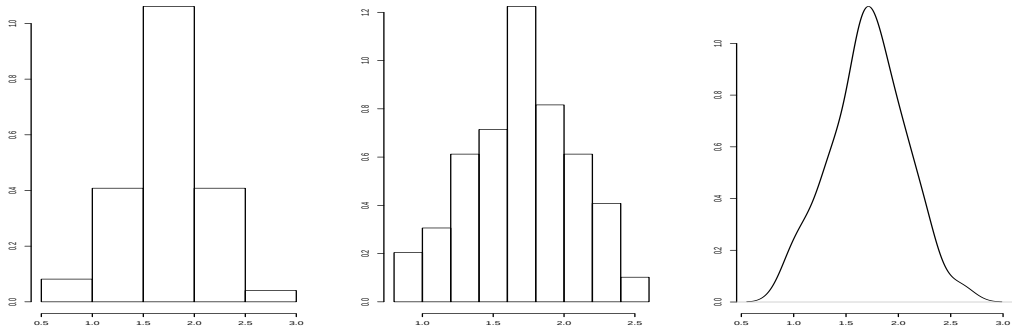
Un industriel fait relever régulièrement la pression dans une cuve de réaction chimique. Voici les mesures obtenues en bar :

2.50	1.37	1.57	1.28	1.64	1.37	2.05	2.11	1.00	1.72
1.96	1.70	2.03	2.26	1.60	1.70	2.09	1.79	1.66	2.09
1.58	1.78	2.02	2.22	1.51	2.59	1.94	1.36	2.31	1.36
0.97	1.07	0.91	2.05	1.70	1.70	1.17	2.02	2.26	1.78
1.60	1.63	1.68	1.34	1.56	1.91	1.86	1.44	2.37	



Cette représentation graphique n'est pas pertinente : on regroupe les valeurs pour tracer un histogramme.

Suivant le nombre de classes choisi, on obtient différents tracés. On peut imaginer de raffiner encore en faisant d'autres mesures avec des instruments de mesure plus précis.



On aurait alors une courbe, d'aire 1 comme les histogrammes, qui représente la manière dont sont réparties les valeurs de la v.a. X (voir les graphes et les données ci-dessus). Cette courbe est la courbe d'une fonction appelée **densité de probabilité** ou simplement **densité**. Une densité f décrit la loi d'une v.a. X en ce sens :

$$\text{pour tous } a, b \in \mathbb{R}, \qquad P[a \leq X \leq b] = \int_a^b f(x)\mathrm{d}x$$

et

$$\text{pour tout } x \in \mathbb{R}, \qquad F(x) = P[X \leq x] = P[X < x] = \int_{-\infty}^{x} f(u)\mathrm{d}u$$

On en déduit qu'une densité doit vérifier

$$\text{pour tout } x \in \mathbb{R},\ f(x) \geq 0 \ \text{ et } \int_{\mathbb{R}} f(x)\mathrm{d}x = 1$$

Définition 34 *On appelle densité de probabilité toute fonction réelle positive, d'intégrale 1.*

Attention ! Pour une v.a. continue X, la densité f ne représente pas la probabilité de l'événement $\{X = x\}$, car $P[X = x] = 0$. Il faut plutôt garder à l'esprit que

$$P[x \leq X \leq x + \Delta x] = f(x).\Delta x$$

Important : La loi d'une v.a. X est donnée par
- sa densité

ou
- les probabilités $P[a \leq X \leq b]$ pour tous a, b

ou
- les probabilités $F(x) = P[X \leq x]$ pour tout x (F est la fonction de répartition).

Remarque : $P[X = x] = 0$ pour tout x.

Proposition 35 *La fonction de répartition F d'une v.a. X de densité f est continue, croissante. Elle est dérivable en tout point x où f est continue et $F'(x) = f(x)$. On a la relation*

$$P[a \leq X \leq b] = P[a < X \leq b] = F(b) - F(a)$$

dès que $b \geq a$.

Définition 36 *L'espérance de la v.a. X est définie par*

$$E[X] = \int_{-\infty}^{+\infty} x f(x) \mathrm{d}x$$

quand cette intégrale a un sens. De plus, la variance de X est $Var(X) = E[X^2] - E[X]^2$.

Proposition 37 *(i) L'espérance d'une fonction $Y = \varphi(X)$ est donnée par*

$$E[Y] = E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x) f(x) \mathrm{d}x$$

(ii) Pour tous réels a et b,

$$E[aX + b] = aE[X] + b$$

$$Var(aX + b) = a^2 Var(X)$$

(iii) Si X et Y sont deux v.a. continues,

$$E[X + Y] = E[X] + E[Y]$$

Si, de plus, elles sont **indépendantes**, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Exemple : Soit X une v.a. de densité f définie par

$$f(x) = \frac{1}{2} \mathbb{1}_{[0,2]}(x) = \begin{cases} \frac{1}{2} & \text{si } x \in [0,2] \\ 0 & \text{sinon} \end{cases}$$

Il s'agit bien d'une densité ($f \geq 0$ et $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$). Calculons $P[1/3 \leq X \leq 2/3]$. Que valent $E[X]$ et $\mathrm{Var}(X)$?

3.2 Loi uniforme

Définition 38 *Une v.a. X suit une loi uniforme sur $[a, b]$, si elle admet pour densité*

$$f(x) = \begin{cases} \frac{1}{b-a} & \textit{si } a < x < b \\ 0 & \textit{sinon} \end{cases}$$

Remarque : on vérifie facilement que f est une densité.
La fonction de répartition F de X est donnée par :

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(x)\mathrm{d}x = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } x > b \end{cases}$$

L'espérance de X est $E[X] = (b-a)/2$ et la variance de X est $\mathrm{Var}(X) = (b-a)^2/12$.
Exercice : soit X de loi uniforme sur $[0, 10]$. Calculer $P[X < 3]$, $P[X > 6]$, $P[3 < X < 8]$.

3.3 La loi normale

3.3.1 Loi normale centrée réduite

C'est la loi la plus importante. Son rôle est central dans de nombreux modèles probabilistes et dans toute la statistique. Elle possède des propriétés intéressantes qui la rendent agréable à utiliser.

Définition 39 *Une v.a. X suit une loi normale (ou loi gaussienne ou loi de Laplace-Gauss) $\mathcal{N}(0,1)$ si sa densité f est donnée par*

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big), \text{ pour tout } x \in \mathbb{R}$$

Remarque : vérifions que f est d'intégrale 1. Notons $I = \int_{\mathbb{R}} f$.

$$\begin{aligned} I^2 &= \Big(\int_{\mathbb{R}} f(x)\mathrm{d}x\Big)\Big(\int_{\mathbb{R}} f(y)\mathrm{d}y\Big) \\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \exp\Big(-\frac{x^2+y^2}{2}\Big)\mathrm{d}x\mathrm{d}y \end{aligned}$$

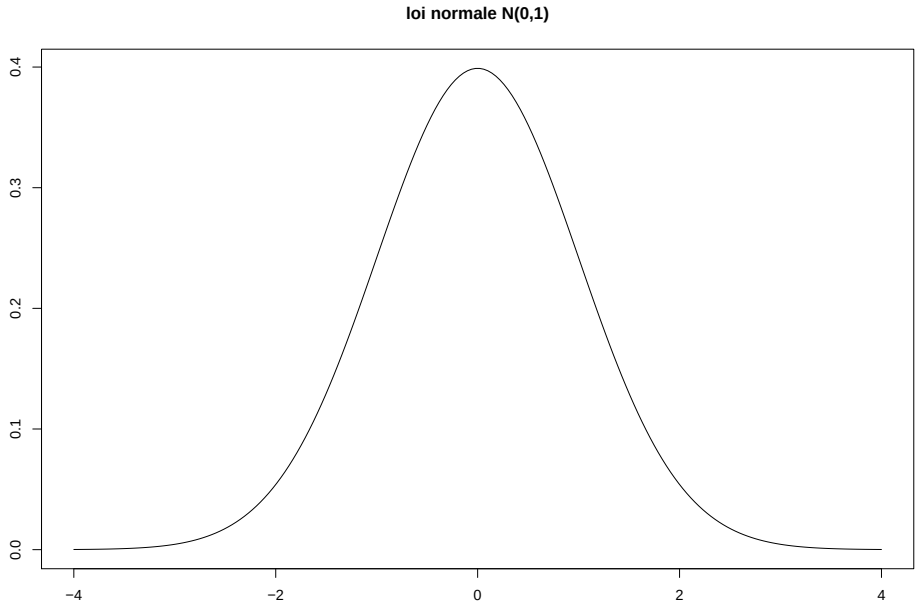
On fait un changement de variable polaire : $x = r\cos\theta$ et $y = r\sin\theta$. Alors $\mathrm{d}x\mathrm{d}y = r\mathrm{d}r\mathrm{d}\theta$ et on obtient :

$$I^2 = \frac{1}{2\pi}\int_0^{+\infty}\int_0^{2\pi} \exp(-r^2/2) r\mathrm{d}\theta\mathrm{d}r = 1$$

Rappel : si X suit une loi normale $\mathcal{N}(0,1)$, alors pour tous $a < b$,

$$P[a \le X \le b] = \frac{1}{\sqrt{2\pi}}\int_a^b \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x = \varphi(b) - \varphi(a)$$

où φ est la fonction de répartition de X. Rappelons que φ est une primitive de la densité f. Mais il n'existe pas de forme analytique de la primitive de f. On doit donc lire les valeurs de φ dans une table disponible en annexe B, ou la faire calculer par un logiciel adapté (ex : excel, matlab, R).
La courbe de la densité de la loi normale $\mathcal{N}(0,1)$ porte le nom de "courbe en cloche". Elle tend vers 0 en l'infini, est croissante sur $\mathbb{R}^-$, puis décroissante. Elle admet donc un maximum en 0. On peut voir aussi qu'elle est symétrique, de centre de symétrie 0.



Proposition 40 *La v.a. X est* **centrée**, *c'est-à-dire de moyenne nulle, et* **réduite**, *c'est-à-dire de variance 1. De plus, $-X$ suit encore une loi normale centrée réduite.*

preuve : le calcul de l'espérance est immédiat quand on a observé que $xf(x)$ est une fonction impaire. Le calcul de la variance se fait par une intégration par parties. Enfin, montrons que X et $-X$ ont même loi. Pour tous $a, b \in \mathbb{R}$,

$$P[a \leq -X \leq b] = P[-b \leq X \leq -a] = \int_{-b}^{-a} f(x)\mathrm{d}x$$

On effectue un le changement de variable : $y = -x$

$$P[a \leq -X \leq b] = \int_{a}^{b} f(-y)\mathrm{d}y = \int_{a}^{b} f(y)\mathrm{d}y$$

par symétrie de la fonction f. D'où

$$P[a \leq -X \leq b] = P[a \leq X \leq b]$$

Utilisation des tables

Pour calculer $P[a \leq X \leq b]$ ou $P[X \leq x]$, on a recours au calcul numérique sur ordinateur ou, plus simplement, à une table qui donne $P[X \leq x]$ pour tout décimal positif x à deux chiffres après la virgule. Puis il faut remarquer que

$$P[a \leq X \leq b] = P[X \leq b] - P[X \leq a]$$

puis on lit les probabilités dans la table si a et b sont positifs. Pour trouver $P[X \leq -x]$ quand $x > 0$, on utilise le fait que X et $-X$ ont même loi :

$$P[X \leq -x] = P[-X \geq x] = P[X \geq x] = 1 - P[X \leq x]$$

Exemple : on cherche à calculer $P[1 \leq X \leq -1]$.

$$\begin{aligned} P[-1 \leq X \leq 1] &= P[X \leq 1] - P[X \leq -1] = P[X \leq 1] - \Big(1 - P[X \leq 1]\Big) \\ &= 2P[X \leq 1] - 1 = 2 \times 0.8413 - 1 = 0,6826 \end{aligned}$$

Exemple : on cherche $u \in \mathbb{R}$ tel que $P[-u \leq X \leq u] = 0.90$.

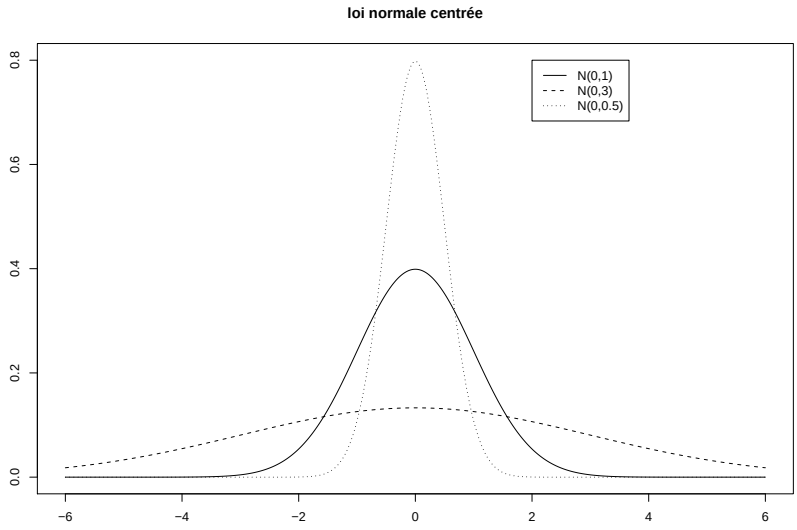
$$P[-u \leq X \leq u] = 2P[X \leq u] - 1$$

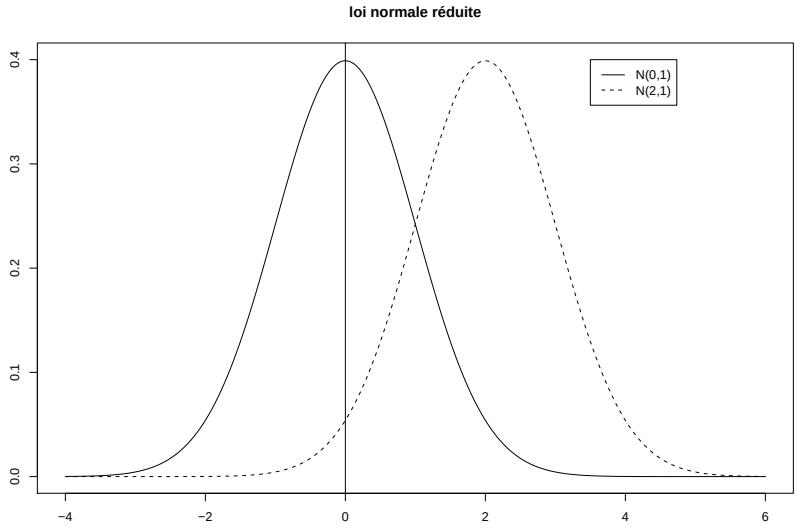
D'où $P[X \leq u] = 0.95$ et $u = 1.6446$.

3.3.2 Loi normale : cas général

Proposition 41 *Soient $m \in \mathbb{R}$ et $\sigma > 0$. Une v.a. X suit une loi normale $\mathcal{N}(m, \sigma)$ si sa densité f est donnée par*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(- \frac{(x-m)^2}{2\sigma^2}\Big), \text{ pour tout } x \in \mathbb{R}$$





Proposition 42 *Si X suit une loi normale $\mathcal{N}(m,\sigma)$, $Z = \frac{X-m}{\sigma}$ suit une loi normale centrée réduite $\mathcal{N}(0,1)$.*

preuve : soient deux réels $a < b$.

$$\begin{aligned} P[a \le Z \le b] &= P\Big[a \le \frac{X-m}{\sigma} \le b\Big] \\ &= P[m + a\sigma \le X \le m + b\sigma] \\ &= \int_{m+a\sigma}^{m+b\sigma} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-m)^2}{2\sigma^2}\Big)\mathrm{d}x \end{aligned}$$

Changement de variable : $z = \frac{x-m}{\sigma}$

$$P[a \le Z \le b] = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z^2}{2}\Big)\mathrm{d}z$$

Cette égalité étant vraie pour tous réels $a < b$, on en déduit que $\frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z^2}{2}\Big)$ est la densité de Z, autrement dit, Z suit une loi normale centrée réduite $\mathcal{N}(0,1)$. □

Proposition 43

$$E[X] = m \text{ et } \mathrm{Var}(X) = \sigma^2$$

preuve : immédiat avec la proposition précédente et la linéarité de l'espérance. □

On peut ainsi facilement se ramener à la loi normale centrée réduite. Comme la fonction de répartition φ de la loi normale centrée réduite est tabulée, on se ramènera même systématiquement à cette loi pour calculer des probabilités. Plus précisément, si X est de loi $\mathcal{N}(m,\sigma)$, et a et b sont deux réels, on calcule $P[a \le X \le b]$ ainsi :

$$P[a \le X \le b] = P\left[\frac{a-m}{\sigma} \le \frac{X-m}{\sigma} \le \frac{b-m}{\sigma}\right] = P\left[\frac{a-m}{\sigma} \le Z \le \frac{b-m}{\sigma}\right]$$

où Z est de loi $\mathcal{N}(0,1)$. Donc

$$P[a \le X \le b] = \varphi\left(\frac{b-m}{\sigma}\right) - \varphi\left(\frac{a-m}{\sigma}\right)$$

On peut aussi remarquer que si Z est de loi $\mathcal{N}(0,1)$, alors $-Z$ aussi et donc, pour $u > 0$,

$$\begin{aligned} P[-u \le Z \le u] &= P[Z \le u] - P[Z \le -u] = P[Z \le u] - P[Z \ge u] \\ &= P[Z \le u] - (1 - P[Z \le u]) = 2P[Z \le u] - 1 \end{aligned}$$

Exemple : Soit X une v.a. de loi $\mathcal{N}(15,4)$. Quelle est la probabilité $P[10 \le X \le 22]$?

$$P[10 \le X \le 22] = P\Big[\frac{10-15}{4} \le \frac{X-15}{4} \le \frac{22-15}{4}\Big] = P[-1.25 \le Y \le 1.75]$$

où Y suit une loi $\mathcal{N}(0,1)$. Ainsi,

$$P[10 \le X \le 22] = P[Y \le 1.75] + P[Y \le 1.25] - 1 = 0.9599 + 0.8944 - 1 = 0.8543$$

Important
Soit Y une v.a. de loi $\mathcal{N}(0,1)$.

$$\begin{aligned} P[-1 \leq Y \leq 1] &= 0.6826 \\ P[-2 \leq Y \leq 2] &= 0.9544 \\ P[-3 \leq Y \leq 3] &= 0.9973 \end{aligned}$$

Soit X une v.a. de loi $\mathcal{N}(m,\sigma)$.

$$\begin{aligned} P[m-\sigma \leq X \leq m+\sigma] &= 0.6826 \\ P[m-2\sigma \leq X \leq m+2\sigma] &= 0.9544 \\ P[m-3\sigma \leq X \leq m+3\sigma] &= 0.9973 \end{aligned}$$

3.4 La loi exponentielle

Soit λ un réel strictement positif.

Définition 44 *Une v.a. X est dite de loi exponentielle $\mathcal{E}(\lambda)$ si sa densité est donnée par*

$$f(x) = \begin{cases} 0 & si\ x < 0 \\ \lambda \exp(-\lambda x) & si\ x \geq 0 \end{cases}$$

La v.a. X de loi $\mathcal{E}(\lambda)$ ne prend que des valeurs positives.
Elle est utilisée dans de nombreuses applications :
- durée de fonctionnement d'un matériel informatique avant la première panne,
- désintégration radioactive,
- temps séparant l'arrivée de deux "clients" dans un phénomène d'attente (guichet, accès à un serveur informatique, arrivée d'un accident du travail...).

Proposition 45 *La fonction de répartition de la loi $\mathcal{E}(\lambda)$ est égale à*

$$F(x) = \begin{cases} 0 & si\ x < 0 \\ 1 - \exp(-\lambda x) & si\ x \geq 0 \end{cases}$$

De plus,

$$E[X] = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

Proposition 46 *La loi exponentielle vérifie la propriété d'absence de mémoire. Soit X un v.a. de loi exponentielle. Alors pour tous $s,t > 0$,*

$$P[X > t+s | X > t] = P[X > s]$$

preuve :

$$P[X > t+s|X > t] = \frac{P[(X > t+s) \cap (X > t)]}{P[X > t]} = \frac{P[X > t+s]}{P[X > t]} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = P[X > s]$$

3.5 Fonction d'une v.a. continue

Soit X une v.a de densité f et ψ une fonction sur $\mathbb{R}$. On pose $Y = \psi(X)$. Cherchons la densité de Y dans le cas où Y est une v.a. continue. Pour cela, on va écrire $P[a \leq Y \leq b]$ sous la forme $\int_a^b g(y)\mathrm{d}y$, pour tous $a < b$. La fonction g ainsi obtenue est la densité de Y. Supposons ψ bijective, de classe C^1, croissante. Dans le cas contraire, on découpera $\mathbb{R}$ en intervalles sur lesquels ψ est bijective (avant de faire le changement de variable ci-dessous). Soit $a < b$ quelconques.

$$P[a \leq Y \leq b] = P[a \leq \psi(X) \leq b] = P[\psi^{-1}(a) \leq X \leq \psi^{-1}(b)] = \int_{\psi^{-1}(a)}^{\psi^{-1}(b)} f(x)\mathrm{d}x$$

On fait le changement de variable $y = \psi(x)$. On obtient

$$P[a \leq Y \leq b] \quad = \quad \int_a^b f\big(\psi^{-1}(y)\big)\frac{1}{\psi'\big(\psi^{-1}(y)\big)}\mathrm{d}y$$

Ainsi, la densité de Y est

$$h(y) = \frac{f\big(\psi^{-1}(y)\big)}{\psi'\big(\psi^{-1}(y)\big)}$$

Exemple : Soit X de densité $f = \frac{1}{2}\mathbb{1}_{[0,2]}$. Soient deux réels c et d avec c non nul. Quelle est la loi de $Y = cX + d$?
Soient $a < b$ quelconques.

$$P[a \leq Y \leq b] \quad = \quad P[a \leq cX + d \leq b] = P\Big[\frac{a-d}{c} \leq X \leq \frac{b-d}{c}\Big] = \frac{1}{2}\int_{\frac{a-d}{c}}^{\frac{b-d}{c}} \mathbb{1}_{[0,2]}(x)\mathrm{d}x$$

On fait le changement de variable $y = cx + d$. Alors $\mathrm{d}y = c\ \mathrm{d}x$ et

$$P[a \leq Y \leq b] = \frac{1}{2c}\int_a^b \mathbb{1}_{[0,2]}((y-d)/c)\mathrm{d}y$$

Or $0 \leq (y-d)/c \leq 2$ ssi $d \leq y \leq 2c + d$, d'où

$$P[a \leq Y \leq b] = \frac{1}{2c}\int_a^b \mathbb{1}_{[d,2c+d]}(y)\mathrm{d}y$$

La densité de Y est donc la fonction g définie par

$$g(y) = \frac{1}{2c}\mathbb{1}_{[d,2c+d]}(y)$$

Théorème 47 (Une transformation bien utile) *Soit X une v.a. continue, de fonction de répartition F. Alors $F(X)$ suit la loi uniforme sur $[0,1]$.*

preuve : même si F n'est pas strictement croissante, on est en mesure de définir une fonction inverse de F par

$$F^{-1}(y) = \min\{x : F(x) \geq y\}$$

Et on obtient, avec la continuité de F :

$$P(F(X) < y) = P(X < F^{-1}(y)) = F(F^{-1}(y)) = y$$

On reconnait la fonction de répartition de la loi uniforme. □

Le résultat suivant permet de simuler de nombreuses lois à partir de nombres uniformément répartis entre 0 et 1.

Théorème 48 *Soit U une v.a. de loi uniforme sur $[0,1]$. Alors $Y = F^{-1}(U)$ a pour fonction de répartition F.*

preuve : D'après la définition de F^{-1}, on a pour tout $0 < x < 1$,

$$F^{-1}(y) \leq x \text{ ssi } F(x) \geq y$$

En remplaçant y par U, on obtient

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

Exemple : pour simuler une réalisation d'une v.a. de la loi exponentielle $\mathcal{E}(\lambda)$, il suffit de calculer $-\log(1-u)/\lambda$ où u est un nombre uniformément réparti entre 0 et 1.

3.6 Exercices

Exercice 1 — Montrer que si X est une variable aléatoire de loi normale $\mathcal{N}(0,1)$, alors la variable aléatoire $Z = \sigma X + m$ suit une loi normale centrée réduite $\mathcal{N}(m,\sigma)$.

Exercice 2 — Soit X une variable aléatoire de loi normale $\mathcal{N}(0,1)$.
Calculer $P[X \leq 1.62]$, $P[X \geq -0.52]$, $P[-1 < X < 1]$. Trouver u tel que $P[X \leq u] = 0.334$.

Exercice 3 — La largeur (en cm) d'une fente dans une pièce fabriquée en aluminium est distribuée selon une loi normale de paramètres $\mu = 2$ et $\sigma = 0,012$. Les limites de tolérance sont données comme étant $2,000 \pm 0,012$. Quel sera le pourcentage de pièces défectueuses ?

Exercice 4 — Soit X une variable aléatoire gaussienne. On sait que :

$$P[X \leq 3] = 0,5517 \qquad \text{et} \qquad P[X \geq 7] = 0,0166$$

Déterminer la moyenne et l'écart-type de X.

Exercice 5 — Dans une usine d'emballage, un automate remplit des paquets de café de 250g. On sait que l'automate verse en fait une quantité de café variable, régie par une loi normale de moyenne réglable et d'écart-type 3. Quelle doit être la moyenne théorique choisie pour que 90% des clients achètent bien au moins 250g de café ?

Exercice 6 — Dans ce problème, les durées des trajets sont supposées de loi normale.
1) Un directeur de société habite dans la ville A. Il part de chez lui à 8h45 et se rend en voiture à son bureau qui ouvre à 9h. La durée de son trajet est, en moyenne, de 13 minutes, avec un écart-type de 3 minutes. Quelle est la probabilité que le directeur arrive en retard ?
2) La secrétaire du directeur habite en A, elle va au bureau avec le train de 8h32 ; elle

descend à la station B. Elle prend ensuite le bus qui part de B à 8h50 (sans attendre le train), pour aller de B à son bureau. La durée du trajet en train a pour moyenne 16 minutes, pour écart-type 2 minutes, et la durée du trajet en bus a pour moyenne 9 minutes et pour écart-type 1 minute. Les durées de trajet en train et en bus sont indépendantes. Quelle est la probabilité que la secrétaire arrive à l'heure ?
3) Quelle est la probabilité pour que le directeur ou la secrétaire (c'est-à-dire l'un au moins des deux) arrive à l'heure, les durées des trajets du directeur ou de la secrétaire étant supposées indépendantes ?

Exercice 7 —Soit X une v.a. continue dont la densité est donnée par

$$f(x) = \begin{cases} a(9x - 3x^2) & \text{pour } 0 < x < 3 \\ 0 & \text{sinon} \end{cases}$$

1) Calculer la constante a.
2) Déterminer $P[X > 1]$ et $P[1/2 < X < 3/2]$.
3) Quelle est la fonction de répartition de X ?
4) Calculer l'espérance et la variance de X.

Exercice 8 — Soit X une v.a. continue de loi uniforme sur $[a, b]$, où a et b sont deux réels positifs tels que $a < b$. Soient c et d positifs avec $a < \sqrt{c} < \sqrt{d} < b$. Écrire $P[c < X^2 < d]$ sous forme d'une intégrale entre c et d. En déduire la densité de la v.a. X^2.

Exercice 9 — Soit X une v.a. de loi exponentielle $\mathcal{E}(1)$. Calculer $P[X \leq 2]$ et $P[X > 0.5]$. Déterminer la densité de $Y = 3X$. De quelle loi s'agit-il ?

Chapitre 4

Théorèmes limites

Considérons une suite $(X_n)_{n\geq 1}$ de v.a. indépendantes et de même loi. Supposons que ces v.a. ont une espérance, notée m et une variance, notée σ^2.

4.1 Loi des grands nombres

Au vu des chapitres précédents, on peut énoncer le théorème suivant.

Théorème 49

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = nm$$

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\sigma^2$$

Théorème 50 (Inégalité de Chebyshev) *Soit X une v.a. et x un réel strictement positif. Alors*

$$P(|X - E(X)| > x) \leq \mathrm{Var}(X)/x^2$$

preuve : Comme $|X - E(X)|^2$ est positif, $|X - E(X)|^2 \geq |X - E(X)|^2 \mathbb{1}_{|X-E(X)|>x}$. On prend maintenant l'espérance (qui conserve l'ordre)

$$\mathrm{Var}(X) = E[|X - E(X)|^2] \geq E[|X - E(X)|^2 \mathbb{1}_{|X-E(X)|>x}] \geq x^2 P(|X - E(X)| > x)$$

et c'est terminé. □

Définition 51 *La moyenne empirique des v.a. $X_1, \ldots, X_n$ est la v.a.*

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

On sait d'ores et déjà que la moyenne empirique a pour espérance m et pour variance σ^2/n. Ainsi, plus n est grand, moins cette v.a. varie. À la limite, quand n tend vers l'infini, elle se concentre sur son espérance, m. C'est la loi des grands nombres.

Théorème 52 (Loi des grands nombres) *Quand n est grand, $\overline{X}_n$ est proche de m avec une forte probabilité. Autrement dit,*

$$\forall \varepsilon > 0, \lim_{n\to\infty} P(|\overline{X}_n - m| > \varepsilon) = 0$$

preuve : soit $\varepsilon > 0$, d'après l'inégalité de Chebyshev,

$$P(|\overline{X}_n - m| > \varepsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \longrightarrow 0$$

On dit que $\overline{X}_n$ converge en probabilité vers m.

Exemple 53 (Fréquence empirique et probabilité) *Il est ici question de rapprocher la probabilité théorique d'un événement A parfois inconnue avec la fréquence empirique de cet événement obtenue en répétant plusieurs fois (mettons n fois) l'expérience aléatoire qui produit A dans les mêmes conditions.*

À la i-ième expérience ($1 \leq i \leq n$), on associe une v.a. Y_i égale à 1 si A est réalisé, 0 sinon. Alors $\sum Y_i$ décrit le nombre de fois où A a été réalisé et $f_A = \frac{1}{n}\sum Y_i$ décrit la fréquence d'observation de A. Comme les Y_i sont supposés de même loi, indépendants, de moyenne $P(A)$ et de variance $P(A)(1 - P(A))$, on peut appliquer la loi des grands nombres pour conclure que f_A converge, quand n tend vers l'infini vers $P(A)$.

Plus généralement, soit X une v.a. dont on ne connaît pas la loi. Alors la loi des grands nombres implique qu'une observation répétée de la v.a. X permet d'avoir une idée plus ou moins précise de la loi de X.

En conséquence, si par exemple on lance un grand nombre de fois un dé, la fréquence d'apparition du 6 doit être proche de 1/6. Des questions restent en suspens : qu'est-ce qu'un "grand nombre de fois", que signifie "proche de 1/6". Le théorème central limite va apporter des réponses.

4.2 Théorème central limite

Théorème 54 *Pour tous réels $a < b$, quand n tend vers $+\infty$,*

$$P\left(a \leq \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \leq b\right) \longrightarrow \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x$$

On dit que $\frac{\overline{X}_n - m}{\sigma/\sqrt{n}}$ converge en loi vers la loi normale $\mathcal{N}(0,1)$. On a le conséquence (presque) immédiate :

Corollaire 55 *Quand n est grand, la loi binomiale $\mathcal{B}(n,p)$ est proche de la loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$*

preuve : Soient $X_1, ..., X_n$ des v.a. de Bernoulli $\mathcal{B}(p)$ indépendantes. Soit $S_n = X_1 + \cdots + X_n$. On sait que S_n suit la loi binomiale $\mathcal{B}(n,p)$. Rappelons que l'espérance commune des X_i est p et leur variance est $p(1-p)$. D'après le théorème central limite, en observant que $\overline{X}_n = S_n/n$, il vient :

$$\frac{S_n - np}{\sqrt{np(1-p)}} \text{ est proche de la loi } \mathcal{N}(0,1)$$

autrement dit, S_n est proche de la loi $\mathcal{N}(np, \sqrt{np(1-p)})$. □

Cette approximation est importante car les probabilités relatives à la loi binomiale sont difficiles à calculer quand n est grand.
Pour améliorer l'approximation, on effectue une "correction de continuité" qui permet de lier loi discrète et loi continue. Ainsi, en pratique, si X est une v.a. de loi binomiale $\mathcal{B}(n,p)$, avec $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, on peut approcher la loi de X par la loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$ de cette manière : pour $0 \leq k \leq n$,

$$P[X \leq k] \simeq P[Y \leq k + 0.5]$$

où Y suit une loi $\mathcal{N}(np, \sqrt{np(1-p)})$.
Et pour tous $a < b$,

$$\begin{aligned} P[a \leq X \leq b] &\simeq P[a - 0.5 \leq Y \leq b + 0.5] \\ &\simeq P\Big[\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}\Big] \\ &\simeq P\Big[\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}\Big] \\ &\simeq \varphi\Big(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\Big) - \varphi\Big(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\Big) \end{aligned}$$

où Z suit une loi normale centrée réduite et φ sa fonction de répartition.
Exemple : on interroge 30 personnes au hasard pour étudier un certain caractère présent dans la population en proportion $p = 0.6$. Soit X le nombre de personnes qui ont ce caractère parmi les 25. Alors X est de loi $\mathcal{B}(30, 0.6)$. Calculons $P[X \leq 15]$. Le calcul exact donne 0.1754. Le calcul approché fournit

$$\varphi((15 + 0.5 - 30 * 0.6)/\sqrt{30 * 0.6 * 0.4}) = 0.1757$$

4.3 Intervalles de confiance

Soit X un caractère (ou variable) étudié sur une population, de moyenne m et de variance σ^2. On cherche ici à donner une estimation de la moyenne m de ce caractère, calculée à partir de valeurs observées sur un échantillon $(X_1, ..., X_n)$.

La fonction de l'échantillon qui estimera un paramètre est appelée **estimateur**, son écart-type est appelé **erreur standard** et est noté SE. L'estimateur de la moyenne m est la moyenne empirique

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

Que savons-nous sur $\overline{X}_n$?
1) $E[\overline{X}_n] = m$,
2) $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$, donc $\mathrm{SE}(\overline{X}_n) = \sigma/\sqrt{n}$,
3) $\overline{X}_n \longrightarrow m$ quand $n \to \infty$,
4) quand n est grand, $\overline{X}_n$ suit approximativement une loi normale $\mathcal{N}(m, \sigma/\sqrt{n})$.

D'après les propriétés de la loi normale, quand n est grand (mettons supérieur à 20), on sait que

$$P[m - 2\sigma/\sqrt{n} \leq \overline{X}_n \leq m + 2\sigma/\sqrt{n}] = 0.954$$

ou, de manière équivalente,

$$P[\overline{X}_n - 2\sigma/\sqrt{n} \leq m \leq \overline{X}_n + 2\sigma/\sqrt{n}] = 0.954$$

Ce qui peut se traduire ainsi : quand on estime m par $\overline{X}_n$, l'erreur faite est inférieure à $2\sigma/\sqrt{n}$, pour 95,4% des échantillons. Ou avec une probabilité de 95,4%, la moyenne inconnue m est dans l'intervalle $[\overline{X}_n - 2\sigma/\sqrt{n}, \overline{X}_n + 2\sigma/\sqrt{n}]$.

Définition 56 *On peut associer à chaque incertitude α, un intervalle, appelé intervalle de confiance de niveau de confiance $1-\alpha$, qui contient la vraie moyenne m avec une probabilité égale à $1-\alpha$.*

Définition 57 *Soit Z une v.a.. Le fractile supérieur d'ordre α de la loi de Z est le réel z qui vérifie*

$$P[Z \geq z] = \alpha$$

Le fractile inférieur d'ordre α de la loi de Z est le réel z qui vérifie

$$P[Z \leq z] = \alpha$$

Proposition 58 *Un intervalle de confiance pour la moyenne, de niveau de confiance $1-\alpha$, est*

$$[\overline{X}_n - z_{\alpha/2}\sigma/\sqrt{n},\ \overline{X}_n + z_{\alpha/2}\sigma/\sqrt{n}]$$

où $z_{\alpha/2}$ est le **fractile** *supérieur d'ordre $\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.*

preuve :

$$\begin{aligned} P[\overline{X}_n - z_{\alpha/2}\sigma/\sqrt{n} \leq m \leq \overline{X}_n + z_{\alpha/2}\sigma/\sqrt{n}] &= P\left[-z_{\alpha/2} \leq \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right] \\ &= P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] \end{aligned}$$

où Z suit une loi normale centrée réduite. Par définition de $z_{\alpha/2}$, cette probabilité vaut bien $1-\alpha$. □

Remarque : soit Z une v.a. de loi $\mathcal{N}(0,1)$, $z_{\alpha/2}$ vérifie

$$P[Z \leq -z_{\alpha/2}] = P[Z \geq z_{\alpha/2}] = \frac{\alpha}{2}, \qquad P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] = 1-\alpha$$

On appelle aussi intervalle de confiance la réalisation de l'intervalle précédent

$$[\bar{x}_n - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{x}_n + z_{\alpha/2}\sigma/\sqrt{n}]$$

Seules quelques valeurs de α sont utilisées habituellement. Les trois valeurs communes sont :
- $\alpha = 0.01$, et $z_{0.005} = 2.58$,
- $\alpha = 0.05$, et $z_{0.025} = 1.96$,
- $\alpha = 0.1$, et $z_{0.05} = 1.645$.

Voici deux exemples, l'un théorique et l'autre pratique.

Exemple 59 *Soit un caractère X qui suit une loi de Poisson de paramètre inconnu λ (ex : nombre de suicides ayant lieu dans le métro lyonnais chaque année). On rappelle que le paramètre d'une loi de Poisson est aussi sa moyenne. On cherche à estimer par un intervalle de confiance le paramètre λ. Supposons que λ vaut 5.2 et générons 25 échantillons de taille 20 de loi de Poisson $\mathcal{P}(5.2)$ pour visualiser des intervalles de confiance qui auraient pu être donnés suite à un échantillonage sur 20 années.*

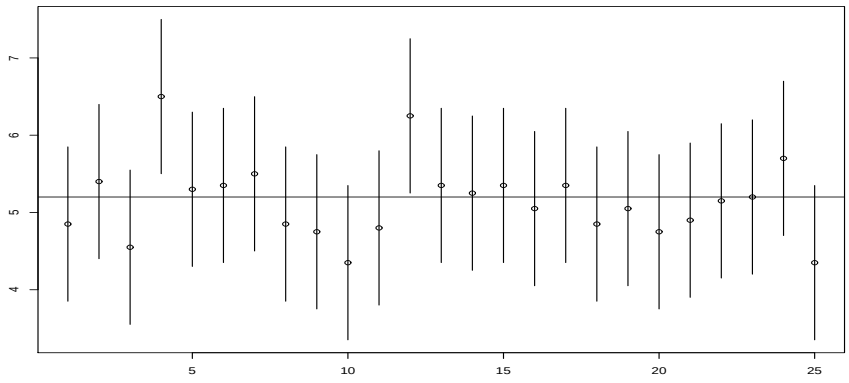


FIGURE 4.1 – Exemple 59 : intervalles de confiance associés à 25 échantillons

Exemple 60 *Voici 30 mesures (en décibels) du bruit occasionné par le trafic routier le long de la nationale 7 :*

57	43	55	59	52	57	50	52	60	49	56	56	52	58	55
58	54	52	56	53	59	50	55	51	54	53	53	56	55	58

Regroupons les différentes modalités

x_i	n_i	f_i	F_i
43	1	0.033	0.033
49	1	0.033	0.066
50	2	0.066	0.133
51	1	0.033	0.167
52	4	0.133	0.3
53	3	0.1	0.4
54	2	0.066	0.466
55	4	0.133	0.6
56	4	0.133	0.733
57	2	0.066	0.8
58	3	0.1	0.9
59	2	0.066	0.966
60	1	0.033	1

Tout d'abord, observons une valeur extrême (43) trop petite : les données n'ont pas l'air

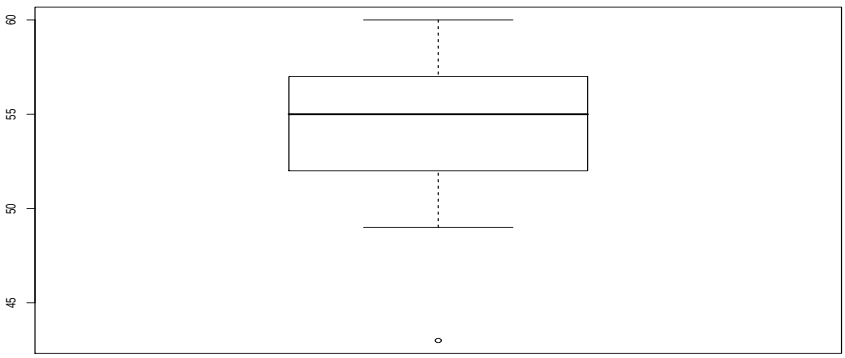


FIGURE 4.2 – Exemple 60

homogène (mesure le dimanche, ou appareil déréglé). Je préfère l'enlever avant l'étude, car elle risque de fausser la moyenne empirique et la variance. Il reste donc 29 observations. On cherche un intervalle de confiance pour le bruit moyen occasionné par le trafic routier. Mais il nous manque l'écart-type de ce bruit. Le calcul est donc impossible.

Quand l'écart-type théorique de la loi du caractère X étudié n'est pas connu, on l'estime par l'écart-type empirique s_{n-1}. Comme on dispose d'un grand échantillon (de taille supérieure à 20), l'erreur commise est petite. L'intervalle de confiance, de niveau de confiance $1-\alpha$ devient :

$$\left[\bar{x}_n - z_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}},\ \bar{x}_n + z_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}}\right]$$

où

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

Exemple 60 (suite) On peut maintenant donner un intervalle de confiance, de niveau de confiance 95%, pour le volume sonore moyen le long de la RN 7. Calculons :

$$\overline{x} = 54.6552, \qquad s_{n-1} = 2.9553$$

d'où l'intervalle

$$\left[54.6552 - \frac{1.96 * 2.9553}{\sqrt{29}}, 54.6552 + \frac{1.96 * 2.9553}{\sqrt{29}}\right] = \left[53.5796, 55.7308\right]$$

Le bruit moyen occasionné par le trafic routier de la nationale 7 est compris entre 53,58 et 55,73, avec un niveau de confiance de 95%.

Tout ce qui concerne les moyennes s'applique aussi aux proportions. On considère une caractéristique que possède une proportion p d'individus dans la population. Soit un échantillon de taille n pris dans la population et X le nombre d'individus dans cet échantillon qui possède la caractéristique étudiée. Alors

$$\hat{p} = \frac{X}{n}$$

est un estimateur de p. De plus, on a déjà vu que X suit une loi binomiale $\mathcal{B}(n,p)$ et que cette loi peut être approchée par une loi normale :

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

suit approximativement une loi normale $\mathcal{N}(0,1)$.
On peut estimer l'écart-type de $\hat{p}$ qui est $\sqrt{p(1-p)/n}$, par $\sqrt{\hat{p}(1-\hat{p})/n}$. On obtient alors un intervalle de confiance pour la proportion p inconnue, de niveau de confiance $1-\alpha$

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Rappel : cet intervalle n'est valide que lorsque $n \geq 20$.

4.4 Exercices

Exercice 1 – Une enquête marketing a pour but de vérifier si les cibles potentielles seraient tentées par un nouveau produit. Il a été montré que 56% des gens sont favorables au nouveau produit. Pour aller plus loin, on interroge à nouveau 200 personnes. Quelle est la loi du nombre de clients potentiels parmi les 200 ? Par quelle loi peut-on l'approcher ? Calculer $P[X > 100]$ et $P[100 \leq X \leq 150]$.

On souhaite maintenant demander des précisions à un grand nombre de personnes favorables au produit, mettons 100 personnes. Déterminer la taille de l'échantillon de personnes à interroger pour que notre échantillon contienne au moins 100 personnes favorables, avec une probabilité supérieure ou égale à 95%.

Exercice 2 – On prélève indépendamment et avec remise n individus d'une population séparée en deux sous-populations A et A^c de proportions respectives p et $1-p$ (clients importants ou petits clients par exemple).
1) Soit K le nombre d'individus de la sous-population A présents dans l'échantillon. Quelle est la loi de K ?
2) Notons $F = K/n$ la fréquence empirique de la catégorie A. Donner l'espérance et la variance de F.
3) Quel est le comportement de F quand n devient grand ?
4) On considère un échantillon de 400 clients d'un organisme de placement financier. Les clients de cet organisme, dans leur ensemble, se répartissent ainsi : 20% possèdent de gros portefeuilles, les autres ne détenant que des portefeuilles modestes. Quelle est la probabilité que la proportion F de gros clients dans l'échantillon soit comprise entre 7% et 23% ? Quelle est la probabilité pour que F soit inférieure à 15% ? Dans l'agence de Lyon (qui compte 400 clients), seuls 15% des clients sont de gros clients. Est-ce acceptable par le siège ?

Exercice 3 – On suppose que le poids d'un nouveau né est une variable normale d'écart-type égal à 0,5 kg. Le poids moyen des 49 enfants nés au mois de janvier 2004 dans l'hôpital de Charleville-Mézières a été de 3,6 kg.
a) Déterminer un intervalle de confiance à 95% pour le poids moyen d'un nouveau né dans cet hôpital.
b) Quel serait le niveau de confiance d'un intervalle de longueur 0,1 kg centré en 3,6 pour ce poids moyen ?

Exercice 4 – On veut étudier la proportion p de gens qui vont au cinéma chaque mois. On prend donc un échantillon de taille $n = 100$. Soit N le nombre de personnes dans l'échantillon qui vont au cinéma mensuellement.
1) Quelle est la loi de N ? Par quelle loi peut-on l'approcher et pourquoi ? En déduire une approximation de la loi de $F = N/n$.
2) On observe une proportion f de gens qui vont chaque mois au cinéma. Donner la forme d'un intervalle de confiance pour p, de niveau de confiance $1 - \alpha$.
3) Applications numériques : $f = 0,1$, $1 - \alpha = 90\%,\ 95\%,\ 98\%$.

Chapitre 5

Tests statistiques

5.1 Tests d'hypothèses

Le but des tests d'hypothèses est de contredire ou au contraire de confirmer une affirmation concernant la population étudiée, en se fondant sur l'observation d'un échantillon. On supposera dans cette section que les échantillons sont de grande taille (mettons supérieure à 20).
Par exemple, on change le traitement médical d'une population de patients. Leur durée de vie s'en trouve-t-elle allongée ? On modifie la composition du carburant d'un moteur. Le rendement est-il significativement amélioré ?

L'hypothèse de travail (durée de vie allongée, rendement amélioré) s'appelle l'**hypothèse alternative** H_1. L'affirmation contraire est appelée **hypothèse nulle** H_0 et elle consiste souvent à supposer que la situation n'a pas évolué. Construire un test revient à établir une règle de décision pour le rejet ou non de H_0 en fonction des observations. Il faut garder à l'esprit que le rejet de H_0 doit être fondé sur des arguments forts, car il devra entraîner une modification des habitudes et donc une importante dépense (par exemple, mettre sur le marché un nouveau médicament, changer de fournisseur de carburant, acheter le nouveau carburant plus cher).

Exemple 61 (Temps de montage) *Un ouvrier spécialisé d'une chaîne de montage passe un temps variable sur chaque pièce. La loi de ce temps a pour moyenne 270 s et pour écart-type 24 s. Une modification technique du montage pourrait diminuer ce temps. Pour le tester, on chronomètre $n = 38$ montages avec la modification technique et on obtient une moyenne empirique $\overline{X}_n$.*
Notons μ le temps moyen que passe l'ouvrier sur chaque pièce en mettant en œuvre la nouvelle technique. La question est de tester l'hypothèse $H_0 : \mu = 270$ contre l'hypothèse $H_1 : \mu < 270$. On doit donc établir quelles sont les valeurs de $\overline{X}_n$ qui vont conduire à rejeter H_0.

Dans cet exemple, la moyenne observée sur l'échantillon devra être significativement différente de 270 pour que H_0 soit rejetée. Dans ce cas, il existera toujours une probabilité de se tromper (H_0 est vraie, mais on la rejette), car le fait de travailler à partir d'un échantillon

entraîne nécessairement une incertitude. On fixe cette probabilité, qui doit être petite, à α avec souvent $\alpha = 0.05$. On dit que **le test est de niveau** α. On pourrait bien sûr diminuer cette probabilité, mais alors la règle de décision nous conduirait à toujours accepter H_0 et ce n'est pas satisfaisant. Il existe par ailleurs une probabilité de se tromper en acceptant H_0, mais nous ne l'étudierons pas.

Il est raisonnable de penser que les échantillons dont la moyenne empirique $\overline{X}_n$ est inférieure à un certain seuil $c < 270$ vont conduire à rejeter H_0. Ainsi, la règle de décision revient à déterminer une région de rejet de H_0 de la forme $\overline{X}_n \leq c$. La probabilité α sera dans ce cas

$$P[\overline{X}_n \leq c] = \alpha \qquad \text{quand } H_0 \text{ est vraie}$$

Une fois α fixée, on peut déterminer c. En effet, on sait que quand H_0 est vraie, $\overline{X}_n$ suit approximativement une loi normale $\mathcal{N}(270, 24/\sqrt{38})$. Ainsi,

$$\alpha = P[\overline{X}_n \leq c] = P\left[\frac{\overline{X}_n - 270}{24/\sqrt{38}} \leq \frac{c-270}{24/\sqrt{38}}\right] = P\left[Z \leq \frac{c-270}{24/\sqrt{38}}\right]$$

où Z suit une loi normale centrée réduite. Or on a noté z_α le fractile supérieur de Z d'ordre α, défini par $P[Z \leq -z_\alpha] = P[Z \geq z_\alpha] = \alpha$. Ainsi, pour $\alpha = 0.05$,

$$\frac{c-270}{24/\sqrt{38}} = -z_\alpha = -z_{0.05} = -1.645$$

ou $c = 263.6$. La région de rejet de H_0 est donc : $\overline{X}_n \leq 263.6$.
On peut aussi décrire la région de rejet en fonction de Z par $Z \leq -z_\alpha$ où $Z = \frac{\overline{X}_n - 270}{24/\sqrt{38}}$.
On appelle Z **statistique de test** (il s'agit d'une grandeur que l'on calcule à partir de l'échantillon et de H_0 et qui permet de prendre la décision).

Une fois la région de rejet définie, il est facile de mesurer, si c'est pertinent, la force avec laquelle on rejette H_0. Reprenons l'exemple précédent. On a fixé un seuil $-z_\alpha$ pour la statistique Z, mais plus la valeur de Z observée sera petite, plus le rejet de H_0 se fera avec force. Ainsi, on calcule la p-valeur (p-value en anglais) qui est la probabilité sous H_0 pour que la statistique de test prenne une valeur plus extrême que celle qu'on observe. Une p-valeur extrêmement petite (inférieure à 0,001) signifie que le test sera extrêmement significatif : on rejette H_0 avec une probabilité de se tromper très proche de 0. Une p-valeur comprise entre 0.001 et 0.01 correspond à un test très significatif. Une p-valeur entre 0.01 et 0.05 correspond à un test significatif et une p-valeur supérieure à 0.05 conduit à accepter H_0.

Exemple 61 Faisons quelques applications numériques. On veut réaliser un test de niveau $\alpha = 0.05$. Alors $-z_\alpha = -1.645$.
- On observe une moyenne empirique calculée sur 38 observations égale à 267. Alors

$$z_{obs} = (267 - 270)/(24/\sqrt{38}) = -0.77 > -z_\alpha$$

donc on accepte H_0.
- On observe une moyenne empirique égale à 260. Alors

$$z_{obs} = (260 - 270)/(24/\sqrt{38}) = -2.57 < -z_\alpha$$

donc on rejette H_0. De plus, la p-valeur vaut $P[Z < -2.57] = 0.005$. Le test est donc très significatif.

Dans la plupart des cas, l'écart-type est inconnu. On l'estime alors, comme pour les intervalles de confiance, par S l'écart-type empirique.

Mise en œuvre d'un test
1- Choix des hypothèse H_0 et H_1.
2- Détermination de la statistique et de la forme de la région de rejet.
3- Choix de α et calcul de la région de rejet.
4- Décision, au vu de l'échantillon et calcul de la p-valeur si nécessaire.

Quand on effectue un test de la moyenne, on peut vouloir tester si la moyenne est différente de la valeur μ_0 contenue dans H_0, ou si elle est supérieure ou inférieure. Dans chacun de ces trois cas, la région de rejet aura une forme différente. Considérons la statistique

$$Z = \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}}$$

où n est la taille de l'échantillon, $\overline{X}_n$ la moyenne empirique et S l'écart-type empirique. Notons z la valeur observée de Z. Voici un récapitulatif :

Hypothèses	Région de rejet	p-valeur
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$Z \geq z_\alpha$	$p = P[Z > z_{obs}]$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$Z \leq -z_\alpha$	$p = P[Z < z_{obs}]$
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$\lvert Z\rvert \geq z_{\alpha/2}$	$p = P[\lvert Z\rvert > z_{obs}]$

5.2 Test d'ajustement du chi-deux

On veut dans cette section tester si une variable X mesurée sur un échantillon suit une loi donnée.

On fixe une partition de $\mathbb{R}$ en K classes notées A_1,...,A_K. On calcule pour chaque classe son effectif théorique et son effectif observé : soit $k = 1, ..., K$,

$$n_{th}(A_k) = nP[X \in A_k] \text{ et } n_{obs}(A_k)$$

Puis on calcule une distance entre ces deux ensembles d'effectifs, définie par

$$d^2_{obs} = \sum_{k=1}^{K} \frac{\Big(n_{obs}(A_k) - n_{th}(A_k)\Big)^2}{n_{th}(A_k)}$$

Si cette distance est grande, on rejettera l'hypothèse H_0. Le seuil est déterminé grâce à une loi continue appelée loi du chi-deux à $K-1$ degrés de liberté et notée χ^2_{K-1}. Ainsi, si c_α est le fractile d'ordre α de cette loi, la région de rejet de H_0 est de la forme

$$d^2_{obs} \geq c_\alpha$$

On peut aussi calculer la p-valeur

$$p = P[\chi^2_{K-1} \geq d^2_{obs}]$$

où χ^2_{K-1} est une v.a. de loi du chi-deux à $K-1$ degrés de liberté et d^2_{obs} est la distance observée.

Comment choisit-on les classes et le nombre K de classes ? Le test est valide quand l'effectif théorique de chaque classe est supérieur à 5. Donc le nombre K de classes doit être inférieur à $n/5$. On choisira $n/5$ si ce nombre est inférieur à 25 et 25 sinon.

Dans un grand nombre de cas, on veut vérifier que l'échantillon provient d'une loi, sans connaître les paramètres de cette loi : on veut vérifier que la loi est normale, ou géométrique, mais sans préciser la valeur de la moyenne ni de la variance. On ne peut plus calculer les effectifs théoriques. On estime alors les paramètres, et on calcule la distance du chi-deux. Le nombre de degrés de liberté sera dans ce cas $K-l-1$ où l est le nombre de paramètres qu'on a dû estimer.

Exemple 62 (Adéquation à une loi de Poisson) *On étudie le nombre de pièces défectueuses dans des lots de 100 pièces. On souhaite tester si ce nombre suit une loi de Poisson. Rappelons qu'une v.a. X suit une loi de Poisson $\mathcal{P}(\lambda)$ si, pour tout $k \in \mathbb{N}$,*

$$P[X = k] = \exp(-\lambda)\frac{\lambda^k}{k!}$$

En pratique, on teste 52 lots et on trouve les effectifs

nombre de pièces défectueuses par lot	*0*	*1*	*2*	*3*	*4*	*5*
nombre de lots	*18*	*18*	*8*	*5*	*2*	*1*

Le paramètre λ d'une loi de Poisson est aussi sa moyenne. On estime donc λ par la moyenne empirique

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{6} n_i x_i = \frac{18\cdot 0 + 18\cdot 1 + 8\cdot 2 + 5\cdot 3 + 2\cdot 4 + 1\cdot 5}{52} = 1.2$$

Dans un premier temps, choisissons la partition

$$A_0 = \{0\}, A_1 = \{1\}, A_2 = \{2\}, A_3 = \{3\}, A_4 = \{4\}, A_5 = \{\textit{le reste}\}$$

Les effectifs théoriques sont

classe	A_0	A_1	A_2	A_3	A_4	A_5
n_{th}	*15.7*	*18.8*	*11.3*	*4.5*	*1.4*	*0.3*

Or chacun des effectifs théoriques doit être supérieur à 5. Il faut donc regrouper les 3 dernières classes et on obtient le tableau

classe	A_0	A_1	A_2	A_3
n_{th}	*15.7*	*18.8*	*11.3*	*6.2*
n_{obs}	*18*	*18*	*8*	*8*

La distance du chi-deux est maintenant facile à calculer

$$d^2_{obs} = \frac{(18-15.7)^2}{15.7} + \frac{(18-18.8)^2}{18.8} + \frac{(8-11.3)^2}{11.3} + \frac{(8-6.2)^2}{6.2} = 1.86$$

Pour tester l'hypothèse que la loi du nombre de pièces défecteuses par lot suit une loi de Poisson, au risque $\alpha = 0.05$, il nous manque encore le fractile du chi-deux à 2 degrés de liberté. Les tables indiquent

$$P[\chi^2_2 \geq 5.99] = 0.05$$

La valeur observée étant plus petite, on accepte l'hypothèse Poisson, au risque 5%.

Exemple 63 (Adéquation à la loi normale) *Le satellite Landsat a mesuré la lumière, dans le proche infrarouge, réfléchie par des zones urbanisées. Voici des points de relevés thermographiques*

mesures	*71*	*72*	*73*	*74*	*75*	*77*	*78*	*79*	*80*	*81*	*82*	*84*
effectifs observés	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*4*	*3*	*3*	*4*	*6*
mesures	*85*	*86*	*87*	*88*	*90*	*91*	*94*					
effectifs observés	*6*	*2*	*1*	*1*	*1*	*1*	*1*					

On veut tester si cette série peut provenir d'une loi normale. On a besoin d'estimer deux paramètres, la moyenne (par la moyenne empirique) et l'écart-type (par l'écart-type empirique).

$$\bar{x}_{40} = 82.075 \text{ et } s_{40} = 4.98$$

On dispose d'un échantillon de taille 40. Le nombre de classes sera donc 8. On les choisit de telle sorte que chacune d'elle ait la probabilité $1/8$, de telle sorte que les effectifs théoriques soient tous égaux à 5.

La partition pour la loi normale $\mathcal{N}(0,1)$ est obtenue en lisant les tables

$$-1.15 \qquad -0.67 \qquad -0.32 \qquad 0 \qquad 0.32 \qquad 0.67 \qquad 1.15$$

ce qui donne pour la loi $\mathcal{N}(82.075, 4.98)$

$$76.35 \qquad 78.72 \qquad 80.49 \qquad 82.075 \qquad 83.66 \qquad 85.43 \qquad 87.8$$

On obtient donc le tableau

classe	n_{th}	n_{obs}
$[-\infty, 76.3]$	*5*	*5*
$[76.3, 78.7]$	*5*	*2*
$[78.7, 80.5]$	*5*	*7*
$[80.5, 82.7]$	*5*	*7*
$[82.7, 83.7]$	*5*	*0*
$[83.7, 85.4]$	*5*	*12*
$[85.4, 87.8]$	*5*	*3*
$[87.8, +\infty]$	*5*	*4*

La distance du chi-deux vaut alors

$$d^2_{obs} = \frac{(5-5)^2}{5}+\frac{(2-5)^2}{5}+\frac{(7-5)^2}{5}+\frac{(7-5)^2}{5}+\frac{(0-5)^2}{5}+\frac{(12-5)^2}{5}+\frac{(3-5)^2}{5}+\frac{(4-5)^2}{5} = 19.2$$

Or $P[\chi^2_5 \geq 11.07] = 0.05$.
On a une valeur observée ici beaucoup plus grande, donc on rejette H_0 *: on décide que la loi observée n'est pas une loi normale. De plus, la p-valeur est égale à*

$$P[\chi^2_5 \geq 19.2] = 0.0018$$

Le test est donc très significatif.

5.3 Test d'indépendance du chi-deux

On étudie deux caractères sur une population et on veut tester si ces caractères sont indépendants. Par exemple, la couleur des yeux et la couleur des cheveux sont-elles indépendantes ? Le fait d'être pour la hausse du tabac et le fait de voter à droite sont-ils indépendants ?

Exemple 64 *On a relevé sur un échantillon de taille 100 la méthode de paiement d'automobilistes et le genre de voitures achetées. En voici la table de contingence*

		genre de voiture		
		neuve	*d'occasion*	
paiement	*comptant*	*15*	*5*	*20*
	à crédit	*45*	*35*	*80*
		60	*40*	*100*

Notons $x_1, ..., x_r$ les r modalités du premier caractère X et $y_1, ..., y_s$ les s modalités du deuxième caractère Y. On suppose les caractères qualitatifs ou discrets. Dans le cas continu, on devra fabriquer des classes comme à la précédente section.

Si les deux caractères sont indépendants, on a pour tout $1 \leq i \leq r$ et $1 \leq j \leq s$

$$P[X = x_i, Y = y_j] = P[X = x_i]P[Y = y_j]$$

et les fréquences observées f_{ij} sont proches de $P[X = x_i]P[Y = y_j]$. On estime ces probabilités (appelées marginales) par $\hat{f}_{i.}$, la fréquence de x_i, soit $n_{i.}/n$ et $\hat{f}_{.j}$ la fréquence de y_j, soit $n_{.j}/n$. Et on calcule la distance du chi-deux

$$D^2 = \sum_{i,j} \frac{(n_{ij} - n\hat{f}_{i.}\hat{f}_{.j})^2}{n\hat{f}_{i.}\hat{f}_{.j}} = \sum_{i,j} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n} = \sum_{\text{cases}} \frac{(n_{obs} - n_{th})^2}{n_{th}}$$

On a dû estimer $s - 1 + r - 1$ probabilités et on travaille sur rs cases, donc cette distance suit une loi du chi-deux à $rs - 1 - s - r + 2 = (r - 1)(s - 1)$ degrés de liberté.

Exemple 64 La distance du chi-deux observée est

$$\begin{aligned} D^2 &= \frac{(15 - 20 * 60/100)^2}{20 * 60/100} + \frac{(5 - 20 * 40/100)^2}{20 * 40/100} \\ &\quad + \frac{(45 - 60 * 80/100)^2}{60 * 80/100} + \frac{(35 - 40 * 80/100)^2}{40 * 80/100} \\ &= \frac{9}{12} + \frac{9}{8} + \frac{9}{48} + \frac{9}{32} = 2.34 \end{aligned}$$

pour 1 degré de liberté. Dans le cas où on a un seul degré de liberté, on effectue une correction dite correction de Yates et on calcule

$$D^2 = \sum_{\text{cases}} \frac{(|n_{obs} - n_{th}| - 0.5)^2}{n_{th}}$$

et on obtient

$$D^2 = \frac{(3 - 0.5)^2}{12} + \frac{(3 - 0.5)^2}{8} + \frac{(3 - 0.5)^2}{48} + \frac{(3 - 0.5)^2}{32} = 1.63$$

De plus, $P[\chi_1^2 \geq 3.84] = 0.05$ et 3.84 est bien plus grand que la valeur observée 1.63. On accepte donc l'hypothèse d'indépendance H_0 : il n'y a pas de relation entre le genre de voiture achetée et le mode de paiement.

5.4 Exercices

Exercice 1 – Dans un centre avicole, des études antérieures ont montré que la masse d'un oeuf choisi au hasard peut être considérée comme la réalisation d'une variable aléatoire normale X, de moyenne m et de variance σ^2. On admet que les masses des oeufs sont indépendantes les unes des autres. On prend un échantillon de $n = 36$ oeufs que l'on pèse. Les mesures sont données (par ordre croissant) dans le tableau suivant :

50, 34	52, 62	53, 79	54, 99	55, 82	57, 67
51, 41	53, 13	53, 89	55, 04	55, 91	57, 99
51, 51	53, 28	54, 63	55, 12	55, 95	58, 10
52, 07	53, 30	54, 76	55, 24	57, 05	59, 30
52, 22	53, 32	54, 78	55, 28	57, 18	60, 58
52, 38	53, 39	54, 93	55, 56	57, 31	63, 15

a) Calculer la moyenne empirique et l'écart-type empirique de cette série statistique. Tracer le boxplot et un histogramme.
b) Donner une estimation des paramètres m et σ.
c) Donner un intervalle de confiance au niveau 95%, puis 98%, de la masse moyenne m d'un oeuf.
d) Tester si la moyenne de cette variable est égale à 56.

Exercice 2 – Un appareil de télécommunications reçoit un signal stocké à chaque (petite) unité de temps dans une suite de variables (X_n). Cet appareil doit détecter un signal effectif, en le différenciant d'un bruit. On suppose que le bruit est une suite de variables indépendantes de loi normale de moyenne nulle et de variance 1. Pour un signal, la moyenne n'est pas nulle.
Aujourd'hui on a observé une suite de 40 variables $(x_1, \ldots, x_{40})$, supposées indépendantes, de variance 1. La moyenne empirique vaut $0,6$. S'agit-il de bruit ? Construire un test pour répondre à cette question.

Exercice 3 – On utilise une nouvelle variété de pommes de terre dans une exploitation agricole. Le rendement de l'ancienne variété était de 41.5 tonnes à l'ha. La nouvelle est cultivée sur 100 ha, avec un rendement moyen de 45 tonnes à l'ha et un écart-type de 11.25. Faut-il, au vu de ce rendement, favoriser la culture de la nouvelle variété ?

Exercice 4 – Dans une agence de location de voitures, le patron veut savoir quelles sont les voitures qui n'ont roulé qu'en ville pour les revendre immédiatement. Pour cela, il y a dans chaque voiture une boîte noire qui enregistre le nombre d'heures pendant lesquelles la voiture est restée au point mort, au premier rapport, au deuxième rapport,..., au cinquième rapport. On sait qu'une voiture qui ne roule qu'en ville passe en moyenne 10% de son temps au point mort, 5% en première, 30% en seconde, 30% en troisième, 20% en quatrième, et 5% en cinquième. On décide de faire un test du χ^2 pour savoir si une voiture n'a roulé qu'en ville ou non.
1) Sur une première voiture, on constate sur 2000 heures de conduite : 210 h au point mort, 94 h en première, 564 h en seconde, 630 h en troisième, 390 h en quatrième, et 112 h en cinquième. Cette voiture n'a-t-elle fait que rester en ville ?
2) Avec une autre voiture, on obtient les données suivantes : 220 h au point mort, 80 h en première, 340 h en seconde, 600 h en troisième, 480 h en quatrième et 280 h en cinquième.

Exercice 5 – Une chaîne d'agences immobilières cherche à vérifier que le nombre de biens vendus par agent par mois suit une loi de Poisson de paramètre $\lambda = 1,5$.
1) On observe 52 agents pendant un mois dans la moitié nord de la France. On trouve la répartition suivante : 18 agents n'ont rien vendu, 18 agents ont vendu 1 bien, 8 agents ont vendu 2 biens, 5 agents ont vendu 3 biens, 2 agents ont vendus 4 biens, et un agent a vendu 5 biens. Avec un test du χ^2, chercher s'il s'agit bien de la loi de Poisson attendue.
2) Répondre à la même question avec les 52 agents dans la moitié sud de la France : 19

agents n'ont rien vendu, 20 agents ont vendu un bien, 7 agents 2 biens, 5 agents 3 biens et 1 agent 6 biens.

Exercice 6 – On teste un médicament X destiné à soigner une maladie en phase terminale. On traite des patients avec ce médicament tandis que d'autres reçoivent un placebo ("contrôle"). On note dans la variable statut si les patients ont survécu plus de 48 jours. Voici le tableau obtenu

	statut	
traitement	non	oui
contrôle	17	29
X	7	38

Conclusion ?

Exercice 7 – On mesure la taille du lobe frontal de 30 crabes *Leptograpsus variegatus*. Voici les 30 longueurs obtenues :

12.6 12.0 20.9 14.2 16.2 15.3 10.4 22.1 19.8 15 12.8 20 11.8 20.6 21.3

11.7 18 9.1 15 15.2 15.1 14.7 13.3 21.7 15.4 16.7 15.6 17.1 7.2 12.6

Est-ce que cette variable suit une loi normale ?

Annexe A

Cardinaux et dénombrement

Définition 65 *Soit A un ensemble fini. Le cardinal de A, noté $|A|$, est le nombre d'éléments que contient A.*

Proposition 66 (Additivité) *Soient A et B deux ensembles finis, disjoints (c'est-à-dire $A \cap B = \emptyset$). Alors*

$$|A \cup B| = |A| + |B|$$

Proposition 67 (Inclusion-exclusion) *Soient A et B deux ensembles finis.*

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Proposition 68 (Multiplicativité) *Soient A et B deux ensembles finis. Alors $|A \times B| = |A| \cdot |B|$*

Corollaire 69 *Soit A un ensemble fini de cardinal n. Le nombre de suites de longueur r constituées d'éléments de A est n^r.*

Théorème 70 (Principe du dénombrement) *On réalise deux expériences qui peuvent produire respectivement n et m résultats différents. Au total, pour les deux expériences prises ensemble, il existe $n.m$ résultats possibles.*

Définition 71 *Soit A un ensemble fini. Une permutation de A est une manière d'ordonner, d'arranger les éléments de A.*

Théorème 72 *Il y a $n!$ permutations d'un ensemble de cardinal n.*

preuve : clair par le principe du dénombrement. ♣
exemple : combien existe-t-il d'anagrammes de PROBA ?

Théorème 73 *Soient n objets distinguables. Le nombre de permutations de r objets, pris parmi les n objets, est*

$$A_n^r = \frac{n!}{(n-r)!}$$

(on dit aussi arrangement de r objets pris parmi n)
preuve : pour la première place, il y a n objets possibles,
pour la seconde, $(n-1)$ objets possibles,
...
pour la dernière, $(n-r+1)$ objets possibles.
Au total, $n(n-1)...(n-r+1)$ possibilités, par le principe du dénombrement. ♣

Théorème 74 *Le nombre de manières de choisir p éléments parmi n (sans tenir compte de l'ordre) est*

$$\binom{n}{p} = \frac{n!}{p!(n-p)!}$$

Autrement dit, c'est le nombre de parties à p éléments pris parmi n éléments. On appelle parfois ces parties des combinaisons de p éléments pris parmi n.

preuve : on regarde le nombre de permutations de ces p éléments et on obtient $p!$ arrangements. Il y a donc $p!$ fois plus d'arrangements que de combinaisons. ♣

Proposition 75 *1)* $\binom{n}{p} = \binom{n}{n-p}$
2) $\binom{n}{p} = \binom{n-1}{p} + \binom{n-1}{p-1}$
3) $(x+y)^n = \sum_{p=0}^{n} \binom{n}{p} x^p y^{n-p}$

Corollaire 76 *Soit Ω un ensemble fini de cardinal n. Le cardinal de $\mathcal{P}(\Omega)$ vaut 2^n.*

preuve : il existe 1 partie à 0 élément,
il existe n parties à 1 élément,
...
il existe $\binom{n}{p}$ parties à p éléments,
...
il existe 1 partie à n éléments.
Finalement, le nombre total de parties est

$$\sum_{p=0}^{n} \binom{n}{p} = \sum_{p=0}^{n} \binom{n}{p} 1^r 1^{n-r} = (1+1)^n = 2^n$$

♣

Théorème 77 *On considère n objets, parmi lesquels n_1 sont indistinguables, n_2 sont indistinguables,..., n_r sont aussi indistinguables. Le nombre de permutations différentes est* $\dfrac{n!}{n_1!n_2!\cdots n_r!}$

exemple : combien d'anagrammes de STAT ? 4 !/2 !=12

Tirage avec remise
On dispose d'une population de n objets, par exemple des boules numérotées de 1 à n dans une urne. On effectue un tirage avec remise de r boules parmi les n boules : on pioche une boule, on note son numéro, on la remet dans l'urne avant de repiocher la boule suivante,

et ainsi de suite. On a n possibilités pour la première boule tirée, n pour la seconde... Finalement, on a n^r tirages différents.

Tirage sans remise

On tire toujours r boules parmi n, mais sans remise : on ne replace pas la boule dans l'urne avant de tirer la suivante. Ainsi, on a n possibilités pour la première boule tirée, $n-1$ pour la seconde... Finalement, on a A_n^r tirages différents.

Tirage exhaustif

Cette fois, on tire r boules d'un coup et on se retrouve avec un "tas" de r boules devant soi. Le nombre de tirages différents est $\binom{n}{r}$.

Il faut remarquer que **les deux derniers tirages sont équivalents** : c'est juste une modélisation différente de la manipulation, une manière différente d'écrire les résultats des tirages.

exemple : résultat du loto (6 numéros parmi 49).

- manière de voir 1 : on regarde en direct le tirage du loto et on obtient un arrangement de 6 nombres pris dans $\{1, ..., 49\}$. On a alors $\omega = (x_1, ..., x_6)$: les 6 nombres sortis avec leur ordre d'arrivée. Quel est le nombre de tirages différents ?

$$A_{49}^6 = 49 * 48 * 47 * 46 * 45 * 44 = 10.068.347.520$$

Mais on peut gagner les 6 bons numéros quel que soit l'ordre de sortie des 6 numéros...

- manière de voir 2 : on regarde les 6 nombres sortis sans s'occuper de l'ordre d'arrivée. On a alors $\omega = \{x_1, ..., x_6\}$. D'où Ω est l'ensemble des combinaisons de 6 nombres pris dans $\{1, ..., 49\}$. Quel est le nombre de tirages différents ?

$$\binom{49}{6} = \frac{49 * 48 * 47 * 46 * 45 * 44}{6 * 5 * 4 * 3 * 2} = 13.983.816$$

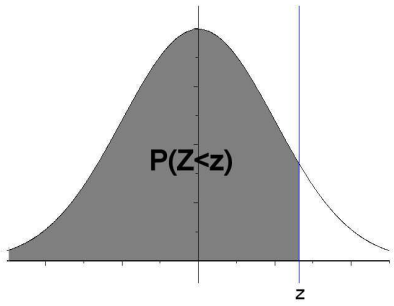
remarque : $(1, 2, 3, 4, 5, 6) \neq (2, 1, 3, 4, 5, 6)$, mais $\{1, 2, 3, 4, 5, 6\} = \{2, 1, 3, 4, 5, 6\}$

Annexe B

Tables statistiques

B.1 Fonction de répartition de la loi normale centrée réduite

$$F(z) = P[Z \leq z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\mathrm{d}x$$



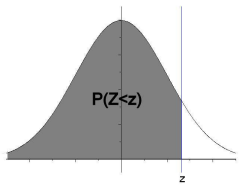
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	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	0.8413	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	0.9772	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

Grandes valeurs de z :

z	3.0	3.1	3.2	3.3	3.4	3.5	4.0	4.5
$F(z)$	0.9987	0.99904	0.99931	0.99952	0.99966	0.99976	0.999968	0.999997

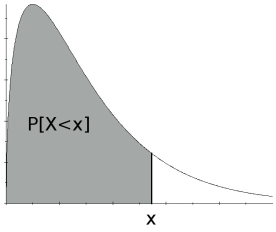
B.2 Fractiles de la loi normale centrée réduite

(attention : si $P[Z < z] < 0.5$, z est négatif)



$P[Z < z]$	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.010	
0.00	Inf	3.0902	2.8782	2.7478	2.6521	2.5758	2.5121	2.4573	2.4089	2.3656	2.3263	0.99
0.01	2.3263	2.2904	2.2571	2.2262	2.1973	2.1701	2.1444	2.1201	2.0969	2.0749	2.0537	0.98
0.02	2.0537	2.0335	2.0141	1.9954	1.9774	1.96	1.9431	1.9268	1.911	1.8957	1.8808	0.97
0.03	1.8808	1.8663	1.8522	1.8384	1.825	1.8119	1.7991	1.7866	1.7744	1.7624	1.7507	0.96
0.04	1.7507	1.7392	1.7279	1.7169	1.706	1.6954	1.6849	1.6747	1.6646	1.6546	1.6449	0.95
0.05	1.6449	1.6352	1.6258	1.6164	1.6072	1.5982	1.5893	1.5805	1.5718	1.5632	1.5548	0.94
0.06	1.5548	1.5464	1.5382	1.5301	1.522	1.5141	1.5063	1.4985	1.4909	1.4833	1.4758	0.93
0.07	1.4758	1.4684	1.4611	1.4538	1.4466	1.4395	1.4325	1.4255	1.4187	1.4118	1.4051	0.92
0.08	1.4051	1.3984	1.3917	1.3852	1.3787	1.3722	1.3658	1.3595	1.3532	1.3469	1.3408	0.91
0.09	1.3408	1.3346	1.3285	1.3225	1.3165	1.3106	1.3047	1.2988	1.293	1.2873	1.2816	0.90
0.10	1.2816	1.2759	1.2702	1.2646	1.2591	1.2536	1.2481	1.2426	1.2372	1.2319	1.2265	0.89
0.11	1.2265	1.2212	1.216	1.2107	1.2055	1.2004	1.1952	1.1901	1.185	1.18	1.175	0.88
0.12	1.175	1.17	1.165	1.1601	1.1552	1.1503	1.1455	1.1407	1.1359	1.1311	1.1264	0.87
0.13	1.1264	1.1217	1.117	1.1123	1.1077	1.1031	1.0985	1.0939	1.0893	1.0848	1.0803	0.86
0.14	1.0803	1.0758	1.0714	1.0669	1.0625	1.0581	1.0537	1.0494	1.045	1.0407	1.0364	0.85
0.15	1.0364	1.0322	1.0279	1.0237	1.0194	1.0152	1.011	1.0069	1.0027	0.9986	0.9945	0.84
0.16	0.9945	0.9904	0.9863	0.9822	0.9782	0.9741	0.9701	0.9661	0.9621	0.9581	0.9542	0.83
0.17	0.9542	0.9502	0.9463	0.9424	0.9385	0.9346	0.9307	0.9269	0.923	0.9192	0.9154	0.82
0.18	0.9154	0.9116	0.9078	0.904	0.9002	0.8965	0.8927	0.889	0.8853	0.8816	0.8779	0.81
0.19	0.8779	0.8742	0.8705	0.8669	0.8633	0.8596	0.856	0.8524	0.8488	0.8452	0.8416	0.80
0.20	0.8416	0.8381	0.8345	0.831	0.8274	0.8239	0.8204	0.8169	0.8134	0.8099	0.8064	0.79
0.21	0.8064	0.803	0.7995	0.7961	0.7926	0.7892	0.7858	0.7824	0.779	0.7756	0.7722	0.78
0.22	0.7722	0.7688	0.7655	0.7621	0.7588	0.7554	0.7521	0.7488	0.7454	0.7421	0.7388	0.77
0.23	0.7388	0.7356	0.7323	0.729	0.7257	0.7225	0.7192	0.716	0.7128	0.7095	0.7063	0.76
0.24	0.7063	0.7031	0.6999	0.6967	0.6935	0.6903	0.6871	0.684	0.6808	0.6776	0.6745	0.75
0.25	0.6745	0.6713	0.6682	0.6651	0.662	0.6588	0.6557	0.6526	0.6495	0.6464	0.6433	0.74
0.26	0.6433	0.6403	0.6372	0.6341	0.6311	0.628	0.625	0.6219	0.6189	0.6158	0.6128	0.73
0.27	0.6128	0.6098	0.6068	0.6038	0.6008	0.5978	0.5948	0.5918	0.5888	0.5858	0.5828	0.72
0.28	0.5828	0.5799	0.5769	0.574	0.571	0.5681	0.5651	0.5622	0.5592	0.5563	0.5534	0.71
0.29	0.5534	0.5505	0.5476	0.5446	0.5417	0.5388	0.5359	0.533	0.5302	0.5273	0.5244	0.70
0.30	0.5244	0.5215	0.5187	0.5158	0.5129	0.5101	0.5072	0.5044	0.5015	0.4987	0.4959	0.69
0.31	0.4959	0.493	0.4902	0.4874	0.4845	0.4817	0.4789	0.4761	0.4733	0.4705	0.4677	0.68
0.32	0.4677	0.4649	0.4621	0.4593	0.4565	0.4538	0.451	0.4482	0.4454	0.4427	0.4399	0.67
0.33	0.4399	0.4372	0.4344	0.4316	0.4289	0.4261	0.4234	0.4207	0.4179	0.4152	0.4125	0.66
0.34	0.4125	0.4097	0.407	0.4043	0.4016	0.3989	0.3961	0.3934	0.3907	0.388	0.3853	0.65
0.35	0.3853	0.3826	0.3799	0.3772	0.3745	0.3719	0.3692	0.3665	0.3638	0.3611	0.3585	0.64
0.36	0.3585	0.3558	0.3531	0.3505	0.3478	0.3451	0.3425	0.3398	0.3372	0.3345	0.3319	0.63
0.37	0.3319	0.3292	0.3266	0.3239	0.3213	0.3186	0.316	0.3134	0.3107	0.3081	0.3055	0.62
0.38	0.3055	0.3029	0.3002	0.2976	0.295	0.2924	0.2898	0.2871	0.2845	0.2819	0.2793	0.61
0.39	0.2793	0.2767	0.2741	0.2715	0.2689	0.2663	0.2637	0.2611	0.2585	0.2559	0.2533	0.60
0.40	0.2533	0.2508	0.2482	0.2456	0.243	0.2404	0.2378	0.2353	0.2327	0.2301	0.2275	0.59
0.41	0.2275	0.225	0.2224	0.2198	0.2173	0.2147	0.2121	0.2096	0.207	0.2045	0.2019	0.58
0.42	0.2019	0.1993	0.1968	0.1942	0.1917	0.1891	0.1866	0.184	0.1815	0.1789	0.1764	0.57
0.43	0.1764	0.1738	0.1713	0.1687	0.1662	0.1637	0.1611	0.1586	0.156	0.1535	0.151	0.56
0.44	0.151	0.1484	0.1459	0.1434	0.1408	0.1383	0.1358	0.1332	0.1307	0.1282	0.1257	0.55
0.45	0.1257	0.1231	0.1206	0.1181	0.1156	0.113	0.1105	0.108	0.1055	0.103	0.1004	0.54
0.46	0.1004	0.0979	0.0954	0.0929	0.0904	0.0878	0.0853	0.0828	0.0803	0.0778	0.0753	0.53
0.47	0.0753	0.0728	0.0702	0.0677	0.0652	0.0627	0.0602	0.0577	0.0552	0.0527	0.0502	0.52
0.48	0.0502	0.0476	0.0451	0.0426	0.0401	0.0376	0.0351	0.0326	0.0301	0.0276	0.0251	0.51
0.49	0.0251	0.0226	0.0201	0.0175	0.015	0.0125	0.01	0.0075	0.005	0.0025	0.0000	0.50
	0.010	0.009	0.008	0.007	0.006	0.005	0.004	0.003	0.002	0.001	0.000	$P[Z < z]$

B.3 Fractiles de la loi du χ^2 (ν = nombre de degrés de liberté)



ν \ P	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.999
1	0.016	0.064	0.148	0.275	0.455	0.708	1.074	1.642	2.706	3.841	5.024	6.635	7.879	10.828
2	0.211	0.446	0.713	1.022	1.386	1.833	2.408	3.219	4.605	5.991	7.378	9.21	10.597	13.816
3	0.584	1.005	1.424	1.869	2.366	2.946	3.665	4.642	6.251	7.815	9.348	11.345	12.838	16.266
4	1.064	1.649	2.195	2.753	3.357	4.045	4.878	5.989	7.779	9.488	11.143	13.277	14.86	18.467
5	1.61	2.343	3	3.655	4.351	5.132	6.064	7.289	9.236	11.07	12.833	15.086	16.75	20.515
6	2.204	3.07	3.828	4.57	5.348	6.211	7.231	8.558	10.645	12.592	14.449	16.812	18.548	22.458
7	2.833	3.822	4.671	5.493	6.346	7.283	8.383	9.803	12.017	14.067	16.013	18.475	20.278	24.322
8	3.49	4.594	5.527	6.423	7.344	8.351	9.524	11.03	13.362	15.507	17.535	20.09	21.955	26.124
9	4.168	5.38	6.393	7.357	8.343	9.414	10.656	12.242	14.684	16.919	19.023	21.666	23.589	27.877
10	4.865	6.179	7.267	8.295	9.342	10.473	11.781	13.442	15.987	18.307	20.483	23.209	25.188	29.588
11	5.578	6.989	8.148	9.237	10.341	11.53	12.899	14.631	17.275	19.675	21.92	24.725	26.757	31.264
12	6.304	7.807	9.034	10.182	11.34	12.584	14.011	15.812	18.549	21.026	23.337	26.217	28.3	32.909
13	7.042	8.634	9.926	11.129	12.34	13.636	15.119	16.985	19.812	22.362	24.736	27.688	29.819	34.528
14	7.79	9.467	10.821	12.078	13.339	14.685	16.222	18.151	21.064	23.685	26.119	29.141	31.319	36.123
15	8.547	10.307	11.721	13.03	14.339	15.733	17.322	19.311	22.307	24.996	27.488	30.578	32.801	37.697
16	9.312	11.152	12.624	13.983	15.338	16.78	18.418	20.465	23.542	26.296	28.845	32	34.267	39.252
17	10.085	12.002	13.531	14.937	16.338	17.824	19.511	21.615	24.769	27.587	30.191	33.409	35.718	40.79
18	10.865	12.857	14.44	15.893	17.338	18.868	20.601	22.76	25.989	28.869	31.526	34.805	37.156	42.312
19	11.651	13.716	15.352	16.85	18.338	19.91	21.689	23.9	27.204	30.144	32.852	36.191	38.582	43.82
20	12.443	14.578	16.266	17.809	19.337	20.951	22.775	25.038	28.412	31.41	34.17	37.566	39.997	45.315
21	13.24	15.445	17.182	18.768	20.337	21.991	23.858	26.171	29.615	32.671	35.479	38.932	41.401	46.797
22	14.041	16.314	18.101	19.729	21.337	23.031	24.939	27.301	30.813	33.924	36.781	40.289	42.796	48.268
23	14.848	17.187	19.021	20.69	22.337	24.069	26.018	28.429	32.007	35.172	38.076	41.638	44.181	49.728
24	15.659	18.062	19.943	21.652	23.337	25.106	27.096	29.553	33.196	36.415	39.364	42.98	45.559	51.179
25	16.473	18.94	20.867	22.616	24.337	26.143	28.172	30.675	34.382	37.652	40.646	44.314	46.928	52.62
26	17.292	19.82	21.792	23.579	25.336	27.179	29.246	31.795	35.563	38.885	41.923	45.642	48.29	54.052
27	18.114	20.703	22.719	24.544	26.336	28.214	30.319	32.912	36.741	40.113	43.195	46.963	49.645	55.476
28	18.939	21.588	23.647	25.509	27.336	29.249	31.391	34.027	37.916	41.337	44.461	48.278	50.993	56.892
29	19.768	22.475	24.577	26.475	28.336	30.283	32.461	35.139	39.087	42.557	45.722	49.588	52.336	58.301
30	20.599	23.364	25.508	27.442	29.336	31.316	33.53	36.25	40.256	43.773	46.979	50.892	53.672	59.703
31	21.434	24.255	26.44	28.409	30.336	32.349	34.598	37.359	41.422	44.985	48.232	52.191	55.003	61.098
32	22.271	25.148	27.373	29.376	31.336	33.381	35.665	38.466	42.585	46.194	49.48	53.486	56.328	62.487
33	23.11	26.042	28.307	30.344	32.336	34.413	36.731	39.572	43.745	47.4	50.725	54.776	57.648	63.87
34	23.952	26.938	29.242	31.313	33.336	35.444	37.795	40.676	44.903	48.602	51.966	56.061	58.964	65.247
35	24.797	27.836	30.178	32.282	34.336	36.475	38.859	41.778	46.059	49.802	53.203	57.342	60.275	66.619
40	29.051	32.345	34.872	37.134	39.335	41.622	44.165	47.269	51.805	55.758	59.342	63.691	66.766	73.402
45	33.35	36.884	39.585	41.995	44.335	46.761	49.452	52.729	57.505	61.656	65.41	69.957	73.166	80.077
50	37.689	41.449	44.313	46.864	49.335	51.892	54.723	58.164	63.167	67.505	71.42	76.154	79.49	86.661

Annexe C

Statistique descriptive univariée

Nous allons voir ici comment décrire simplement des données en trois phases :
- **présentation des données**,
- **représentations graphiques**,
- calcul de **résumés numériques**.

On considère une **population** $\mathcal{P}$ qui est l'ensemble des **individus** concernés par l'étude statistique. Un **échantillon** est un sous-ensemble de cette population. Les données statistiques sont des mesures effectuées sur les individus de l'échantillon.

On cherche à étudier un **caractère** ou **variable** de cette population à partir des observations faites sur l'échantillon. Une variable peut être **qualitative** (oui/non, homme/femme...) ou **quantitative** (à valeurs dans $\mathbb{R}$). On mesure cette variable sur l'échantillon et on obtient ainsi les observations ou données.

Exemple : si l'échantillon est un groupe de TD de l'université,
- un individu est un étudiant,
- la population peut être l'ensemble des étudiants de l'université, l'ensemble des étudiants en France, l'ensemble des habitants du Grand Lyon, etc.
- la variable étudiée peut être la mention obtenue au baccalauréat, la filière choisie, l'âge, la taille, etc.

Dans la suite, on distinguera trois cas suivant que la variable étudiée est une :
- **variable quantitative discrète** (elle ne prend qu'un nombre fini ou dénombrable de valeurs ; en général il s'agit d'entiers)
- **variable quantitative continue** (variable quantitative qui n'est pas discrète)
- **variable qualitative**.

C.1 Variable quantitative discrète

Le plus souvent, une variable quantitative discrète ne prend qu'un nombre fini et même petit (moins de 20) de valeurs. Par exemple, on peut s'intéresser au nombre d'enfants par famille, au nombre d'années d'étude des étudiants après le bac...

Exemple 78 *"The World Almanac and Book of Facts" (1975) a publié le nombre des grandes inventions mises au point chaque année entre 1860 et 1959, soit*

5 3 0 2 0 3 2 3 6 1 2 1 2 1 3 3 3 5 2 4 4 0 2 3 7 12 3 10 9 2 3 7 7
2 3 3 6 2 4 3 5 2 2 4 0 4 2 5 2 3 3 6 5 8 3 6 6 0 5 2 2 2 6 3 4 4
2 2 4 7 5 3 3 0 2 2 2 1 3 4 2 2 1 1 1 2 1 4 4 3 2 1 4 1 1 1 0 0 2 0

(source : base de données du logiciel R)

Notons $y = (y_1, ..., y_n)$ la suite des observations rangées par ordre croissant, n étant la taille de l'échantillon. Des données de ce type sont à présenter dans un tableau statistique dont la première colonne est l'ensemble des r observations distinctes (ou **modalités**), classées par ordre croissant et notées $x_1, ..., x_r$. Puis on leur fait correspondre dans une seconde colonne leurs **effectifs**, c'est-à-dire leurs nombres d'occurence, notés $n_1, ..., n_r$. Alors $\sum_i n_i = n$. On indique aussi les **fréquences** $f_i = n_i/n$ et les **fréquences cumulées** $F_i = \sum_{j=1}^{i} f_j$ $(1 \leq i \leq r)$. La plupart du temps, on donnera les fréquences sous forme de pourcentage. Pour notre exemple 78, on obtient

x_i	n_i	f_i	F_i
0	9	0.09	0.09
1	12	0.12	0.21
2	26	0.26	0.47
3	20	0.20	0.67
4	12	0.12	0.79
5	7	0.07	0.86
6	6	0.06	0.92
7	4	0.04	0.96
8	1	0.01	0.97
9	1	0.01	0.98
10	1	0.01	0.99
12	1	0.01	1

Venons-en maintenant aux représentations graphiques de base : le diagramme en bâtons et le diagramme cumulatif.

Le **diagramme en bâtons** se construit avec les modalités en abscisse et les effectifs en ordonnée. Quant au **diagramme cumulatif**, il s'obtient à partir des fréquences cumulées et c'est le graphe d'une fonction appelée **fonction de répartition empirique** et définie ainsi :

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_1 \\ F_i & \text{si } x_i \leq x < x_{i+1} \qquad (i = 1, ..., r-1) \\ 1 & \text{si } x \geq x_r \end{cases}$$

(*cf.* les graphes ci-dessous)

Un certain nombre de grandeurs, qui forment le résumé statistique de l'échantillon, participent aussi à l'analyse des données. On peut les classer en deux catégories : les

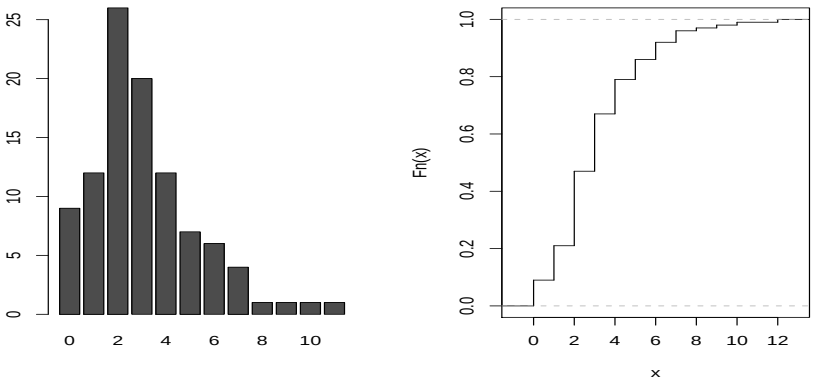


FIGURE C.1 – Diagramme en bâtons et diagramme cumulatif de l'exemple 78

paramètres de position et les paramètres de dispersion. Il existe bien sûr de nombreux autres paramètres qui permettraient une analyse plus fine. On peut se reporter à un ouvrage de statistique descriptive pour plus de détails.

Commençons par les **paramètres de position** que sont la moyenne et la médiane. Ces grandeurs donnent un "milieu", une position moyenne autour desquels les données sont réparties.

Définition 79 *La* **moyenne empirique** *est la moyenne arithmétique des observations :*

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{r} n_i x_i$$

La **médiane** *partage l'échantillon ordonné en deux parties de même effectif : la moitié au moins des observations lui sont inférieures ou égales et la moitié au moins lui sont supérieures ou égales. Quand les observations sont triées, si le nombre d'observations n est impair, la médiane est $y_{(n+1)/2}$. Si n est pair, n'importe quel nombre de l'intervalle $[y_{n/2}, y_{n/2+1}]$ vérifie la définition. On convient la plupart du temps de prendre le milieu de cet intervalle pour médiane. La médiane est aussi $F_n^{-1}(1/2)$.*

La moyenne de l'échantillon de l'exemple 1 est 3,1. Sa médiane est 3. Ici la médiane et la moyenne sont proches, mais ce n'est pas toujours le cas (ex : 0,100,101).

Comme **paramètres de dispersion**, on peut citer la variance, l'écart interquartile et l'amplitude. Ils servent à connaître la variabilité des données autour de la position moyenne.

Définition 80 *La* **variance empirique** *des observations est la moyenne du carré des écarts à la moyenne :*

$$s_n^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{r} n_i(x_i - \bar{y})^2$$

On utilise aussi le variance empirique modifiée :

$$s^2_{n-1}(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{r} n_i(x_i - \bar{y})^2$$

*L'***écart-type** *est la racine de la variance et on le note souvent s_n ou s_{n-1}, suivant qu'il est associé à la variance empirique ou à la variance empirique modifiée.*
Les **quartiles** *sont à rapprocher de la médiane : il divisent l'échantillon en quatre sous-ensembles de même effectif. Un quart (au moins) des observations sont inférieures ou égales au premier quartile et trois quarts (au moins) des observations lui sont supérieures. On remarque que le premier quartile est $F_n^{-1}(1/4)$. Le deuxième quartile est la médiane. Le troisième quartile est supérieur à trois quarts des observations et est inférieur à un quart. C'est aussi $F_n^{-1}(3/4)$. La différence entre le troisième quartile et le premier quartile est l'***écart interquartile***.*
*L'***amplitude** *(ou l'étendue) est la différence $y_n - y_1$.*

Pour les données de l'exemple 78, on obtient : $s^2_{n-1}(y) = 5,081$, $s_{n-1} = 2,254$, le premier quartile est 2, le troisième 4 et l'espace interquartile vaut donc 2. L'amplitude est 12.

Le **diagramme en boîte** (parfois appelé diagramme en boîte à moustaches ou **boxplot** ou box-and-whisker plot) concentre des informations liées aux quartiles, à la médiane et aux valeurs extrêmes. On définit les valeurs extrêmes comme s'écartant du quartile le plus proche d'au moins 1,5 fois l'espace interquartile. Il faut toujours ajuster les moustaches aux valeurs observées. Dans notre exemple, les valeurs extrêmes sont des valeurs strictement inférieures à -1 (ou plutôt 0) ou strictement supérieures à 7.
Attention : en pratique, pour tracer un boxplot, il faut
1) Calculer la médiane, le premier quartile $Q1$ et le troisième quartile $Q3$
2) Calculer la taille maximale des moustaches : $Q1 - 1.5 * (Q3 - Q1)$ et $Q3 + 1.5 * (Q3 - Q1)$
3) Raser les moustaches, c'est-à-dire les ajuster en les rétrécissant jusqu'à atteindre une valeur observée
4) Repérer les valeurs extrêmes, c'est-à-dire les valeurs qui sortent des moustaches.
Ne ratez pas une étape !

C.2 Variable quantitative continue

Quand une variable discrète prend un grand nombre de valeurs distinctes, le travail est légèrement différent.
Soit $y = (y_1, ..., y_n)$ un échantillon d'une variable continue. Il n'est pas pertinent de calculer les fréquences empiriques et les fréquences cumulées. Par contre, le résumé numérique et le diagramme cumulatif sont les mêmes que dans le cas discret. Ainsi, la moyenne est donnée par

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

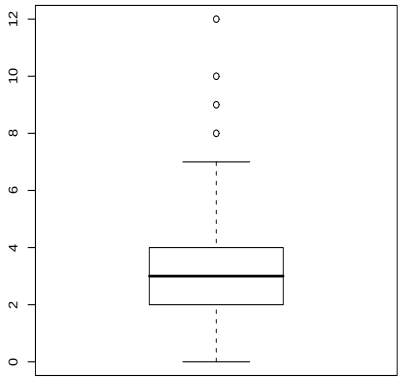


FIGURE C.2 – Boxplot de l'exemple 78

La variance empirique, la variance empirique modifiée et l'écart-type sont

$$s_n^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad s_n(y) = \sqrt{s_n^2(y)}$$

$$s_{n-1}^2(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad s_{n-1}(y) = \sqrt{s_{n-1}^2(y)}$$

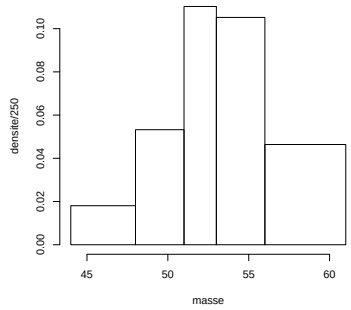
Mais à la place du diagramme en bâtons, on trace un **histogramme**. Pour ce faire, notons a et b respectivement la valeur observée minimale et la valeur observée maximale. On découpe l'intervalle $[a, b]$ en k petits intervalles appelés classes, avec k compris entre 6 et 12 (en gros $1 + \ln n / \ln 2$). Notons ces classes $[a_0, a_1[,...,[a_{k-1}, a_k]$, avec $a = a_0$ et $b = a_k$. On peut présenter les données dans un tableau, en indiquant : les classes rangées par ordre croissant, les effectifs n_i des classes (nombres d'observations dans les classes), les amplitudes L_i des classes ($L_i = a_i - a_{i-1}$), les **densités des observations** dans chaque classe, $h_i = \frac{n_i}{L_i}$.

Exemple 81 *On obtient, sur un échantillon de 250 œufs qu'on a pesé,*

Masse des œufs en gramme	*effectif*	*amplitude*	*densité*
$44 - 48$	19	4	$4,75$
$48 - 51$	42	3	14
$51 - 53$	58	2	29
$53 - 56$	83	3	$27,67$
$56 - 61$	48	5	$9,6$

L'histogramme est composé de rectangles dont
- les bases sont les classes
- les hauteurs sont les densités des classes (ou les hauteurs sont proportionnelles aux densités, ce qui donne le même graphique).

On peut choisir des classes de même largeur, ou des classes de même effectif !



C.3 Variable qualitative

Les modalités d'une variable qualitative ne sont pas numériques. De fait, on ne peut pas calculer les grandeurs statistiques telles que la moyenne... On peut faire un tableau pour présenter les données en indiquant pour chaque modalité l'effectif et la fréquence. Les trois représentations graphiques sont les diagrammes en colonne, les diagrammes en barre et le diagramme en secteur ("camembert").

Exemple 82 *On a interrogé des étudiants qui se rongent les ongles. Voici les circonstances pendant lesquelles ils pratiquent cette mauvaise habitude.*

activité	*fréquence*
regarder la télévision	*58*
lire un journal	*21*
téléphoner	*14*
conduire une auto	*7*
faire ses courses	*3*
autre	*12*

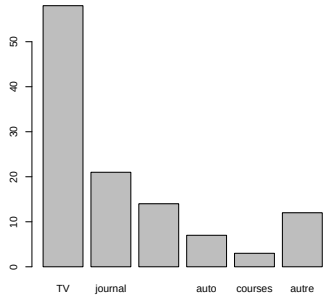


FIGURE C.3 – Diagramme des ongles rongés (ex 82)

Bibliographie

[1] Pierre Dagnélie. *Statistique théorique et appliquée.* De Boeck Université, 1998.

[2] Rick Durrett. *Elementary probability for applications.* Cambridge university press, 2009.

[3] Richard Arnold Johnson et Gouri K. Bhattacharyya. *Statistics : principles and methods.* Wiley, 1996.

[4] Aurelio Mattei. *Inférence et décision statistiques : théorie et application à la gestion des affaires.* P. Lang, 2000.

[5] Sheldon M. Ross. *Initiation aux probabilités.* Presses polytechniques et universitaires romandes, 2007.

[6] Gilbert Saporta. *Probabilités, analyse des données et statistique.* Technip, 1990.