Chapitre 1

Modélisation probabiliste

> La théorie des probabilités n'est au fond que le bon sens réduit au calcul.
>
> Pierre-Siméon Laplace

> Il est prouvé que fêter les anniversaires est bon pour la santé. Les statistiques montrent que les personnes qui en fêtent le plus deviennent les plus vieilles.
>
> Den Hartog

1.1 Introduction

Examinons la situation suivante. On veut étudier la manière dont un certain caractère est réparti parmi les **individus** d'une **population**, qui peut être de toute nature (êtres humains, animaux, plantes, microbes, territoires, périodes historiques, *etc*). Nous supposerons que le caractère en question est quantitatif (dont la mesure donne une valeur numérique) plutôt que qualitatif (beauté, intelligence, goût pour les maths...).

La méthode opératoire classique consiste, quand la population est trop grande, à en prélever un échantillon. On opère sur chaque individu une mesure. La liste des résultat s'appelle une **série statistique** et peut contenir de nombreuses valeurs. On veut comparer cette série avec les résultats obtenus sur d'autres échantillons. Comment faire ?

La *statistique descriptive* répond à cette question en fournissant des outils clés en main de deux types :
- D'une part des modes représentation graphiques (diagrammes en bâton, histogrammes, camembert) généralement connus et sur lesquels nous n'insisterons pas.
- D'autres part des **paramètres** tels que la moyenne, la variance, l'écart-type, la médiane, les quartiles *etc*, faciles à comparer d'une série statistique à une autre, et supposés condenser l'essentiel de l'information

Un problème différent, et plus épineux, est de fournir un modèle théorique permettant de *prévoir*, non pas le résultat d'une mesure, mais la probabilité que ce résultat se situe dans une certaine fourchette.

C'est l'objet de la *théorie des probabilités*. Quand une telle provision est possible, la grandeur mesurée s'appellera une **variable aléatoire**. On peut la voir comme une fonction, qui à chaque individu de la population générale associe le résultat *que donnerait la mesure si on l'effectuait* sur cet individu, autrement dit comme une fonction X de la population Ω dans l'ensemble $\mathbf{R}$ des réels (ou dans l'ensemble $\mathbf{N}$ des entiers naturels, si la variable aléatoire ne prend que des valeurs entières). On le note :

$$X : \Omega \longrightarrow \mathbf{R}$$

Exemple 1.1.1 La taille des hommes adultes en France cette année est une variable aléatoire : on ne peut prédire à l'avance quelle est la taille d'une personne prise au hasard, mais on peut estimer de façon précise la probabilité que cette taille soit comprise par exemple entre 1m62 et 1m87.

Pour cela la théorie des probabilités dispose de tout un arsenal théorique que nous ne pourrons pas expliquer ici. Nous en retiendrons simplement le mode opératoire : étant donnée une variable aléatoire X, on appelle **loi** de X une application P_X qui donne, pour tout intervalle I de $\mathbf{R}$ la probabilité que X se situe dans l'intervalle I. Cette probabilité $P_X(I)$ est notée :

$$P(X \in I)$$

Exemple 1.1.2 On ne peut savoir à l'avance quel est le résultat d'un lancer de dés à 6 faces. Mais si le dés est bien équilibré on sait qu'il y a une chance sur 6 d'obtenir un 1, un 2, un 3, *etc.* On sait aussi qu'il y a une chance sur deux d'obtenir un nombre pair. Le résultat d'un lancer de dés est donc une variable aléatoire, dont la loi est donnée par :

$$P(X \in I) = \frac{\sharp\,(I \cap \{1, 2, \ldots, 6\})}{6}$$

où $\sharp\, E$ désigne le **cardinal** de l'ensemble E, c'est-à-dire le nombre d'éléments de E. Une telle loi porte un nom : c'est une loi uniforme discrète sur l'ensemble $\{1, 2, \ldots, 6\}$.

Dans ce cours, nous donnerons sans les justifier un certains nombre de critères permettant de déterminer facilement quelle loi classique suit une variable aléatoire. Nous verrons ensuite comment les utiliser pour faire des prédictions extrêmement poussées.

> **Avertissement :** *Il est important de souligner que toutes ces prédictions ou estimations, aussi précises soient-elles, reposent sur le choix d'une loi de probabilité pour décrire un phénomène donnée, ce qu'on appelle un* **modèle probabiliste**. *Quelque soit le soin apporté dans le choix de ce modèle, celui-ci ne constitue pas « la réalité ».*

1.2 Statistique descriptive sur une population finie

Un série statistique peut être donnée sous diverses formes, dont les plus courantes sont :

– La liste des valeurs observées, avec répétition (si une valeur est observée 3 fois, elle apparaît 3 fois dans la liste). Exemple :

11 11 21 21 21 35 48 48 92

– La liste des couples (valeur x observée, nombre n de fois où x a été observée). Exemple :

$$(11;2)\ (21;3)\ (35;1)\ (48;2)\ (92;1)$$

Dans une telle liste $(x_1, n_1), \ldots, (x_r, n_r)$ les x_i sont appelés les **modalités** et les n_i leurs **effectifs**. On suppose toujours les x_i rangés par ordre croissant :

$$x_1 < x_2 < \cdots < x_r$$

L'**effectif total** est $N = n_1 + \cdots + n_r$, ce qu'on note $N = \sum_{1 \leq i \leq r} n_i$ ou encore $N = \sum_{i=1}^{n_r} n_i$.

La **fréquence** d'une modalité x_i est $f_i = \dfrac{n_i}{N}$.

L'**effectif cumulé** d'une modalité x_i est $N_i = \sum_{1 \leq j \leq i} n_j$.

La **fréquence cumulée** de x_i est $F_i = \sum_{1 \leq j \leq i} f_j = \dfrac{N_i}{N}$.

Pour *décrire* notre série statistique (ce qui peut permettre de la comparer plus facilement à une autre, par exemple en comparant deux moyennes) on peut en calculer différents paramètres dont les principaux sont :

– Sa **moyenne** : $\overline{x} = \sum_{i=1}^{r} f_i x_i = \dfrac{1}{N} \sum_{i=1}^{r} n_i x_i$

– Sa **variance** : $V = \sum_{i=1}^{r} f_i (x_i - \overline{x})^2 = \dfrac{1}{N} \sum_{i=1}^{r} n_i (x_i - \overline{x})^2$

– Son **écart-type** : $\sigma = \sqrt{V}$

La moyenne est un *paramètre de position*. Elle indique autour de quelle valeur théorique sont centrées les valeurs observées.

La variance (comme l'écart-type) est un *paramètre de dispersion*. Plus les modalités ayant un effectif important seront resserrées autour de la moyenne et plus la variance sera petite.

D'autres paramètres de position (médiane, quartiles, *etc*) ou de dispersion (étendue, écart-moyen, écart inter-quartiles, *etc*) peuvent venir affiner cette description sommaire. Commençons par les plus simples :

– L'**étendue** est la différence entre la plus grande et la plus petite modalité.

– L'**écart moyen** est la moyenne des écarts (en valeurs absolue) entre les modalités et la moyenne $\overline{x}$:

$$\frac{1}{N} \sum_{i=1}^{r} n_i |x_i - \overline{x}|$$

Les autres paramètres descriptifs que nous rencontrerons sont tous basés sur la notion de q-quantile, où $q \geq 2$ est un entier.

Techniquement, pour définir et calculer **le k-ème q-quantile** $Q_{k/q}$ (où k est un entier entre 1 et $q-1$), on commence par ranger les modalités par ordre croissant et par déterminer la plus grande modalité x_i dont la fréquence cumulée $F_i \leq k/q$. Alors $Q_{k/q}$ est, par définition, l'abcisse de l'unique point d'ordonnée k/q situé sur le segment joignant les points de coordonnées (x_i, F_i) et (x_{i+1}, F_{i+1}).

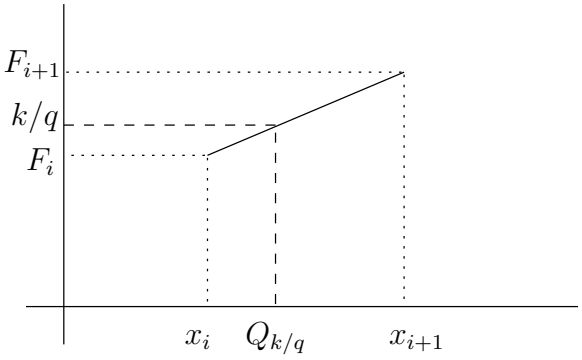


FIG. 1.1 – Représentation graphique du quantile $Q_{k/q}$.

On le calcule comme suit. Le théorème de Thales nous dit que les rapports des ordonnées et des abcisses sont égaux. Notons θ ce rapport :

$$\frac{k/q - F_i}{F_{i+1} - F_i} = \theta = \frac{Q_{k/q} - x_i}{x_{i+1} - x_i}$$

La première égalité permet de calculer θ à partir des fréquences cumulées F_i et F_{i+1}. De la seconde égalité on déduit alors facilement que $Q_{k/q} = x_i + \theta(x_{i+1} - x_i)$.

Un point important est qu'une telle définition ne fonctionne que si $k/q < F_1 = f_1$, la fréquence de la plus petite modalité. *Le k-ème q-quantile n'est donc défini que pour $k/q \geq F_1$.* D'autre part il résulte facilement de la définition que :

$$Q_{1/q} < Q_{2/q} < \cdots < Q_{(q-1)/q}$$

Certains q-quantiles portent des noms spécifiques. L'unique 2-quantile est appelé la **médiane**. Les trois 4-quantiles sont appelés **quartiles**. Notons que le deuxième quartile $Q_{2/4}$ est égal à la médiane $Q_{1/2}$ puisque $2/4 = 1/2 = 0,5$. On devine sans peine ce que sont les **déciles** ($q = 10$) et les **centiles** ($q = 100$). Ces derniers sont souvent appelés **percentiles** par anglomanie.

Tous les q-quantiles sont des paramètres de position, mais l'**écart inter-quartile** défini par $Q_{3/4} - Q_{1/4}$ est quant à lui un paramètre de dispersion.

• **Représentation graphique.** Pour comprendre ce que représentent vraiment les q-quantiles, il est commode d'utiliser le **polygone des fréquences cumulées croissantes**. Celui-ci est une représentation graphique des fréquences cumulées, qui consiste à placer dans un repère adapté les points de coordonnées (x_i, F_i) et à tracer une ligne polygonale joignant ces points (dans l'ordre des x_i croissants).

Exemple 1.2.1 Voici la liste des notes obtenues en L1-SVG par un étudiant de l'an dernier, dans l'ordre croissant :

$$7, 7, 7, 8, 8, 9, 9, 9, 12, 13, 13, 18$$

Son effectif total est 12, sa moyenne 10, son écart-type environ $3,24$. On calcule :

Modalités	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$x_1 = 7$	3	3	0,25	0,25
$x_2 = 8$	2	5	0,17	0,42
$x_3 = 9$	3	8	0,25	0,67
$x_4 = 12$	1	9	0,08	0,75
$x_5 = 13$	2	11	0,17	0,92
$x_6 = 18$	1	12	0,08	1,00

On trace alors :

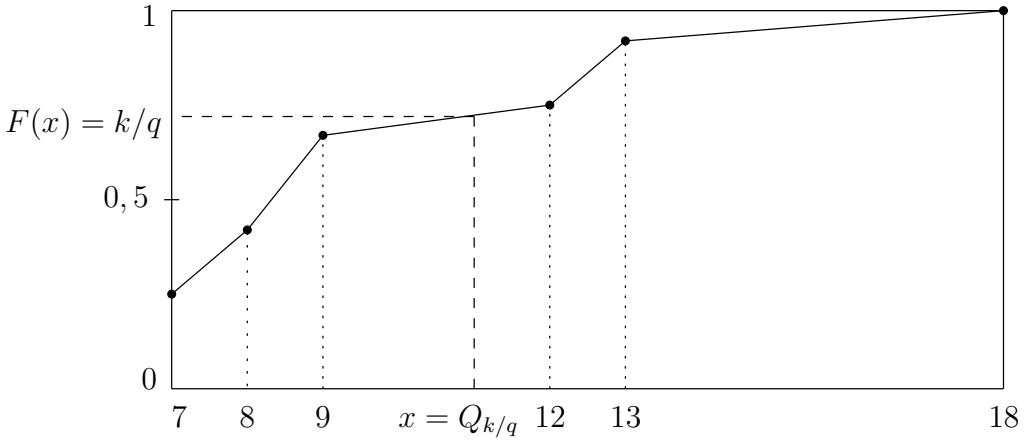


FIG. 1.2 – Polygone des fréquences.

Le polygone des fréquences rend possible de définir pour tout réel x compris entre la plus petite et la plus grande modalité (ici entre 7 et 18) ce qu'on pourrait appeler une « fréquence cumulée théorique[1] » $F(x)$ ayant pour valeur l'ordonnée de l'unique point d'abcisse x sur le polygone des fréquences (voir figure 1.2).

Le polygone des fréquences n'est rien d'autre que le graphe de la fonction F, laquelle est donc continue, croissante, à valeurs dans $[0, 1]$. Elle est définie uniquement sur le plus petit intervalle contenant toutes les modalités (ici l'intervalle $[7, 18]$).

Cette définition, parfaitement générale, permet de redéfinir le k-ème q-quantile comme *l'unique réel x tel que $F(x) = k/q$* (celui-ci n'est bien sûr défini que pour k/q entre F_1 et 1). Notons que comme F est croissante on a aussi[2] :

$$F(x) \leq k/q \iff x \leq Q_{k/q}$$

Ainsi la médiane se lit-elle sur la figure 1.2, comme l'abcisse de l'unique point d'ordonnée $0,5$ sur le polygone des fréquences. Dans notre exemple on voit qu'elle se situe entre $x_2 = 8$ et $x_3 = 9$. Pour la

[1] Cette terminologie est particulière à ce cours. Toutefois cette fonction F est très proche d'une notion standard que nous rencontrerons plus loin sous le nom de « fonction de répartition ». Les deux notions coïncident même parfaitement pour les variables aléatoires continues.

[2] Autrement dit $Q_{k/q}$ est le plus grand réel x tel que $F(x) \leq k/q$. C'est aussi, par un argument similaire, le plus petit réel x tel que $F(x) \geq k/q$.

déterminer on commence par calculer le rapport $\theta = (0,5 - 0,42)/(0,67 - 0,42) = 0,3478$. La médiane de cette série statistique est donc :

$$Q_{1/2} = x_2 + 0,3478(x_3 - x_2) \simeq 8,35$$

Ici les deux autres quartiles se lisent directement sur le tableau des fréquences cumulées[3] puisque $F_1 = 0,25$ et $F_4 = 0,75$. On a donc respectivement $Q_{1/4} = x_1 = 7$ et $Q_{3/4} = x_4 = 12$, d'où un écart interquartile de $12 - 7 = 5$.

On note que les deux premiers déciles $q_{1/10}$ et $q_{2/10}$ ne sont pas définis pour cette série statistique puisque $2/10 < 0,25 = F_1$.

Remarque 1.2.2 Certains auteurs donnent une autre définition, non équivalente, des q-quantiles : le k-ème q-quantile est pour eux la première modalité x_i telle que la fréquence cumulée $F_i \geq k/q$. L'avantage de cette notion alternative est de ne nécessiter aucune interpolation par le polygone des fréquences et d'être définie pour tous les k entre 1 et q, pour tout q. En outre elle ne fournit que des valeurs effectivement observées. La différence est minime en pratique, dès que l'on dispose d'une série statistique ayant des modalités assez finement distribuées. Quand au contraire les deux notions donnent des valeurs très différentes, alors les quantiles interpolés (ceux que nous avons définis plus haut) sont *plus significatifs* comme paramètres de position. C'est pourquoi ce sont ceux-là seuls que nous retiendrons.

• **Regroupement en classes.** Lorsqu'une série statistique possède un trop grand nombre de modalités, il peut être commode de les regrouper en classes. Les **classes** sont des intervalles disjoints $I_1, \ldots, I_s$ recouvrant l'ensemble des modalités. On construit alors une nouvelle série statistique dont les modalités y_j (pour j entre 1 et s) sont les *moyennes* des modalités $x_i \in I_j$, et dont les effectifs sont les sommes des effectifs des $x_i \in I_j$.

Notons que le regroupement en classes fait perdre de l'information : on ne sait plus comment étaient répartis les x_i dans chaque classe I_j. Il modifie aussi les paramètres de position et de dispersion, à l'exception notable de la moyenne.

Exemple 1.2.3 On reprend la série statistique de l'exemple précédent, et on opère un regroupement en classes $[0, 8[$, $[8, 10[$, $[10, 20]$. Le résultat est une série y_1, y_2, y_3 dont les effectifs sont présentés dans le tableau ci-après :

Modalités x_i	Effectifs	Modalités y_j	Effectifs
$x_1 = 7$	3	$y_1 = 7$	3
$x_2 = 8$	2		
$x_3 = 9$	3	$y_2 = 8,6$	5
$x_3 = 12$	1		
$x_4 = 13$	2		
$x_5 = 18$	1	$y_3 = 14$	4

[3] Il s'agit bien sûr d'une particularité de notre exemple. Dans le cas général, il faudrait calculer $Q_{1/4}$ et $Q_{3/4}$ par interpolation, comme on l'a fait pour la médiane $Q_{1/2}$.

Cette nouvelle série statistique a la même moyenne que la série initiale mais un écart-type nettement inférieur (environ $2,90$ pour les y_j au lieu de $3,24$ pour les x_i).

1.3 Probabilités et variables aléatoires

Nous distinguerons deux grands types de variables aléatoires $X : \Omega \to \mathbf{R}$:

- Si les valeurs prises par X sont isolées les unes des autres (typiquement parce que X ne prend que des valeurs entières, ou qu'un nombre fini de valeurs) on parle de **variable aléatoire discrète**.
- Si au contraire les modalités de X forment un intervalle (non réduit à un singleton !) nous dirons que X est une **variable aléatoire continue**.

Dans tous les cas, on a les propriétés suivantes.

Proposition 1.3.1 *Soient $X : \Omega \to \mathbf{R}$ une variable aléatoire et $I, J.$ deux intervalles quelconques, ou plus généralement deux parties de $\mathbf{R}$ qui sont réunion d'un nombre fini d'intervalles. Alors :*

$$P(X \in I \text{ ou } X \in J) = P(X \in I) + P(Y \in J) - P(X \in I \text{ et } Y \in J)$$

Comme évidemment la probabilités que $X \in \emptyset$ est nulle, et la probabilité que $X \in \mathbf{R}$ est 1, on a aussi :
- *Si $I \cap J = \emptyset$ alors $P(X \in I \cup J) = P(X \in I) + P(X \in J)$.*
- *Si $J \subseteq I$ alors $P(X \in I \setminus J) = P(X \in I) - P(X \in J)$.*
- *En particulier, $P(X \notin I) = 1 - P(X \in I)$.*

1.3.2 Variables aléatoires discrètes

Pour décrire la loi d'une variable aléatoire discrète $X : \Omega \to \mathbf{R}$, il suffit de se donner une énumération $(x_k)_{k\in K}$ des modalités de X (indexée sur $K \subseteq \mathbf{N}$) et une suite $(p_k)_{k\in K}$ de réels positifs ou nuls telle que [4] $\sum_{k\in K} p_k = 1$. On pose alors pour tout intervalle I de $\mathbf{R}$:

$$P(X \in I) = \sum_{\substack{k\in K \\ x_k \in I}} p_k$$

En particulier, on aura $P(X = x_k) = p_k$ pour tout $k \in K$.

En pratique, nos variables aléatoires discrètes seront le plus souvent à valeurs dans $\mathbf{N}$, et on se donnera simplement la suite $(p_k)_{k\in\mathbf{N}}$. Plutôt que la formule ci-dessus on écrira alors :

$$P(X \in I) = \sum_{k\in I\cap\mathbf{N}} p_k$$

Et on aura donc $P(X = k) = p_k$ pour tout $k \in \mathbf{N}$.

[4] Si l'index de sommation K est infini, par exemple si $K = \mathbf{N}$, alors une expression comme $\sum_{k\in\mathbf{N}} p_k$ n'a pas de sens *stricto sensu*, puisqu'on ne peut calculer que des sommes portant sur un nombre fini de termes. Toutefois, comme on a supposé les $p_k \geq 0$ il s'ensuit que la suite $s_n = \sum_{k\leq n} p_k$ est croissante ($s_n - s_{n-1} = p_n \geq 0$). Elle tends donc nécessairement soit vers une limite finie, soit vers plus l'infini (elle ne peut pas ne pas avoir de limite). C'est cette limite finie ou non que l'on note abusivement $\sum_{k\in\mathbf{N}} p_k$.

Exemple 1.3.3 Reprenons la fonction $X : \Omega \to \mathbf{R}$ qui modélise le résultat d'un lancer de dés. Si on veut calculer par exemple la probabilité de l'événement "X est pair", autrement dit "$X \in \{2,4,6\}$" on écrira :

$$P(X \in \{2,4,6\}) = P(X=2) + P(X=4) + P(X=6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

1.3.4 Variable aléatoire continue

Pour décrire la loi d'une variable aléatoire continue, nous nous restreindrons au cas particulier ou celles-ci peut s'exprimer sous la forme d'une intégrale. Autrement dit nous nous donnerons une fonction $f_X : \mathbf{R} \to \mathbf{R}$ continue par morceaux, positive ou nulle, telle que[5] $\int_{-\infty}^{+\infty} f_X(t)\mathrm{d}t = 1$ et nous poserons par définition pour tout intervalle I de $\mathbf{R}$ d'extrémités $a < b \in \mathbf{R} \cup \{\pm\infty\}$:

$$P(X \in I) = \int_a^b f_X(t)\mathrm{d}t$$

Une telle fonction f_X est appelée une **densité de probabilité** pour X. Soulignons deux conséquences immédiates mais importantes de cette définition, qui distinguent bien le cas continu du cas discret :
- $P(X = a) = \int_a^a f_X(t)\mathrm{d}t = 0$ pour tout $a \in \mathbf{R}$.
- $P(X \leq a) = P(X < a)$ pour tout $a \in \mathbf{R}$.

1.3.5 Fonction de répartition

On appelle **fonction de répartition** d'une variable aléatoire $X : \Omega \to \mathbf{R}$ (qu'elle soit discrète ou continue) la fonction $F_X : \mathbf{R} \to \mathbf{R}$ définie par :

$$F_X(x) = P(X \leq x)$$

Cette fonction est nécessairement croissante, tend vers 0 en $-\infty$ et vers 1 en $+\infty$, ce qui autorise à écrire par abus de notation :

$$F_X(-\infty) = 0 \qquad \text{et} \qquad F_X(+\infty) = 1$$

- Si X prend un nombre fini de valeurs $x_1 < x_2 < \cdots < x_N$, alors $F_X(x_i)$ est exactement la fréquence cumulée de x_i.
- Si X est à valeurs dans $\mathbf{N}$, et $P(X = k) = p_k$ pour tout entier k, alors on aura pour tout $n \in \mathbf{N}$:

$$F_X(n) = \sum_{0 \leq k \leq n} p_k$$

- Si au contraire X est une variable aléatoire continue, et f_X une densité de probabilité de X alors pour tout $x \in \mathbf{R}$:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$$

  Une telle fonction est toujours continue (d'où le nom de variable aléatoire continue). Si de plus la densité de probabilité f_X est continue (ce qui sera toujours le cas pour nous en pratique) alors F_X est dérivable sur $\mathbf{R}$, de dérivée f_X.

[5] L'intégrale $\int_a^b f_X(t)\mathrm{d}t$ n'est bien définie que pour $a, b \in \mathbf{R}$, mais l'hypothèse que $f_X(t) \geq 0$ pour tout $t \in \mathbf{R}$ permet de définir $\int_{-\infty}^{+\infty} f_X(t)\mathrm{d}t$ à l'aide d'un passage aux limites que nous ne détaillerons pas.

Dans tous les cas, le principal intérêt de la fonction de répartition est qu'elle détermine entièrement la loi de X. Pour s'en convaincre, remarquons tout d'abord qu'on peut calculer la probabilité de tout évènement du type "$X \in]a, b]$" grâce à la formule suivante.

Proposition 1.3.6 *Pour tous $a < b \in \mathbf{R} \cup \{\pm\infty\}$:*

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

Plus généralement, la fonction de répartition permet de calculer la probabilité de tout évènement "$X \in I$" où I est un intervalle quelconque (voir TD). En particulier si deux variables aléatoires X et Y ont la même fonction de répartition alors $P(X \in I) = P(Y \in I)$ pour tout intervalle I, et il est facile de voir qu'il en est de même pour toute réunion d'intervalles, donc que X et Y ont la même loi. Réciproquement, si X et Y ont la même loi alors il est immédiat qu'elles ont la même fonction de répartition (il suffit de regarder la définition de F_X). En résumé :

Proposition 1.3.7 *Deux variables aléatoires quelconques ont le même loi si et seulement si elles ont la même fonction de répartition.*

Si les fonctions de répartitions de X et Y sont seulement approximativement égales (en pratique à moins de 1 % près), alors nous dirons souvent que Y « modélise » X. Noter que X et Y ne sont pas nécessairement définies sur la même population. Tout l'intérêt d'une telle modélisation est de remplacer une variable aléatoire X dont la fonction de répartition est mal connue ou difficile à calculer par une variable aléatoire Y dont la fonction de répartition, connue *a priori*, est approximativement égale à celle de X et se calcule facilement à l'aide de tables spécialement dédiées. En particulier il est fréquent de modéliser une variable aléatoire discrète par une variable aléatoire continue, comme nous le verrons plus loin.

1.3.8 Espérance et variance d'une variable aléatoire

En statistiques descriptives, nous n'avons considéré que des populations Ω finies. Dans ce cas toute fonction $X : \Omega \to \mathbf{R}$ prend un nombre fini de modalités, que l'on peut ranger en ordre croissant $x_1 < \cdots < x_r$. Pour faire de X une variable aléatoire, il faut se donner un loi de probabilités, et dans un tel contexte il est naturel de poser :

$$P(X = x_i) = \frac{\sharp \text{“}X = x_i\text{”}}{\sharp \Omega}$$

On peut remarquer que l'effectif n_i de toute modalité x_i est exactement le cardinal de l'évènement "$X = x_i$", et on a donc :

$$P(X = x_i) = \frac{n_i}{\sharp \Omega} = f_i$$

Pour tout intervalle I de $\mathbf{R}$ on aura donc :

$$P(X \in I) = \frac{\sharp \text{“}X \in I\text{”}}{\sharp \Omega}$$

En particulier la fréquence cumulée F_i est exactement $P(X \leq x_i)$, autrement dit $F_i = F_X(x_i)$.

On en déduit que la moyenne $\overline{x}$ et la variance V de la série statistique des (x_i, n_i) peuvent s'exprimer directement en fonction de X par les formules suivantes :

– $\overline{x} = \sum_{i=1}^{r} x_i P(X = x_i)$.

– $V = \sum_{i=1}^{r} (x_i - \overline{x})^2 P(X = x_i)$.

Si de plus X est à valeurs dans $\mathbf{N}$, ces deux sommes peuvent s'écrire sous la forme :

– $\overline{x} = \sum_{k \in \mathbf{N}} k P(X = k)$.

– $V = \sum_{k \in \mathbf{N}} (k - \overline{x})^2 P(X = k)$..

En effet dans ces sommes les seuls termes qui comptent sont ceux pour lesquels $P(X = k) \neq 0$, autrement dit ceux pour lesquels k est l'une des modalités de X.

Ces observations ouvrent la voie à une généralisation aux variables aléatoires discrètes ou continues des paramètres fondamentaux des statistiques descriptives. Pour des raisons historiques, la notion statistique de « moyenne » est rebaptisée « espérance » en théorie des probabilités.

Si $X : \Omega \to \mathbf{R}$ est une variable aléatoire discrète à valeurs dans $\mathbf{N}$, on appelle :

– **Espérance** de X le nombre $E(X) = \sum_{k \in \mathbf{N}} k p_k$.

– **Variance** de X le nombre $V(X) = \sum_{k \in \mathbf{N}} (k - E(X))^2 p_k$.

L'**écart-type** de X sera naturellement $\sigma(X) = \sqrt{V(X)}$.

Pour généraliser ces notions au cas des variables aléatoires continues, nous allons nous baser sur une analogie. Nous avons vu que la fonction de répartition de X s'écrit comme une intégrale sur f_X quand X est continue, et comme une somme sur les $p_k = P(X = k)$ si X est discrète à valeurs dans $\mathbf{N}$. Autrement dit l'intégration de la fonction f_X joue dans le cas continu le même rôle que la sommation de la suite $(p_k)_{k \in \mathbf{N}}$. C'est cette analogie qui justifie les définitions suivantes.

Si $X : \Omega \to \mathbf{R}$ est une variable aléatoire continue etf_X une densité de probabilité pour X, on appelle :

– **Espérance** de X le nombre $E(X) = \int_{-\infty}^{+\infty} t f_X(t) \mathrm{d}t$.

– **Variance** de X le nombre $V(X) = \int_{-\infty}^{+\infty} (t - E(X))^2 f_X(t) \mathrm{d}t$.

Là encore l'**écart-type** de X est $\sigma(X) = \sqrt{V(X)}$.

Tous ces paramètres sont des réels positifs ou nuls, ou $+\infty$. Nous admettrons qu'ils ont les propriétés suivantes.

Proposition 1.3.9 *Soient $X : \Omega \to \mathbf{R}$, $Y : \Omega \to \mathbf{R}$ deux variables aléatoires et $a, \lambda \in \mathbf{R}$.*

$$\begin{array}{ll} E(X + a) = E(X) + a & V(X + a) = V(X) \\ E(\lambda X) = \lambda E(X) & V(\lambda X) = \lambda^2 V(X) \\ E(X + Y) = E(X) + E(Y) & \end{array}$$

Proposition 1.3.10 *Soit* $X : \Omega \to \mathbf{R}$.

$$V(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$$

1.4 Lois classiques

1.4.1 Loi uniforme discrète

On dit qu'une variable aléatoire X **suit une loi uniforme discrète** si :
– L'ensemble $\{x_1, \ldots, x_n\}$ de ses modalités est fini.
– Pour tout k, $P(X = x_k) = \dfrac{1}{n}$.
Autrement dit, X suit une loi uniforme si toutes ses modalités sont observées avec la même probabilité. L'espérance et la variance d'une telle variable aléatoire sont exactement la moyenne et la variance de la série statistique des $(x_i, 1)$.

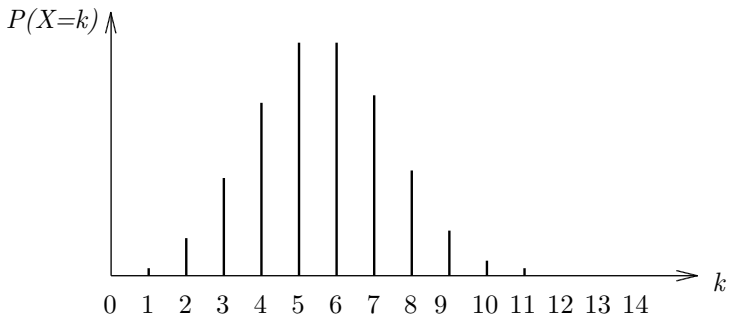
Les exemples typiques sont ceux d'un lancer de dés (6 modalités équiprobables), d'un jeu de pile ou face (2 modalités équiprobables), d'un tirage de loto (pour 7 chiffres entre 1 et 49, C_{49}^7 modalités équiprobables), *etc.*

Rappelons en passant que pour tout entier n, k avec $0 \leq k \leq n$, on note C_n^k le nombre de parties à k éléments d'un ensemble à n éléments. Il se calcule ainsi :

$$\mathrm{C}_n^k = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$$

1.4.2 Loi binomiale $\mathcal{B}(n, p)$

On dit qu'une variable aléatoire X **suit une loi binomiale de paramètres** n **et** p si, en posant $q = 1 - p$:
– X prend ses valeurs parmi $\{0, 1, 2, \ldots, n\}$.
– Pour tout k, $P(X = k) = \mathrm{C}_n^k p^k q^{n-k}$.
On l'écrit en abrégé « X suit une loi $\mathcal{B}(n, p)$ » ou encore $X \sim \mathcal{B}(n, p)$.



On peut calculer :

$$E(X) = np \quad \text{et} \quad V(X) = npq$$

On rencontre cette loi dans la situation suivante :

– Une expérience consiste à répéter n fois un même test.
– Chacun de ces tests ne peut fournir que deux résultats (positif ou négatif) avec probabilité p et $q = 1-p$ respectivement.
– Ces tests sont indépendants les uns des autres [6].

On introduit alors la variable aléatoire X qui à tout n-uplet de tests ω associe le nombre de tests positifs parmis ceux de ω.

Théorème 1.4.3 *Dans les conditions ci-dessus, la variable aléatoire X suit une loi $\mathcal{B}(n,p)$.*

Exemple 1.4.4 On plante dans un champs 2000 graines d'une certaine espèce, dont on sait par ailleurs que chacune a une probabilité 0,85 de germer. Quelques semaines plus tard, on compte le nombre de graines ayant germé. La variable aléatoire associée à ce décompte est donc la fonction G qui à toute plantation ω de 2000 graines de cette espèce associe le nombre de graines ayant germé. Les « individus » ω sont ici les plantations, et la « population » Ω l'ensemble des plantations possibles.

Pour une plantation ω donnée, le décompte de $G(\omega)$ consiste à répéter sur chacune des $n = 2000$ graines semées dans un champs ω le test : « est-ce que la graine a germé ? ». Chacun de ces tests a une probabilité 0,85 de fournir un résultat positif, et les tests sont indépendants. La variable aléatoire G suit donc une loi $\mathcal{B}(2000; 0,85)$.

Il s'ensuit que le nombre moyen de graines germées dans une telle plantation sera de :

$$E(G) = 2000 \times 0,85 = 1700$$

avec un écart-type de : $\sigma(G) = \sqrt{2000 \times 0,85 \times (1-0,85)} = 15,9687\cdots \approx 16$.

Soit une fourchette allant de 1684 à 1716 graines germées. Noter que le rendement moyen, $E(G)/2000 = 0,85$ correspond bien à ce qui était intuitivement prévisible.

On peut maintenant se demander quelle est la probabilité d'avoir un rendement compris dans cette fourchette. Pour cela il « suffit » de calculer :

$$P(1684 \leq G \leq 1716) = \sum_{k=1684}^{1716} \mathrm{C}_{2000}^{k} \times 0,85^{k} \times 0,15^{2000-k}$$

On touche ici aux limites de ce modèle, car calculer une telle somme nécessite l'usage d'un ordinateur et d'un bon logiciel de calcul. Il est néanmoins possible d'obtenir presque sans calcul une valeur approchée du résultat, grâce à l'approximation de la loi binomiale par la loi normale, que nous verrons plus loin.

1.4.5 Loi hypergéométrique

On dit qu'une variable aléatoire X **suit une loi hypergéométrique de paramètres** n, K, N avec n et K inférieurs ou égaux à N si :
– X prend ses valeurs parmi $\{0, 1, 2, \ldots, l\}$ où $l = \min(n, K)$.

[6] Nous ne donnons pas ici de définition précise de ce qu'on entend par « tests indépendants ». L'idée est que si le résultats des premiers tests n'a aucune influence sur les suivants, alors les tests sont indépendants (voir à ce sujet la remarque 1.4.7).

– Pour tout k, $P(X=k) = \dfrac{\mathrm{C}_K^k \mathrm{C}_{N-K}^{n-k}}{\mathrm{C}_N^n}$.

En posant $p = K/N$ et $q = 1-p$, on peut calculer (même si nous ne l'utiliserons guère) :

$$E(X) = np \qquad V(X) = npq\frac{1-n/N}{1-1/N}$$

Cette loi se rencontre dans une situation très voisine de celle de la loi binomiale :
– L'expérience consiste à répéter un même test sur n éléments d'un ensemble qui en compte N.
– Chacun de ces tests ne peut fournir que deux résultats (positif ou négatif).
– On connaît à l'avance le nombre K d'éléments dans cet ensemble pour lesquels le test est positif.
On considère alors la variable aléatoire X qui à tout n-uplet ω d'éléments associe le nombre de résultats positifs parmi les tests effectués sur les éléments de ω.

Théorème 1.4.6 *Dans les conditions ci-dessus, la variable aléatoire X suit une loi hypergéométrique de paramètres n, K, N.*

Remarque 1.4.7 Comme pour la loi binomiale, chaque individu pris isolément dans Ω a une probabilité K/N de donner un résultat positif. Mais ici les tests ne sont pas indépendants ! En effet, supposons que dans un échantillon de 2 individus ω_1 et ω_2, le test a donné un résultat positif sur ω_1. Comme ω_2 appartient à la population $\Omega' = \Omega \setminus \{\omega_1\}$, qui comporte $N-1$ individus dont $K-1$ donnent un résultat positif, on peut en déduire que la probabilité pour que le test soit positif sur ω_2 est $(K-1)/(N-1)$ et non plus K/n. On voit ici comment le résultat du premier test peut avoir une incidence sur les suivants.

Toutefois on peut montrer que dès que N est « grand » devant n (ce qui correspond à la situation expérimentale d'une vaste population dont on extrait un échantillon raisonnable de n individus), la loi $\mathcal{B}(n,p)$ avec $p = K/N$ fournit une bonne approximation de la loi hypergéométrique de paramètres n, K, N. C'est pourquoi nous utiliserons peu cette loi, la remplaçant par la loi binomiale chaque fois que ce sera possible.

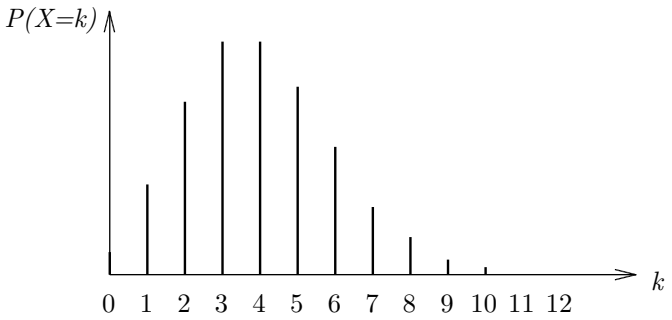
Théorème 1.4.8 *Soit X une variable aléatoire suivant une loi hypergéométrique de paramètres n, K, N. Si $n < N/10$ alors pour tout k :*

$$P(X=k) \approx \mathrm{C}_n^k p^k q^{1-k}$$

où $p = K/N$ et $q = 1-p$.

1.4.9 Loi de Poisson $\mathcal{P}(\lambda)$

On dit qu'une variable aléatoire X **suit une loi de Poisson de paramètre** λ si c'est une variable aléatoire discrète, à valeurs dans $\mathbf{N}$, et si :
– Pour tout k, $P(X=k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$.
On l'écrit en abrégé « X suit une loi $\mathcal{P}(\lambda)$ » ou encore $X \sim \mathcal{P}(\lambda)$.



Loi de Poisson $\mathcal{P}(4)$

On calcule alors :

$$E(X) = \lambda \qquad V(X) = \lambda$$

Cette loi non expérimentale (puisqu'elle autorise une infinité de modalités) fournit néanmoins une bonne approximation de la loi binomiale $\mathcal{B}(n,p)$ de même espérance, dans les conditions suivantes :

Théorème 1.4.10 *Soit X une variable aléatoire suivant une loi $\mathcal{B}(n,p)$. Si $n \geq 30$, $p \leq 0,1$ et $np \leq 10$ alors pour tout k :*

$$P(X = k) \approx e^{-\lambda}\frac{\lambda^k}{k!}$$

où $\lambda = np$.

À cause de ce théorème, la loi de Poisson est souvent utilisée pour modéliser les variables aléatoires mesurant le nombre d'observations d'un évènement *rare* (quand n est grand et p est petit).

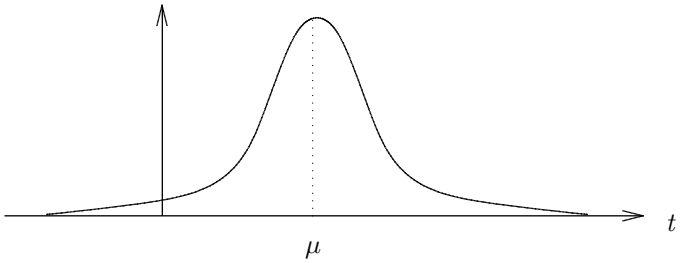
1.4.11 Loi normale (ou Gaussienne) $\mathcal{N}(\mu,\sigma)$

On dit qu'une variable aléatoire X **suit une loi normale (ou Gaussienne)** de paramètres μ et σ si :

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

On l'écrit en abrégé « X suit une loi $\mathcal{N}(\mu,\sigma)$ » ou encore $X \sim \mathcal{N}(\mu,\sigma)$.

La courbe représentative de la densité de probabilité $f_X(t)$ prend alors la forme caractérisique d'une courbe « en cloche » :



Si X suit la loi $\mathcal{N}(\mu,\sigma)$ on peut calculer :

$$E(X) = \mu \qquad \sigma(X) = \sigma$$

Si de plus $\mu = 0$ et $\sigma = 1$ on dit que X **suit loi normale (ou Gaussienne) centrée réduite**.

Une variable aléatoire X qui suit une loi normale est évidemment continue. Cette loi joue néanmoins un rôle majeur dans la modélisation des variables aléatoires expérimentales en raison des deux faits suivants.

- Quand une variable aléatoire X mesure une grandeur expérimentale qui est la résultante de nombreuses petites actions indépendantes, alors X suit à peu près une loi normale.
- Il suffit d'une table des valeurs approchées de F_N où N suit une loi $\mathcal{N}(0,1)$ pour connaître les valeurs approchées de F_X pour toute variable aléatoire X suivant une loi $\mathcal{N}(\mu,\sigma)$.

Le premier résultat ci-dessus s'énonce sous la forme plus précise du théorème « de la limite centrale ». Malheureusement nous ne disposons pas des outils nécessaires pour le présenter ici. Nous nous contenterons donc d'admettre comme une vérité empirique son application pratique la plus directe, à savoir qu'un très grand nombre de variables aléatoires expérimentales peuvent être modélisées par une Gaussienne.

Le second résultat est simplement une application du théorème suivant.

Théorème 1.4.12 *Une variable aléatoire X suit une loi $\mathcal{N}(\mu,\sigma)$ si et seulement si la variable aléatoire $N = (X-\mu)/\sigma$ suit une loi $\mathcal{N}(0,1)$.*

Pour tout réel a on a évidemment :

$$X \leq a \iff \frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}$$

Autrement dit "$X \leq a$" et "$(X-\mu)/\sigma \leq (a-\mu)/\sigma$" sont un seul et même évènement. En notant $N = (X-\mu)/\sigma$ on aura donc :

$$F_X(a) = F_N\left(\frac{a-\mu}{\sigma}\right)$$

Le théorème 1.4.12 assure que si X suit une loi $\mathcal{N}(\mu,\sigma)$ alors N suit une loi $\mathcal{N}(0,1)$, et donne donc un moyen de calculer les valeurs de F_X à partir d'une table des valeurs de F_N pour la loi $\mathcal{N}(0,1)$. En outre, les valeurs de $F_N(a)$ pour $a \leq 0$ se déduisent de celles pour $a \geq 0$ par la formule :

$$F_N(-a) = 1 - F_N(a)$$

Signalons enfin que la loi normale permet aussi de modéliser la loi binomiale $\mathcal{B}(n,p)$ de même moyenne et de même écart-type, lorsque n est assez grand et que p n'est pas trop proche de 0 ou de 1.

Théorème 1.4.13 (Approximation d'une loi binomiale par une loi normale.) *Soit X une variable aléatoire suivant une loi $\mathcal{B}(n,p)$. Posons $q = 1-p$. Si $n \geq 30$ et si $\frac{5}{n} \leq p \leq 1 - \frac{5}{n}$ alors pour tous $k, l \in \mathbf{N}$:*

$$P(k \leq X \leq l) \approx P(k - 0,5 \leq Y \leq l + 0,5)$$

où Y suit la loi $\mathcal{N}(np, \sqrt{npq})$.

Remarque 1.4.14 L'élargissement de l'intervalle $[k,l]$ en $[k-0,5, l+0,5]$ est dû au fait qu'on modélise une variable aléatoire discrète, à valeurs entières, par une variable aléatoire continue. On appelle cet élargissement une *correction de continuité*.

Exemple 1.4.15 Dans l'exemple 1.4.4 nous avions utilisé une loi binomiale $\mathcal{B}(2000; 0,85)$ pour modéliser la variable aléatoire mesurant le nombre $G(\omega)$ de graines ayant germé dans une plantation ω. Le calcul de $E(G) = 1700$ et de $\sigma(G) \approx 16$ s'était avéré facile, alors que celui de $P(1684 \leq G \leq 1716)$ nécessitait d'évaluer une somme passablement compliquée comportant 33 termes. Ici $n = 2000$ est supérieur à 30, et $p = 0,85$ est compris entre $5/2000 \approx 0,0025$ et $1 - 5/2000 \approx 0,9975$. On peut donc considérer que G suit sensiblement une loi $\mathcal{N}(1700; 16)$, autrement dit que la variable aléatoire $N = (X - 1700)/16$ suit une loi $\mathcal{N}(0, 1)$:

$$\begin{aligned} & P(1683 \leq G \leq 1716) \\ \approx\ & P(1683,5 \leq 16N + 1700 \leq 1716,5) \\ =\ & P\left(\frac{1683,5 - 1700}{16} \leq N \leq \frac{1716,5 - 1700}{16}\right) \\ =\ & P(-1,03 \leq N \leq 1,03) \\ =\ & F_N(1,03) - F_N(-1,03) \end{aligned}$$

Il suffit alors de recourir à une table pour lire $F_N(1,03) \approx 0,848\,5$. Comme $F_N(-1,03) = 1 - F_N(1,03)$ il vient :

$$P(1684 \leq G \leq 1716) \approx 2F_N(1,03) - 1 \approx 0,697\,0$$

Autrement dit, le modèle gaussien que nous avons choisi, prédit environ 70% de chances que le nombre de graines qu'on verra germer se situe entre 1684 et 1716.

Remarque 1.4.16 Le calcul de $P(1684 \leq G \leq 1716)$ par la formule de l'exemple 1.4.6 à l'aide d'un ordinateur donne en fait $0,698\,6$. On touche ici du doigt l'écart, modeste mais pas nul, entre le modèle binomial et son approximation gaussienne.

Chapitre 2

Estimation

Pour beaucoup d'esprits, une probabilité calculée à sept ou huit décimale près est beaucoup plus convaincante qu'un argument fondé sur des considérations qualitatives. Ces esprits oublient que si le calcul en question est fondé sur des éléments statistiques, qui ne sont donc pas numériquement précis, le nombre de décimales est une pure illusion.

René Thom

La statistique est la première des sciences inexactes.

Edmond et Jules de Goncourt

2.1 Généralités sur l'estimation ponctuelle

On veut estimer la valeur γ d'une grandeur Γ sur une population Ω. Cette population étant trop grande, on est contraint de se restreindre à un échantillon de n individus. La mesure de Γ sur cet échantillon donne une valeur g qu'on appelle **estimation ponctuelle** de γ (sous-entendu : sur cet échantillon).

Question 2.1.1 Dans quelle mesure peut-on se fier à cette estimation ponctuelle ?

Bien entendu la valeur g pourra être très différente d'un échantillon à l'autre, et aussi très différente de γ. Néanmoins, naïvement, on est en droit d'espérer que la moyenne de g (moyenne prise sur l'ensemble de *tous* les échantillons de taille n) soit assez proche de γ.

On considère donc la population $\Omega[n]$ de tous les échantillons de taille n extraits de Ω, et la variable aléatoire $G_n : \Omega[n] \to \mathbf{R}$ qui à un échantillon ω de taille n associe la valeur de Γ sur cet échantillon [1]. Si

[1] Attention ici à ne pas prendre ω pour un individu de Ω. C'est au contraire un individu de $\Omega[n]$, c'est-à-dire que ω lui-même est un échantillon de n individus de Ω.

$E(G_n) = \gamma$ on dit alors que G_n est un **estimateur sans biais** de γ, sinon on dit parle d'**estimateur biaisé**. Nous allons voir quelques grandeurs Γ pour lesquels l'estimateur G_n construit comme ci-dessus sera sans biais, et une grandeur pour laquelle il sera biaisé.

2.2 Estimation ponctuelle d'une fréquence

La valeur à estimer est la probabilité p qu'un individu de la population Ω possède un caractère A donné, autrement dit la fréquence de ce caractère dans la population Ω. On considère $F_n : \Omega[n] \to \mathbf{R}$ qui à tout échantillon ω de taille n associe la fréquence du caractère A dans cet échantillon.

Théorème 2.2.1 $E(F_n) = p$

Autrement dit F_n est un estimateur sans biais de p.

Exemple 2.2.2 On veut estimer la faculté germinative p d'une espèce donnée. La population Ω de toutes les graines de cette espèce (présentes, passées et futures) étant illimitée, on en extrait un échantillon de 200 graines que l'on plante de façon appropriée. Quelques semaines plus tard on constate que 163 d'entre elles on germé. Une estimation ponctuelle de p est alors donnée par $f = 163/200 = 81,5\,\%$.

2.3 Estimation ponctuelle d'une moyenne

La valeur à estimer est la moyenne μ d'une variable aléatoire $X : \Omega \to \mathbf{R}$. On considère $\overline{X}_n : \Omega[n] \to \mathbf{R}$ qui à tout échantillon ω de taille n associe la moyenne $\overline{x}$ des valeurs de X sur ω.

Théorème 2.3.1 $E(\overline{X}_n) = \mu$

Autrement dit $\overline{X}_n$ est un estimateur sans biais de μ.

Exemple 2.3.2 Pour estimer le poids moyen des français à partir d'un échantillon donné de 500 personnes, on se contentera de calculer la moyenne des poids de ces 500 personnes.

2.4 Estimation ponctuelle d'une variance

La valeur à estimer est la variance σ^2 d'une variable aléatoire $X : \Omega \to \mathbf{R}$. On considère $S_n : \Omega[n] \to \mathbf{R}$ qui à tout échantillon ω de taille n associe l'écart-type de X sur ω. Puisque $S_n(\omega)^2$ est la variance de X sur ω, on peut espérer que S_n^2 soit un estimateur sans biais de σ^2. Et bien il n'en est rien !

Théorème 2.4.1 $E(S_n^2) = \dfrac{n-1}{n}\sigma^2$

Posons $\widetilde{S}_n = \sqrt{\dfrac{n}{n-1}} S_n$. Le théorème ci-dessus entraîne que :

$$E(\widetilde{S}_n^2) = E\left(\frac{n}{n-1} S_n^2\right) = \frac{n}{n-1} E(S_n^2) = \sigma^2$$

Autrement dit $\widetilde{S}_n^2$ est un estimateur sans biais de σ^2. Pour cette raison, on appelle **variance débiaisée de X sur un échantillon de taille n** le nombre :

$$\text{variance débiaisée} = \frac{n}{n-1} \times (\text{variance de } X \text{ sur l'échantillon de taille } n)$$

La racine carrée de la variance débiaisée fournit une estimation de l'écart-type σ et sera appelée pour cette raison l'**écart-type estimé** de X sur cet échantillon :

$$\text{écart-type estimé} = \sqrt{\frac{n}{n-1}} \times (\text{écart-type de } X \text{ sur l'échantillon de taille } n)$$

Remarque 2.4.2 Sur certaines calculatrices, la touche $\boxed{\sigma_n}$ calcule la variance tandis que la touche $\boxed{\sigma_{n-1}}$ calcule la variance débiaisée.

Exemple 2.4.3 La mesure des tailles de 10 étudiants de L2-SVG pris au hasard a donné les résultats suivants (en cm) :

$$168, 172, 173, 175, 176, 176, 177, 180, 183, 188$$

Cette série statistique a une moyenne et un écart-type de :

$$\overline{x} = \frac{168 + 172 + 173 + 175 + 2 \times 176 + 177 + 180 + 183 + 188}{10} = 176,8$$

$$s = \sqrt{\frac{(168 - 176,8)^2 + (172 - 176,8)^2 + \cdots + (188 - 176,8)^2}{10}} \approx 5,42$$

La taille moyenne de tous les étudiants de L2-SVG, sera donc estimée à 176,8 cm tandis que son écart-type sera estimé à $\widetilde{s} = \sqrt{10/9} \times 5,42 \approx 5,71$.

2.5 Généralités sur l'estimation par intervalle de confiance

Lors de l'estimation ponctuelle d'une probabilité p (resp. de la moyenne μ d'une variable aléatoire X) nous avons préconisé d'utiliser l'estimateur F_n (resp. $\overline{X}_n$) parce que le calcul de leur espérance à montré qu'ils étaient sans biais. Dans ces deux cas on peut aussi calculer :

$$V(F_n) = \frac{p(1-p)}{n} \qquad \left(\text{resp. } V(\overline{X}_n) = \frac{1}{n} V(X)\right)$$

Ceci fournit une indication de l'écart à craindre, pour un échantillon ω_0 fixé de taille n, entre l'estimation $F_n(\omega_0)$ et p (resp. entre $\overline{X}_n(\omega_0)$ et μ) mais rien ne garantit que l'écart ne soit pas plus important.

Plus généralement, quand on cherche à estimer une valeur γ à l'aide d'un estimateur sans biais G_n portant sur les échantillons de taille n, on aimerait pouvoir déterminer pour tout $\alpha \in]0,1[$ et toute modalité g

de G_n un intervalle $]a_\alpha(g), b_\alpha(g)[$ tel que $\gamma \in]a_\alpha(g), b_\alpha(g)[$ avec une probabilité $1-\alpha$. On dit alors que $]a_\alpha(g), b_\alpha(g)[$ est un **intervalle de confiance pour γ au coefficient de risque α** (ou **au coefficient de sécurité $1-\alpha$**) **relatif**[2] **à une observation** g.

En pratique, toute la difficulté va être de ramener le problème à la lecture d'une table qui, étant donnée une variable aléatoire X suivant une loi donnée, fournit les valeurs approchées de $P(|X| \geq x)$ pour différentes valeurs de x. Nous disposons en particulier de ces tables lorsque :
– X suit une loi $\mathcal{N}(0,1)$ (table 2) ;
– X suit une loi de Student à $\nu \leq 30$ degrés de libertés (table 3) ;
– X suit une loi du χ^2 à $\nu \leq 30$ degrés de libertés (table 4).
Comme nous allons le voir, une telle réduction n'a rien d'élémentaire. Elle passe à chaque fois par des théorèmes spécifiques, que nous ne donnerons pas mais qui nécessitent dans chaque cas d'introduire quelques hypothèses supplémentaires (raisonnables en pratique).

2.6 Intervalle de confiance d'une probabilité

La valeur à estimer est la probabilité p qu'un individu de la population Ω possède un caractère A donné. On dispose de l'estimateur sans biais F_n que nous avons introduit en 2.2 et on veut déterminer un intervalle de confiance pour p au risque α.

Hypothèse supplémentaire. Nous supposerons que n est assez grand pour que les hypothèses du théorème 1.4.13 soient satisfaites : $n \geq 30$ et $5/n \leq p \leq 1-5/n$.

Pour tout échantillon ω de taille n, $nF_n(\omega)$ est le nombre d'individus ayant le caractère A dans ω. La variable aléatoire nF_n suit donc une loi $\mathcal{B}(n,p)$. Notre hypothèse supplémentaire sur n permet alors, grâce au théorème 1.4.13, d'affirmer que nF_n suit sensiblement une loi $\mathcal{N}(np, \sqrt{npq})$ (où $q = 1-p$). Autrement dit nous pouvons supposer que la variable aléatoire N définie par :

$$N = \frac{nF_n - np}{\sqrt{npq}} = \frac{F_n - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

suit la loi $\mathcal{N}(0,1)$. Sur la table 2 on peut donc lire u_α tel que $P(|N| \geq u_\alpha) \approx \alpha$, c'est-à-dire :

$$P(|N| < u_\alpha) \approx 1-\alpha$$

On en déduit :

$$P\left(|F_n - p| < u_\alpha \sqrt{\frac{p(1-p)}{n}}\right) \approx 1-\alpha$$

D'où enfin :

$$P\left(F_n - u_\alpha\sqrt{\frac{pq}{n}} < p < F_n + u_\alpha\sqrt{\frac{pq}{n}}\right) \approx 1-\alpha$$

En pratique, nous disposons d'un unique échantillon ω_0 de taille n, et *nous ne connaissons pas* la valeur de p, qu'il s'agit justement d'estimer. En revanche nous connaissons la valeur de $f = F_n(\omega_0)$. On peut

[2]Par abus de langage, on parle le plus souvent d'intervalle de confiance au risque α, sans mentionner l'observation g.

alors démontrer (c'est un théorème) que l'erreur introduite en substituant f à p dans la formule ci-dessus est en fait négligeable. Ceci permet de conclure que :

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

est un intervalle de confiance de p au risque α relatif à l'observation f.

2.7 Intervalle de confiance d'une moyenne

On considère une variable aléatoire $X : \Omega \to \mathbf{R}$ d'espérance μ et d'écart-type σ. Nous disposons de l'estimateur sans biais $\overline{X}_n$ que nous avons introduit en 2.3 et nous voulons déterminer un intervalle de confiance pour μ au risque α. Pour y parvenir il va nous falloir nous autoriser quelques hypothèses supplémentaires.

Hypothèses supplémentaires. Nous supposerons que n n'est pas trop petit ($n > 30$) ou que X suit une loi normale. Dans ce cas, on sait montrer (c'est un théorème) que $\overline{X}_n$ suit sensiblement une loi normale. Nous avons déjà vu que $E(\overline{X}_n) = \mu$ et $V(\overline{X}_n) = V(X)/n$ d'où $\sigma(\overline{X}_n) = \sigma/\sqrt{n}$. La variable aléatoire N définie par :

$$N = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

suit donc une loi $\mathcal{N}(0,1)$. On peut donc lire sur la table 2 un réel u_α tel que $P(|N| \geq u_\alpha) \approx \alpha$, c'est-à-dire :

$$P(|N| < u_\alpha) \approx 1 - \alpha$$

On en déduit :

$$P\left(|\overline{X}_n - \mu| < u_\alpha\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

D'où :

$$P\left(\overline{X}_n - u_\alpha\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + u_\alpha\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

2.7.1 Cas particulier : σ est connu

En pratique, nous disposons d'un unique échantillon ω_0 de taille n, sur lequel X prend la valeur $\overline{x} = \overline{X}_n(\omega_0)$. Si $n > 30$ ou si X suit une loi normale, et si de plus l'écart-type σ de X est connu, alors le raisonnement ci-dessus montre que :

$$\left]\overline{x} - u_\alpha\frac{\sigma}{\sqrt{n}}, \overline{x} + u_\alpha\frac{\sigma}{\sqrt{n}}\right[$$

est un intervalle de confiance pour μ au risque α relatif à l'observation $\overline{x}$.

Le hic est qu'en général σ n'est absolument pas connu. Dans ce cas, il est tentant de remplacer dans la formule ci-dessus σ par $\widetilde{s}$, la racine carrée de la variance débiaisée de X sur ω_0, puisque celle-ci est connue et qu'on sait qu'elle fournit une estimation ponctuelle (non biaisée) de σ^2. Encore faut-il pouvoir contrôler l'erreur introduite par cette substitution. Ceci nous amène à distinguer deux cas.

2.7.2 Cas des grands échantillons ($n > 30$, X quelconque)

On considère la variable aléatoire :

$$\widetilde{N} = \frac{\overline{X}_n - \mu}{\widetilde{S}_n/\sqrt{n}}$$

où $\widetilde{S}_n : \Omega[n] \to \mathbf{R}$ est la variable aléatoire introduite en 2.4 ($\widetilde{S}_n^2$ mesure la variance débiaisée de X sur les échantillons de taille n).

Comme $n > 30$ on peut montrer (c'est un théorème) que quelle que soit la loi suivie par X, $\widetilde{N}$ suit encore une loi $\mathcal{N}(0,1)$. Sur la table 2 on peut donc lire u_α tel que $P(|\widetilde{N}| \geq u_\alpha) \approx \alpha$, c'est-à-dire :

$$P(|\widetilde{N}| < u_\alpha) \approx 1 - \alpha$$

Le même calcul que précédemment, en remplaçant σ par $\widetilde{S}_n$, donne alors :

$$P\left(\overline{X}_n - u_\alpha \frac{\widetilde{S}_n}{\sqrt{n}} < \mu < \overline{X}_n + u_\alpha \frac{\widetilde{S}_n}{\sqrt{n}}\right) \approx 1 - \alpha$$

En pratique, on dispose d'un unique échantillon ω_0 de taille n, sur lequel X a une valeur moyenne $\overline{x} = \overline{X}_n(\omega_0)$ et une variance débiaisée $\widetilde{s}^2 = \widetilde{S}_n^2(\omega_0)$. Sachant que $n > 30$, nous pouvons conclure de ce qui précède que :

$$\left]\overline{x} - u_\alpha \frac{\widetilde{s}}{\sqrt{n}}, \overline{x} + u_\alpha \frac{\widetilde{s}}{\sqrt{n}}\right[$$

est un intervalle de confiance pour μ au risque α relatif à l'observation $\overline{x}$.

2.7.3 Cas des petits échantillons ($n \leq 30$, X normale)

Comme $n \leq 30$, $\widetilde{N}$ s'écarte trop de la loi normale pour que l'intervalle ci-dessus reste un intervalle de confiance. Cependant comme X suit une normale, on peut démontrer (c'est un théorème) que $\widetilde{N}$ suit une loi classique, appelée **loi de Student à $n-1$ degrés de libertés**. On peut lire sur la table 3 un réel t_α tel que $P(|\widetilde{N}| \geq t_\alpha) = \alpha$, c'est-à-dire :

$$P(|\widetilde{N}| < t_\alpha) = 1 - \alpha$$

On en déduit :

$$P\left(|\overline{X}_n - \mu| < t_\alpha \frac{\widetilde{S}_n}{\sqrt{n}}\right) \approx 1 - \alpha$$

D'où :

$$P\left(\overline{X}_n - t_\alpha \frac{\widetilde{S}_n}{\sqrt{n}} < \mu < \overline{X}_n + t_\alpha \frac{\widetilde{S}_n}{\sqrt{n}}\right) \approx 1 - \alpha$$

En pratique, nous disposons d'un unique échantillon ω_0 de taille n, sur lequel X à une valeur moyenne $\overline{x} = \overline{X}_n(\omega_0)$ et une variance débiaisée $\widetilde{s}^2 = \widetilde{S}_n^2(\omega_0)$. Sachant que $n \leq 30$ et que X suit une loi normale, nous pouvons conclure de ce qui précède que :

$$\left]\overline{x} - t_\alpha \frac{\widetilde{s}}{\sqrt{n}}, \overline{x} + t_\alpha \frac{\widetilde{s}}{\sqrt{n}}\right[$$

est un intervalle de confiance pour μ au risque α relatif à l'observation $\overline{x}$.

2.8 Intervalle de confiance d'une variance

On considère à nouveau une variable aléatoire $X : \Omega \to \mathbf{R}$ d'espérance μ et d'écart-type σ. Nous disposons des variables aléatoires S_n et $\widetilde{S}_n$ que nous avons introduit en 2.4 et nous voulons déterminer un intervalle de confiance pour σ au risque α. Comme pour μ, nous n'y parviendrons pas sans hypothèse supplémentaire.

Hypothèse supplémentaire. Cette fois nous supposons d'emblée que X suit une loi normale. On considère alors la nouvelle variable aléatoire :

$$Y = \frac{\sqrt{n-1}}{\sigma}\widetilde{S}_n = \frac{\sqrt{n}}{\sigma}S_n$$

2.8.1 Cas des grands échantillons ($n - 1 > 30$, X normale)

Dans ce cas on peut montrer (c'est un théorème) que Y suit sensiblement une loi $\mathcal{N}(\sqrt{(2n-3)/2}, 1/\sqrt{2})$, donc que la variable aléatoire :

$$N = \frac{Y - \sqrt{(2n-3)/2}}{1/\sqrt{2}} = \sqrt{2}Y - \sqrt{2n-3}$$

suit une loi $\mathcal{N}(0,1)$. Sur la table 2 on peut donc lire u_α tel que $P(|N| \geq u_\alpha) \approx \alpha$, c'est-à-dire :

$$P(|N| < u_\alpha) \approx 1 - \alpha$$

Autrement dit $P(|\sqrt{2}Y - \sqrt{2n-3}| < u_\alpha) \approx 1 - \alpha$, d'où l'on déduit :

$$P\left(\frac{\sqrt{2n-3} - u_\alpha}{\sqrt{2}} < Y < \frac{\sqrt{2n-3} - u_\alpha}{\sqrt{2}}\right) \approx 1 - \alpha$$

En remplaçant Y par $\dfrac{\sqrt{n-1}}{\sigma}\widetilde{S}_n$ il vient, après quelques calculs :

$$P\left(\frac{\sqrt{2(n-1)}\widetilde{S}_n}{\sqrt{2n-3}+u_\alpha} < \sigma < \frac{\sqrt{2(n-1)}\widetilde{S}_n}{\sqrt{2n-3}-u_\alpha}\right) \approx 1-\alpha$$

En pratique, on dispose d'un échantillon ω_0 de taille n et on connaît la variance débiaisée $\widetilde{s}^2 = \widetilde{S}_n(\omega_0)^2$ de X sur ω_0. Sachant que $n-1 > 30$ et que X suit une loi normale, le raisonnement ci-dessus montre alors que :

$$\left]\frac{2(n-1)\widetilde{s}^2}{(\sqrt{2n-3}+u_\alpha)^2}, \frac{2(n-1)\widetilde{s}^2}{(\sqrt{2n-3}-u_\alpha)^2}\right[$$

est un intervalle de confiance pour σ^2 au risque α relatif à l'observation $\widetilde{s}^2$.

2.8.2 Cas des petits échantillons ($n-1 \leq 30$, X normale)

Dans ce cas on peut montrer (c'est un théorème) que Y^2 suit une loi classique, appelée **loi du χ^2 à $n-1$ degrés de libertés**. Pour $1 \leq n-1 \leq 30$ la table 4 permet de lire a_α et b_α tels que :

$$P(Y^2 \geq b_\alpha) = \frac{\alpha}{2} \qquad \text{et} \qquad P(Y^2 \geq a_\alpha) = 1-\frac{\alpha}{2}$$

Alors $P(a_\alpha < Y^2 < b_\alpha) = (1-\alpha/2) - \alpha/2 = 1-\alpha$. En remplaçant Y^2 par $(n-1)\widetilde{S}_n^2/\sigma^2$ il vient après de laborieux (mais élémentaires) calculs :

$$P\left(\frac{(n-1)\widetilde{S}_n^2}{b_\alpha} < \sigma^2 < \frac{(n-1)\widetilde{S}_n^2}{a_\alpha}\right) \approx 1-\alpha$$

En pratique, on dispose d'un unique échantillon ω_0 de taille n sur lequel on calcule la variance débiaisée $\widetilde{s}^2 = \widetilde{S}_n^2(\omega_0)$. Sachant que $n-1 \leq 30$ et que X suit une loi normale, le raisonnement ci-dessus permet de conclure :

$$\left]\frac{(n-1)\widetilde{s}^2}{b_\alpha}, \frac{(n-1)\widetilde{s}^2}{a_\alpha}\right[$$

est un intervalle de confiance pour σ^2 au risque α relatif à l'observation $\widetilde{s}^2$.

2.9 Estimation par seuil

Dans certains cas, on ne veut pas vraiment estimer si un paramètre θ se trouve dans un certain intervalle, mais plutôt s'il est ou non supérieur à un certain seuil. Par exemple, lors d'une élection où s'affrontent

deux candidats, ce qui est déterminant n'est pas tant d'estimer (ponctuellement ou par intervalle) la proportion p du corps électoral favorable au candidat Dupont, que de d'évaluer avec un risque donné si p est ou non supérieur au seuil fatidique des 50 %.

Ce type d'estimation par seuil se fait exactement comme l'estimation par intervalle, à une petite (mais cruciale) différence près. Dans l'estimation par intervalle, que ce soit pour les paramètres p, μ ou σ, nous avons introduit un variable aléatoire N (ou $\widetilde{N}$) dépendant du problème donné, qui suivait une loi $\mathcal{N}(0,1)$. La détermination d'un intervalle de confiance se faisait alors en fonction du nombre u_α tel que $P(|N| \geq u_\alpha) \approx \alpha$. Dans le cas d'une estimation par seuil, ce n'est pas l'écart en valeur absolue mais l'écart tout court qui nous intéresse. On cherchera donc plutôt un nombre u'_α tel que $P(N \geq u'_\alpha) \approx \alpha$, c'est-à-dire $P(N \leq u'_\alpha) \approx 1-\alpha$ (on lit u'_α sur la table 1, alors que u_α se lisait sur la table 2). Le reste du calcul se fait exactement comme dans le cas d'une estimation par intervalle.

À titre d'exemple, nous traitons ici le cas d'une estimation par seuil de la proportion p d'une population Ω qui possède un caractère A. Les autres cas (estimation par seuil de μ et σ) sont laissés en exercice au lecteur. Le risque α est fixé arbitrairement. Tout comme en 2.6, on introduit la variable aléatoire $F_n : \Omega[n] \to \mathbf{R}$ qui à tout échantillon ω de taille n associe la fréquence du caractère A dans cet échantillon, et la variable aléatoire N définie par :

$$N = \frac{F_n - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

Tout comme en 2.6, on suppose que $n \geq 30$ et $5/n \leq p \leq 1 - 5/n$. Le théorème 1.4.13 garantit alors que N suit sensiblement la loi $\mathcal{N}(0,1)$. On lit donc sur la table 1 un réel u'_α tel que $P(N \leq u'_\alpha) \approx 1-\alpha$, et on en déduit apres calcul :

$$P\left(F_n - u'_\alpha\sqrt{\frac{p(1-p)}{n}} \leq p\right) \approx 1-\alpha$$

En pratique, nous disposons d'un unique échantillon ω_0 de taille n, et *nous ne connaissons pas* la valeur de p. En revanche nous connaissons la valeur de $f = F_n(\omega_0)$. On peut démontrer (c'est un théorème) que l'erreur introduite en substituant f à p dans la formule ci-dessus est en fait négligeable.

Conclusion : Au risque α de se tromper, on peut affirmer que $p \geq f - u'_\alpha\sqrt{\dfrac{f(1-f)}{n}}$.

Chapitre 3

Tests statistiques

> Ne croyez pas ce que disent les statistiques avant d'avoir soigneusement examiné ce qu'elles ne disent pas.
>
> William W. Watt

> Je ne crois aux statistiques que lorsque je les ai moi-même falsifiées.
>
> Winston Churchill

3.1 Généralités sur les tests statistiques

Un test statistique est un mécanisme qui permet de trancher entre deux hypothèses (H_0) et (H_1) au vu des résultats d'un échantillon. La décision aboutira à choisir (H_0) ou (H_1), à tort ou à raison. Il y a donc 4 possibilités avec les probabilités correspondantes :

	(H_0) vraie	(H_1) vraie
(H_0) retenue	$1-\alpha$	β
(H_1) retenue	α	$1-\beta$

α est appelé **risque de première espèce**. Il mesure la probabilité d'écarter (H_0) à tort (faux négatif). Le **risque de deuxième espèce**, que l'on note ici β, mesure quant à lui la probabilité de conserver (H_0) alors qu'elle est fausse (faux positif).

Contrairement à ce qu'on pourrait naïvement supposer, les hypothèses (H_0) et (H_1) ne jouent pas des rôles symétriques. D'une part l'hypothèse alternative (H_1), souvent implicite, n'est pas nécessairement le contraire de l'hypothèse nulle (H_0). Mais surtout (H_0) est privilégiée par tous les tests statistiques que nous rencontrerons, ce qui revient à considérer que le risque de la rejeter si elle est vraie est beaucoup plus coûteux que celui de la conserver à tort. Le test est donc construit uniquement en fonction du risque

α, fixé arbitrairement et assez petit ($\alpha \leq 0,1$). À l'inverse le risque β est calculé : il mesure la probabilité que le test *construit en fonction du risque* α amène un faux positif (voir exemple ci-après).

Exemple 3.1.1 Des relevés effectués pendant de nombreuses années ont permis d'établir que le niveau naturel des pluies dans la Beauce en millimètres par an suit approximativement une loi $\mathcal{N}(600, 100)$. Des entrepreneurs, surnommés faiseurs de pluie, prétendaient pouvoir augmenter en moyenne le niveau de pluie annuel de 50 mm par insémination des nuages au moyen d'iodure d'argent. Leur procédé fut mis à l'essai entre 1951 et 1959 et on releva les hauteurs de pluie suivantes :

Année	1951	1952	1953	1954	1955	1956	1957	1958	1959
mm	510	614	780	512	501	534	603	788	650

Le niveau moyen de pluie annuel sur cette période fut donc d'environ $\overline{x} = 610$ mm. Que pouvait-on en conclure ?

Deux hypothèses s'affrontaient ici : ou bien l'insémination était sans effet (H_0), ou bien elle augmentait réellement de 50 mm en moyenne le niveau de pluie annuel (H_1). Attention cependant à ne pas se méprendre sur le sens de ces hypothèses : les moyennes en question n'étaient pas les moyennes sur 9 ans, sans quoi la réponse eut été immédiate. Pour être plus précis, commençons par modéliser.

La variable aléatoire $X : \omega \to \mathbf{R}$ dont il est question ici est évidemment celle qui mesure la pluviométrie annuelle, la « population » étant l'ensemble des années. Les données expérimentales ont permis d'affirmer que X suit sensiblement une loi $\mathcal{N}(600, 100)$. Quand les faiseurs de pluie affirment que leur procédé augmente en moyenne la pluviométrie de 50 mm par an, c'est de l'espérance de X qu'ils parlent, et non de sa valeur moyenne sur un échantillon de 9 années. Ils n'affirment évidemment pas non plus qu'il va pleuvoir 650 mm de pluie chaque année. Les deux hypothèses s'énoncent ainsi :

(H_0) L'insémination des nuages est sans effet notable, et X continue donc à suivre une loi $\mathcal{N}(600, 100)$ même quand ce procédé est utilisé.

(H_1) L'insémination augmente l'espérance de X de 50 mm/an, et X suit donc avec ce procédé une loi $\mathcal{N}(650, 100)$.

Le fait que sur les neuf années de l'expérience la pluviométrie moyenne ait été de 610 mm/an ne contredit donc pas directement l'hypothèse (H_1). Elle ne la rend pas non plus ridicule puisqu'on peut montrer que la probabilité, sous l'hypothèse (H_1), de constater un pluviométrie moyenne inférieure ou égale à 610 mm/an sur 9 ans est d'environ 12 % : c'est peu, mais ce n'est pas négligeable. La difficulté de trancher entre les deux hypothèses vient de ce que l'écart entre les valeurs moyennes dans les deux cas est nettement inférieur à l'écart-type.

Les agriculteurs, hésitant à opter pour le procédé onéreux des faiseurs de pluie, penchaient naturellement pour l'hypothèse (H_0). Pour les convaincre, il fallait que l'expérience contredise nettement la validité de (H_0), autrement dit que les niveaux de pluie observés pendant la durée de l'expérience traduisent une éventualité très improbable compte tenu de (H_0).

Ils choisirent $\alpha = 0,05$ comme risque de première espèce, c'est-à-dire qu'ils se déclarèrent prêts à accepter (H_1) si le résultat obtenu s'avérait faire partie d'une éventualité qui n'avait que 5 % de chance de se produire si (H_0) était vraie.

Puisque nous voulons mesurer la cohérence entre l'hypothèse (H_0) et la valeur moyenne $\overline{x} = 610$ constatée pour X sur cet échantillon de $n = 9$ années, on est amené à considérer la population $\Omega[9]$ de tous les échantillons de taille 9 extraits de Ω, et la variable aléatoire $\overline{X}_9 : \Omega[9] \to \mathbf{R}$ qui mesure la valeur moyenne de X sur chacun de ces échantillons. Parce que X suit une loi $\mathcal{N}(600, 100)$, on peut montrer (c'est un théorème, que nous avons déjà rencontré au chapitre 2) que $\overline{X}_9$ suit une loi $\mathcal{N}(600, 100/\sqrt{9})$. Autrement dit, la variable aléatoire :

$$N = \frac{\overline{X}_9 - 600}{100/3}$$

suit une loi $\mathcal{N}(0, 1)$. Pour que les agriculteurs soient convaincus, il aurait fallu que sur l'échantillon de 9 années qu'a duré l'expérience, non seulement $\overline{X}_9 \geq 600$ c'est-à-dire $N \geq 0$ mais que N atteigne au moins la plus petite valeur $u_{0,05}$ telle que :

$$P(N \geq u_{0,05}) \approx 0,05$$

Autrement dit :

$$P(N \leq u_{0,05}) \approx 0,95$$

Puisque N suit une loi $\mathcal{N}(0, 1)$ on lit sur la table 1 que $P(N \leq x) \geq 0,95$ si et seulement si $x \geq 1,65$. On retiendra donc $u_{0,05} = 1,65$ comme valeur critique pour N.

$$N \geq 1,65 \iff \overline{X}_9 = \frac{100}{3}N + 600 \geq 655$$

Conclusion : Sous l'hypothèse (H_0) on avait moins de 5 % de chances d'observer sur 9 ans un niveau moyen de pluie annuel supérieure ou égale à 655 mm/an. Si un tel évènement avait été observé, les agriculteurs auraient rejeté l'hypothèse (H_0) et accepté (H_1). Puisqu'au contraire on a constaté une pluviométrie moyenne de $\overline{x} = 610$ mm/an, la conclusion a été de conserver (H_0) : les valeurs observées pouvaient être dues au hasard, en l'absence de toute influence de l'iodure d'argent.

Cependant, rien ne dit que conserver H_0 n'était pas une erreur. Le risque de seconde espèce nous renseigne utilement là-dessus. Sous l'hypothèse (H_1) en effet, $\overline{X}_9$ suit une loi $\mathcal{N}(650, 100/\sqrt{9}$ et donc :

$$N_1 = \frac{\overline{X}_9 - 650}{100/3}$$

suit une loi $\mathcal{N}(0, 1)$. Même si l'insémination avait bien l'effet annoncé, la probabilité que les faiseurs de pluie « passent le test » n'était donc, compte tenu de la correction de continuité, que de :

$$P(\overline{X}_9 \geq 655,5) = P\left(\frac{\overline{X}_9 - 650}{100/3} \geq \frac{650 - 655,5}{100/3}\right) = P(N_1 \geq -0,135) = 1 - P(N_1 \leq 0,165) \approx 44\,\%$$

Autrement dit le risque de seconde espèce était ici $\beta = 56\,\%$, ce qui est considérable !

Signalons que l'usage, en biologie, semble être d'ignorer superbement le risque de seconde espèce. En dehors de l'exemple ci-dessus, nous nous conformerons donc à cet usage.

3.1.2 Tests du χ^2

Dans ce chapitre, nous commencerons par présenter trois tests généraux : un test de conformité, un test d'homogénéité et un test d'indépendance. Ces trois tests ont en commun de résoudre le problème

de décision par l'introduction d'une variable aléatoire suivant une loi du χ^2, et sont pour cette raison appelés tous les trois des « tests du χ^2 ». En outre, contrairement aux tests paramétriques de conformité et d'homogénéité que nous présenterons ensuite, ces tests robustes ne nécessitent aucune hypothèse supplémentaire sur la loi des variables aléatoires intervenant dans la modélisation.

3.1.3 Tests paramétriques de conformité

Nous présenterons ensuite les principaux **tests paramétriques**. On appelle ainsi les tests portant sur un paramètre d'une variable aléatoire (fréquence, moyenne, variance). Il s'agira à chaque fois de déterminer avec un risque fixé si une valeur θ_0 observée sur un échantillon est conforme à la valeur θ de ce paramètre sur la population totale, c'est-à-dire si la différence n'est pas significative et peut être attribuée aux fluctuations d'échantillonnages.

Partant de l'hypothèse (H_0) que cette différence n'est pas significative, et d'un risque de première espèce α fixé arbitrairement, nous construirons deux types de tests :
- Les tests **bilatéraux** où l'on ne suppose aucune relation entre θ et θ_0. L'hypothèse (H_0) sera rejetée avec un risque α si la différence $\theta - \theta_0$ sort d'un certain intervalle.
- Les test **unilatéraux** où l'on suppose *a priori* que θ est supérieur (ou inférieur) à θ_0. L'hypothèse (H_0) sera rejetée avec un risque α si $\theta - \theta_0$ se trouve au-delà (resp. en deçà) d'un certain seuil.

Dans l'exemple 3.1.1 le test portait sur la valeur moyenne de la variable aléatoire X mesurant le niveau de pluie annuel. On supposait *a priori* la pluviométrie plus importante avec insémination que sans (c'est-à-dire que toute valeur inférieure à 600 mm/an aurait évidemment conduit à conserver (H_0)). Il s'agissait donc d'un test paramétrique unilatérale sur le paramètre $\mu = E(X)$. L'hypothèse (H_0) aurait été rejetée avec un risque de 5 % si la différence $\overline{x} - \mu$ s'était trouvée au-delà du seuil critique de 55 mm/an.

3.1.4 Tests paramétrique d'homogénéité

En fin de chapitre nous aborderons également quelques tests d'homogénéités pour lesquels on étudie un certain paramètre sur deux populations (qui peuvent aussi bien être la même population à deux moments différents). On dispose d'un échantillon de chacune de ces populations, sur lesquels on observe que ce paramètre prend les valeurs θ_1 et θ_2. La question est de savoir si la différence entre θ_1 et θ_2 est significative d'une différence entre les populations, ou si elle s'explique par des fluctuations d'échantillonnages.

Exemple 3.1.5 On compare le taux de réussite au bac d'un groupe de filles et d'un groupe de garçons. Avec un risque de 5 %, peut-on affirmer que la différence constatée entre ces deux groupes démontre globalement un écart significatif entre garçons et filles ?

Exemple 3.1.6 Un même groupe de 30 copies est soumis à une double correction. Avec un risque de 5 %, peut-on affirmer que l'écart entre les moyennes des deux correcteurs est significatif ?

Dans ce type de test, l'hypothèse (H_0) sera toujours que la différence n'est pas significative, autrement dit que les deux populations sont homogènes du point de vue du paramètre étudié.

3.2 Test du χ^2 de conformité à une loi théorique

Une variable aléatoire empirique $X : \Omega \to \mathbf{R}$ prends ses valeurs dans r intervalles $I_1, \ldots, I_r$ deux à deux disjoints et recouvrant $\mathbf{R}$. On dispose seulement d'un échantillon de taille n extrait de Ω. Pour chaque $i \leq r$ on compte sur cet échantillon le nombre de fois où X prend ses valeurs dans I_i, c'est-à-dire le nombre N_i d'individus ω de cet échantillon tels que $X(\omega) \in I_i$.

L'examen de cette répartition amène à construire un modèle théorique, c'est-à-dire à introduire une variable aléatoire $Y : \Omega \to \mathbf{R}$ dont la fonction de répartition est connue, et semble fournir une bonne approximation de celle de X. On formule alors l'hypothèse suivante :

(H_0) Les données statistiques recueillies sur l'échantillon sont conformes au modèle théorique qui prédit que $P(X \in I) \approx P(Y \in I)$ pour tout intervalle I.

Sous l'hypothèse (H_0), pour chaque intervalle I_i la fréquence constatée N_i/n de l'évènement "$X \in I_i$" devrait donc être proche de $p_i = P(Y \in I_i)$, autrement dit :

$$N_i \approx np_i$$

Les np_i sont appelés **effectifs théoriques** ou **effectifs calculés**. Pour évaluer dans quelle mesure l'hypothèse (H_0) est correcte, on compare les effectifs calculés aux effectifs constatés à l'aide de l'expression suivante :

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

La valeur D^2 mesure l'écart entre les N_i et les np_i sur cet échantillon. On aurait pu construire d'autres mesures de cet écart, mais l'intérêt de celle-ci apparaîtra dans le théorème 3.2.1.

Si l'hypothèse (H_0) est exacte, la valeur de D^2 doit être faible *en général*, c'est-à-dire sur la plupart des échantillons de taille n. Néanmoins il existera toujours des échantillons, en principe non représentatifs, pour lesquels la valeur de D^2 sera plus élevée. Comment définir un seuil à partir duquel la valeur de D^2 sera jugée trop élevée pour qu'on puisse conserver (H_0) ?

Remarquons que dans l'expression de D^2 donnée ci-dessus ce sont les N_i qui varient quand on passe d'un échantillon à l'autre (en gardant la même taille n). Les effectifs théoriques np_i, eux, sont fixés. D^2 est donc une variable aléatoire sur l'ensemble $\Omega[n]$ des échantillons de taille n extraits de Ω.

Théorème 3.2.1 *Sous l'hypothèse (H_0), et si les effectifs théoriques np_i ne sont pas trop petits (en pratique si les $np_i \geq 5$), alors la variable aléatoire D^2 suit sensiblement une loi du χ^2 à $n-1$ degrés de liberté.*

Pour tester l'hypothèse (H_0), on se donne un coefficient $\alpha \in [0, 1]$ arbitrairement choisi (en général petit) qui représente le risque qu'on accepte de prendre en écartant (H_0), autrement dit le risque de première espèce. Puisque le théorème 3.2.1 nous donne la loi de D^2, on peut lire sur la table 4 une valeur b_α telle que $P(D^2 \geq b_\alpha) \approx \alpha$.

Conclusion : On peut écarter l'hypothèse (H_0) avec une probabilité α de se tromper, si et seulement si la valeur constatée de D^2 sur l'échantillon considéré (valeur qu'on note souvent χ^2_c) est supérieure ou égale à b_α.

Remarque 3.2.2 Signalons que l'hypothèse sur les np_i n'est nullement contraignante en pratique puisqu'on dispose d'une totale liberté sur le choix des intervalles I_i. La seule contrainte à respecter est qu'ils restent deux à deux disjoints et recouvrent $\mathbf{R}$. Si par exemple X prends des valeurs entières entre 3 et 12, on pourra considérer les intervalles $I_0 =]-\infty; 3,5]$, $I_i =]i-0,5; i+0,5]$ pour $1 \leq i \leq 12$ et $I_{13} =]12,5; +\infty[$. Mais si les effectifs théoriques np_i pour certains intervalles I_i sont trop petits, rien n'interdit de faire des regroupements d'intervalles contiguës pour pouvoir appliquer le théorème 3.2.1. On prendra garde toutefois à faire le moins de regroupements possibles : plus le nombre d'intervalles retenus sera important, plus le test sera pertinent.

Cas où la loi théorique n'est pas entièrement connue. Le théorème 3.2.1 ne s'applique que quand la loi théorique à laquelle on veut tester la conformité de X est entièrement connue. En pratique, il pourra arriver que ce ne soit pas le cas, et que sa détermination complète passe par l'estimation de certains paramètres. Typiquement, on pense que X suit un loi normale $\mathcal{N}(\mu, \sigma)$ mais il faut d'abord estimer μ et/ou σ (qui dans la population générale ne sont pas connus) grâce à une estimation ponctuelle réalisée sur l'échantillon dont on dispose, comme il a été vu au chapitre 2. Dans ce cas on applique le théorème suivant :

Théorème 3.2.3 *Sous l'hypothèse (H_0), et si les effectifs théoriques np_i ne sont pas trop petits (en pratique si les $np_i \geq 5$), alors la variable aléatoire D^2 suit sensiblement une loi du χ^2 à $n-d-1$ degrés de liberté, où d est le nombre de paramètres de la loi théorique qu'il a fallu estimer.*

Notons que quand la loi théorique est connue on n'a besoin d'estimer aucun paramètre, autrement dit $d = 0$ et l'énoncé ci-dessus se ramène alors à celui du théorème 3.2.1.

3.3 Test du χ^2 d'homogénéité entre échantillons

On considère comme précédemment une variable aléatoire X prenant ses valeurs dans r intervalles disjoints $I_1, \ldots, I_r$. Cette fois on dispose de s échantillons $E_1, \ldots, E_s$ de tailles $n_1, \ldots, n_s$ respectivement. On constate que sur l'échantillon E_j, X prend N_{ij} fois ses valeurs dans I_i. L'examen de ces différentes distributions amène à formuler l'hypothèse suivante :

(H_0) La distribution des valeurs de X est homogène dans les différents échantillons. Les différences observées sont des fluctuations normales d'échantillonnage.

On réunit les différents échantillons en un seul gros échantillon de taille $n = n_1 + \cdots + n_s$. Sur ce gros échantillon l'évènement "$X \in I_i$" survient donc avec une probabilité :

$$p_i = \frac{1}{n} \sum_{j=1}^{s} N_{ij}$$

L'hypothèse (H_0) prédit que sur chaque échantillon E_j l'effectif N_{ij} constaté pour l'évènement "$X \in I_j$" est sensiblement égal à $n_j p_i$. Les $n_j p_i$ sont donc appelés **effectifs théoriques** ou **effectifs calculés**.

Pour mesurer l'écart entre cette prédiction et la réalité, on introduit :

$$D^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - n_j p_i)^2}{n_j p_i}$$

Ceci définit une variable aléatoire D^2 sur l'ensemble des s-uplets d'échantillons de tailles respectives $n_1, \ldots, n_s$.

Théorème 3.3.1 *Sous l'hypothèse (H_0), et si les effectifs théoriques ne sont pas trop petits (en pratique si les $n_j p_i \geq 5$), alors D^2 suit une loi du χ^2 à $(r-1)(s-1)$ degrés de liberté.*

Étant donné un risque de première espèce α arbitrairement fixé, le théorème 3.3.1 permet de lire sur la table 4 une valeur b_α telle que $P(D^2 \geq b_\alpha) \approx \alpha$.

Conclusion : On peut écarter l'hypothèse (H_0) avec un risque α de se tromper, si et seulement si la valeur constatée pour D^2 sur les échantillons donnés est supérieure ou égale à b_α.

Remarque 3.3.2 Si certains effectifs théoriques $n_j p_i$ sont < 5 on procédera à des regroupements d'intervalles.

3.4 Test du χ^2 d'indépendance de deux caractères

On étudie cette fois deux variables aléatoires expérimentales X et Y sur une même population Ω, prenant leurs valeurs dans des intervalles $I_1, \ldots, I_r$ et $J_1, \ldots, J_s$ respectivement. Sur un échantillon E de taille n extrait de Ω, on compte, pour tout $i \leq r$ et tout $j \leq s$, le nombre N_{ij} d'individus ω de Ω tels que $X(\omega) \in I_i$ et $Y(\omega) \in J_k$. La somme de tous les N_{ij} est donc le nombre n d'individus dans l'échantillon considéré. En outre :

- $m_i = \sum_{j=1}^{s} N_{ij}$ = le nombre d'individus ω dans l'échantillon tels que $X(\omega) \in I_i$.
- $n_j = \sum_{i=1}^{r} N_{ij}$ = le nombre d'individus ω dans l'échantillon tels que $Y(\omega) \in J_j$.

Mathématiquement, on dit que les caractères mesurés par les variables aléatoires X et Y sont **indépendants** si pour tout intervalle I et J de $\mathbf{R}$:

$$P(X \in I \text{ et } Y \in J) = P(X \in I) \times P(Y \in J) \tag{3.1}$$

Dans ce cas, si l'échantillon est représentatif de la population Ω, on devrait avoir pour tout i et j :

$$\frac{N_{ij}}{n} \approx \frac{m_i}{n} \times \frac{n_j}{n}$$

Autrement dit, l'hypothèse (H_0) prédit :

$$N_{ij} \approx \frac{m_i n_j}{n}$$

On appelle cette fois **effectifs constatés** ou **calculés** les $n_{ij} = \frac{m_i n_j}{n}$.

(H_0) Les caractères mesurés par X et Y sont indépendants. L'écart constaté entre les N_{ij} et les n_{ij} s'explique par les fluctuations d'échantillonnage.

Pour tester cette hypothèse, on introduit donc la variable aléatoire suivante, qui mesure l'écart entre effectifs théoriques et constatés sur tout échantillon de taille n :

$$D^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - n_{ij})^2}{n_{ij}}$$

Là encore, quand on passe d'un échantillon à l'autre les effectifs constatés N_{ij} peuvent varier tandis qu'on laisse fixes les effectifs théoriques n_{ij}.

Théorème 3.4.1 *Sous l'hypothèse* (H_0)*, et si les* n_{ij} *ne sont pas trop petits (en pratique si les* $n_{ij} \geq 5$*), alors* D^2 *suit une loi du* χ^2 *à* $(r-1)(s-1)$ *degrés de libertés.*

Étant donné un risque de première espèce α arbitrairement fixé, le théorème 3.4.1 permet de lire sur la table 4 une valeur b_α telle que $P(D^2 \geq b_\alpha) \approx \alpha$.

Conclusion : On peut écarter l'hypothèse (H_0) avec un risque α de se tromper, si et seulement si la valeur constatée pour D^2 sur l'échantillon donné est supérieure ou égale à b_α.

Remarque 3.4.2 Comme précédemment, si certains effectifs théoriques $n_j p_i$ sont < 5 on procédera à des regroupements d'intervalles.

3.5 Test de conformité d'une fréquence

On étudie dans une population Ω la probabilité p qu'un individu possède un caractère A donné. Sur un échantillon de taille n, on a constaté que ce caractère apparaissait avec une fréquence f. La question est de savoir si cet échantillon est représentatif de la population pour le caractère A, c'est-à-dire si la différence entre f et p est explicable par les fluctuations d'échantillonnage. On veut donc tester l'hypothèse suivante, avec un risque de première espèce α arbitrairement choisi dans l'intervalle $[0, 1]$:

(H_0) La fréquence f observée sur l'échantillon est conforme à la fréquence p sur la population.

Pour cela on considère la population $\Omega[n]$ des échantillons de taille n extraits de Ω, et la variable aléatoire $F_n : \Omega[n] \to \mathbf{R}$ estimateur sans biais de p (voir chapitre 2, section 2.2). On note ω_0 l'échantillon considéré. Par définition on a donc $f = F_n(\omega_0)$. Enfin on fixe un risque de première espèce $\alpha \in [0, 1]$.

Pour tester (H_0) il suffit, par exemple, de trouver un intervalle $]a_\alpha(p), b_\alpha(p)[$ tel que la probabilité que $F_n - p$ n'appartienne pas à cet intervalle soit environ égale à α. Dans ce cas le test consiste à vérifier si la valeur $f - p$ observée sur l'échantillon appartient à cet intervalle. Dans le cas contraire, on pourra écarter l'hypothèse (H_0) avec une probabilité au plus α de se tromper.

Remarque 3.5.1 Bien que ce test ressemble à s'y méprendre à la construction d'un intervalle de confiance, il y a une différence importante : ici p est connu ! Le but n'est donc plus d'estimer une

fréquence p inconnue à l'aide d'un intervalle (dont les bornes ne peuvent pas dépendre de p) mais de fournir un test à l'hypothèse (H_0) à l'aide d'un intervalle (dont les bornes peuvent dépendre de p puisque celui-ci est connu).

Hypothèses supplémentaires. On suppose n assez grand pour que $n \geq 30$ et $5/n \leq p \leq 1 - 5/n$. Tout comme dans la section 2.6 du chapitre 2 on sait qu'alors la variable aléatoire $N : \Omega[n] \to \mathbf{R}$ définie par :

$$N = \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$$

suit la loi $\mathcal{N}(0,1)$. Sur l'échantillon dont nous disposons, N prend la valeur :

$$u = \frac{f - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Test bilatéral. Sur la table 2, nous pouvons lire u_α tel que $P(|N| \geq u_\alpha) \approx \alpha$. On a donc :

$$P\left(|F_n - p| \geq u_\alpha \sqrt{\frac{p(1-p)}{n}}\right) \approx \alpha$$

L'hypothèse (H_0) peut donc être rejetée avec un risque α de se tromper, si et seulement si f n'appartient pas à l'intervalle $]p - u_\alpha\sqrt{p(1-p)/n}, p + -u_\alpha\sqrt{p(1-p)/n}[$. En pratique toutefois, il est plus simple de faire porter le test directement sur u plutôt que sur f.

Conclusion : l'hypothèse (H_0) peut être rejetée avec un risque α de se tromper, si et seulement si $u \notin]-u_\alpha, u_\alpha[$.

Test unilatéral. Cette fois on suppose *a priori* que $f \geq p$, autrement dit que $u \geq 0$. On peut lire sur la table 1 une valeur u'_α telle que $P(N \leq u'_\alpha) \approx 1 - \alpha$, autrement dit $P(N \geq u'_\alpha) \approx \alpha$.

Conclusion : l'hypothèse (H_0) peut être rejetée avec un risque α de se tromper, si et seulement si $u \geq u'_\alpha$.

Remarque 3.5.2 Le cas d'un test unilatéral où l'on suppose *a priori* que $f \leq p$ se traite de la même façon. Avec la même valeur u'_α que ci-dessus, comme $P(N \leq -u'_\alpha) = P(N \geq u'_\alpha)$ on peut conclure que : l'hypothèse (H_0) peut être rejetée avec un risque α de se tromper, si et seulement si $u \leq -u'_\alpha$.

3.6 Test de conformité d'une moyenne

On étudie une variable aléatoire $X : \Omega \to \mathbf{R}$ d'espérance μ et d'écart-type σ. Sur un échantillon de taille n extrait de Ω on a constaté une valeur moyenne de X égale à $\overline{x}$. Connaissant μ et $\overline{x}$, mais pas forcément σ, le problème est de savoir si l'échantillon est représentatif de la population. Il faut donc évaluer dans quelle mesure la différence entre μ et $\overline{x}$ est significative, ou peut être imputée à des fluctuations d'échantillonnage.

(H_0) La moyenne $\overline{x}$ constatée sur l'échantillon est conforme à l'espérance μ de X sur la population globale.

Pour tester (H_0) avec un risque de première espèce α, on considère une nouvelle fois la variable aléatoire $\overline{X}_n : \Omega[n] \to \mathbf{R}$ introduite au chapitre 2.

Hypothèses supplémentaires. Nous supposerons que n n'est pas trop petit ($n > 30$) ou que X suit une loi normale. Dans ce cas, comme nous l'avons vu au chapitre 2, on sait montrer que la variable aléatoire N définie par :

$$N = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

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

3.6.1 Cas particulier : σ est connu

Alors N prend sur notre échantillon une valeur u connue :

$$u = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

Test bilatéral. Sur la table 2 on peut lire u_α tel que $P(|N| \geq u_\alpha) \approx \alpha$ ou encore :

$$P(-u_\alpha < N < u_\alpha) \approx 1 - \alpha$$

Conclusion : l'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $u \notin]-u_\alpha, u_\alpha[$.

Test unilatéral. Cette fois on suppose *a priori* $\overline{x} \geq \mu$, autrement dit $N \geq 0$. Sur la table 1 on peut lire u'_α tel que $P(N \leq u'_\alpha) \approx 1 - \alpha$ ou encore $P(N \geq u'_\alpha) \approx \alpha$.

Conclusion : l'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $u \geq u'_\alpha$.

Remarque 3.6.2 Le cas d'un test unilatéral où l'on suppose *a priori* que $\overline{x} \geq \mu$ se traite de la même façon. Avec la même valeur u'_α que ci-dessus, comme $P(N \leq -u'_\alpha) = P(N \geq u'_\alpha)$ on peut conclure que : l'hypothèse (H_0) peut être rejetée avec un risque α de se tromper, si et seulement si $u \leq -u'_\alpha$.

Naturellement, dans le cas général, σ n'est pas connu et on doit se reporter à l'un ou l'autre des deux cas suivants.

3.6.3 Cas des grands échantillons ($n > 30$, X quelconque)

Comme σ n'est pas connu, on considère au lieu de N la variable aléatoire $\widetilde{N}$ définie par :

$$\widetilde{N} = \frac{\overline{X}_n - \mu}{\widetilde{S}_n/\sqrt{n}}$$

où $\widetilde{S}_n^2 : \Omega[n] \to \mathbf{R}$ mesure la variance débiaisée sur les échantillons de taille n. Sur l'échantillon dont nous disposons, cette variance débiaisée $\widetilde{s}^2$ est connue et donc aussi la valeur $\widetilde{u}$ de $\widetilde{N}$:

$$\widetilde{u} = \frac{\overline{x} - \mu}{\widetilde{s}/\sqrt{n}}$$

Comme $n > 30$ on a vu au chapitre 2 que $\widetilde{N}$ suit encore sensiblement une loi $\mathcal{N}(0,1)$.

Test bilatéral. Sur la table 2 on peut lire u_α tel que $P(|\widetilde{N}| \geq u_\alpha) \approx \alpha$ ou encore :

$$P(-u_\alpha < \widetilde{N} < u_\alpha) \approx 1 - \alpha$$

Conclusion : l'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $\widetilde{u} \notin]-u_\alpha, u_\alpha[$.

Test unilatéral. Cette fois on suppose *a priori* $\overline{x} \geq \mu$, autrement dit $\widetilde{N} \geq 0$. Sur la table 1 on peut lire u'_α tel que $P(\widetilde{N} \leq u'_\alpha) \approx 1 - \alpha$ ou encore $P(\widetilde{N} \geq u'_\alpha) \approx \alpha$.

Conclusion : l'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $\widetilde{u} \geq u'_\alpha$.

Remarque 3.6.4 Si au contraire on suppose *a priori* $\overline{x} \leq \mu$ alors, symétriquement, l'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $\widetilde{u} \leq -u'_\alpha$.

3.6.5 Cas des petits échantillons ($n \leq 30$, X normale)

Comme $n \leq 30$ et X suit une loi normale, on a vu au chapitre 2 que $\widetilde{N}$ suit une loi de Student à $n-1$ degrés de liberté.

Test bilatéral. Sur la table 3 on peut lire t_α tel que $P(|\widetilde{N}| \geq t_\alpha) \approx \alpha$ ou encore :

$$P(-t_\alpha < \widetilde{N} < t_\alpha) \approx 1 - \alpha$$

Conclusion : l'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $\widetilde{u} \notin]-t_\alpha, t_\alpha[$.

Test unilatéral. Cette fois on suppose *a priori* $\overline{x} \geq \mu$, autrement dit $\widetilde{N} \geq 0$. Sur la table 1 on peut lire t'_α tel que $P(\widetilde{N} \leq t'_\alpha) \approx 1 - \alpha$ ou encore $P(\widetilde{N} \geq t'_\alpha) \approx \alpha$.

Conclusion : l'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $\widetilde{u} \geq t'_\alpha$.

Remarque 3.6.6 Si au contraire on suppose *a priori* $\overline{x} \leq \mu$ alors, symétriquement, l'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $\widetilde{u} \leq -t'_\alpha$.

3.7 Test de conformité d'une variance

On étudie une variable aléatoire $X : \Omega \to \mathbf{R}$ d'espérance μ et de variance σ^2. Sur un échantillon de taille n extrait de Ω on a constaté pour X une valeur moyenne $\overline{x}$ et une variance débiaisée $\widetilde{s}$. Le problème est de savoir dans quelle mesure la différence entre $\widetilde{s}$ et σ est significative, ou peut être imputée à des fluctuations d'échantillonnage.

(H_0) La variance débiaisée $\widetilde{s}^2$ constatée sur l'échantillon est conforme à la variance σ^2 de X sur la population.

Pour tester (H_0) avec un risque de première espèce α, on considère la variable aléatoire :

$$Y = \frac{\sqrt{n-1}}{\sigma}\widetilde{S}_n = \frac{\sqrt{n}}{\sigma}S_n$$

introduite au chapitre 2.

Hypothèses supplémentaires. On suppose que X est normale. Dans ce cas on sait, par théorème, que la variable aléatoire Y^2 suit sensiblement une loi :
– χ^2 à $n-1$ degré de libertés si $n-1 \leq 30$;
– $\mathcal{N}(\sqrt{(2n-3)/2}, 1/\sqrt{2})$ si $n-1 > 30$.

3.7.1 Cas des grands échantillons ($n-1 > 30$, X normale)

Dans ce cas $N = \sqrt{2}Y - \sqrt{2n-3}$ suit sensiblement une loi $\mathcal{N}(0,1)$ et prend sur notre échantillon une valeur u connue :

$$u = \frac{\sqrt{2(n-1)}}{\sigma}\widetilde{s} - \sqrt{2n-3}$$

Test bilatéral. On lit sur la table 2 le nombre u_α tel que $P(|N| \geq u_\alpha) \approx \alpha$.

Conclusion : L'hypothèse (H_0) peut être rejetée avec un risque de première espèce α si et seulement si $u \notin]-u_\alpha, u_\alpha[$.

Test unilatéral. On suppose *a priori* $\widetilde{s} \geq \sigma$. On lit sur la table 1 le nombre u'_α tel que $P(N \leq u'_\alpha) \approx 1-\alpha$ et donc $P(N \geq u'_\alpha) \approx \alpha$.

Conclusion : L'hypothèse (H_0) peut être rejetée avec un risque de première espèce α si et seulement si $u \geq u'_\alpha$.

Remarque 3.7.2 Si au contraire on suppose *a priori* $\widetilde{s} \leq \sigma$ alors, symétriquement, l'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $u \leq -u_\alpha$.

3.7.3 Cas des petits échantillons ($n-1 \leq 30$, X normale)

Dans ce cas Y^2 suit sensiblement une loi du χ^2 à $n-1$ degrés de liberté, et prend sur notre échantillon une valeur y^2 connue :

$$y^2 = \frac{n-1}{\sigma^2}\widetilde{s}^2$$

Test bilatéral. On lit sur la table 4 les nombres a_α et b_α tels que :

$$P(Y^2 \geq b_\alpha) \approx \frac{\alpha}{2} \quad \text{et} \quad P(Y^2 \geq a_\alpha) \approx 1 - \frac{\alpha}{2}$$

On a donc :

$$P(a_\alpha \leq Y^2 \leq b_\alpha) \approx 1 - \alpha$$

Conclusion : L'hypothèse (H_0) peut être rejetée avec un risque de première espèce α si et seulement si $y^2 \notin]a_\alpha, b_\alpha[$.

Test unilatéral. On suppose *a priori* $\widetilde{s} \geq \sigma$. On lit sur la table 4 le nombre b'_α tel que $P(Y^2 \geq b'_\alpha) \approx \alpha$.

Conclusion : L'hypothèse (H_0) peut être rejetée avec un risque de première espèce α si et seulement si $y^2 \geq b'_\alpha$.

Remarque 3.7.4 Si au contraire on suppose *a priori* $\widetilde{s} \leq \sigma$ alors on lit sur la table 4 le nombre a'_α tel que $P(Y^2 \geq a'_\alpha) \approx 1 - \alpha$, c'est-à-dire $P(Y^2 \leq a'_\alpha) \approx \alpha$. L'hypothèse (H_0) peut être écartée avec un risque α de se tromper, si et seulement si $y^2 \leq a'_\alpha$.

3.8 Homogénéité des fréquences dans deux échantillons

Après les tests paramétrique de conformité, nous abordons maintenant un premier exemple de test paramétrique d'homogénéité.

Dans deux populations Ω_1 et Ω_2 on étudie la probabilité p_1 et p_2 respectivement, qu'un individu possède un caractère A donné. Sur un échantillon de taille n_1 (resp. n_2) extrait de Ω_1 (resp. Ω_2), on constate que le caractère A apparaît avec une fréquence f_1 (resp. f_2). Contrairement au cas du test de conformité (section 3.5), p_1 et p_2 ne sont pas nécessairement connus. On veut tester avec un risque de première espèce α fixé, l'hypothèse suivante :

(H_0) $p_1 = p_2$. Autrement dit la différence entre f_1 et f_2 n'indique pas une différence significative entre p_1 et p_2.

Exemple 3.8.1 L'exemple 3.1.5, où l'on étudie la taux de réussite au bac chez les garçons et chez les filles à partir de deux échantillons, est de ce type.

Remarque 3.8.2 Le test du χ^2 d'homogénéité peut aussi tout-à-fait s'appliquer à ce type de situation.

Soit $F_1 : \Omega_1[n_1] \to \mathbf{R}$ qui mesure la fréquence du caractère A sur chaque échantillon de taille n_1 extrait de Ω_1. On sait que f_1 suit une loi binomiale $\mathcal{B}(n_1, p_1)$. De même la variable aléatoire $F_1 : \Omega_2[n_1] \to \mathbf{R}$ définie de manière analogue suit une loi $\mathcal{B}(n_2, p_2)$.

Hypothèses supplémentaires. On suppose n_1 et n_2 assez grands pour que F_1 et F_2 suivent sensiblement une loi normale (voir le théorème 1.4.13 du chapitre 1). Sous l'hypothèse (H_0), p_1 et p_2 sont égaux. La meilleure estimation de la valeur commune est :

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

On peut alors montrer que[1] $F_1 - F_2$ suit une loi :

$$\mathcal{N}\left(0, \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

où $\hat{q} = 1 - \hat{p}$. Autrement dit la variable aléatoire N définie par :

$$N = \frac{F_1 - F_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

suit une loi $\mathcal{N}(0, 1)$. Sa valeur u sur le couple d'échantillon considéré est connue :

$$u = \frac{f_1 - f_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Test bilatéral. On lit sur la table 2 un réel u_α tel que $P(|N| \geq u_\alpha) \approx \alpha$.

Conclusion : On rejette l'hypothèse (H_0) avec un risque de première espèce α, si et seulement si $u \notin]-u_\alpha, u_\alpha[$.

Test unilatéral. Si on suppose *a priori* que $f_1 \geq f_2$, on lit sur la table 1 un réel u'_α tel que $P(N \leq u_\alpha) \approx 1 - \alpha$, c'est-à-dire $P(N \geq u_\alpha) \approx \alpha$.

Conclusion : On rejette l'hypothèse (H_0) avec un risque de première espèce α, si et seulement si $u \geq u_\alpha$.

Exemple 3.8.3 Sur 96 pièces venant d'un fournisseur A, 12 sont défectueuses. Sur 55 pièces venant d'un fournisseur B, 15 sont défectueuses. Peut-on affirmer avec un risque de 5 % que les proportions de pièces défectueuses chez ces deux fournisseurs sont significativement différentes ? Et avec un risque de 1 % ?

On calcule $f_1 \approx 0,13$ et $f_2 \approx 0,27$ ainsi que $\hat{p} = (12 + 15)/(96 + 55) \approx 0,18$, d'où :

$$u = \frac{f_1 - f_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx -2,28$$

[1] Les fonctions F_1 et F_2 n'étant pas définies sur le même ensemble, la notation $F_1 - F_2$ est abusive. Elle désigne ici la fonction de $\Omega_1[n_1] \times \Omega_2[n_2]$ dans $\mathbf{R}$ qui à un *couple* d'échantillons (ω_1, ω_2) associe $F_1(\omega_1) - F_2(\omega_2)$.

Sur la table 2 on lit $u_{0,05} \approx 1,96$ et $u_{0,01} \approx 2,58$. On peut donc rejeter l'hypothèse que les deux fournisseurs ont globalement le même pourcentage de pièces défectueuses, avec un risque de 5 % de se tromper mais pas avec un risque de 1 % de se tromper.

3.9 Homogénéité des moyennes d'échantillons appariés

Un même échantillon d'individus est soumis à deux mesures successives d'une même grandeurs. On veut tester l'hypothèse que les moyennes des deux séries de mesures sont semblables.

Exemple 3.9.1 L'exemple 3.1.6 d'un paquet de copies ayant subit une double correction est de ce type. On peut aussi penser aux mesures effectuées chez des cobayes avant et après un traitement donné.

Bien que les deux mesures portent ici sur le même groupe d'individus, il peut être commode de considérer comme deux échantillons distincts : d'une part ce groupe d'individus lors de la première mesure, et d'autre part ce même groupe lors de la seconde mesure. Dans ce cas on parle d'**échantillons appariés**, pour indiquer que les deux échantillons sont en fait constitués des mêmes individus.

Un autre point de vue, que nous adopterons, est de considérer que nous n'avons qu'un seul échantillon de taille n mais deux variables aléatoires X_1 et X_2 correspondant à chacune des deux mesures effectuées.

Sur notre échantillon on a constaté pour X_1 et X_2 une valeur moyenne de $\overline{x}_1$ et $\overline{x}_2$ respectivement. L'hypothèse à tester, avec un risque de première espèce α est donc :

(H_0) La différence entre $\overline{x}_1$ et $\overline{x}_2$ n'est pas significative.

Pour cela on considère la variable aléatoire $D = X_1 - X_2$, dont la valeur moyenne sur l'échantillon considéré est donc $\overline{d} = \overline{x}_1 - \overline{x}_2$. Sous l'hypothèse (H_0) la variable aléatoire D doit avoir une espérance nulle puisque $E(D) = E(X_1) - E(X_2)$. L'hypothèse (H_0) peut donc se reformuler ainsi :

(H_0) La moyenne $\overline{d}$ constatée sur l'échantillon est conforme à $E(D) = 0$.

On est ainsi ramené à un test de conformité sur une moyenne, voir section 3.6 :
– Si $n > 30$ on utilise une loi normale.
– Si $n \leq 30$ et si X_1, X_2 suivent une loi normale, alors D aussi et on utilise une loi de Student.

Annexe A

Quelques preuves ou idées de preuve

A.1 Loi binomiale

Nous avons rencontré au chapitre 1 le théorème 1.4.3 suivant.

Théorème.
On considère une expérience dans laquelle :
- *un même test est répété n fois ;*
- *chacun des tests peut fournir deux résultats (positif ou négatif) avec probabilité p et $1-p$ respectivement ;*
- *tous ces tests sont indépendants.*

Alors la variable aléatoire X qui compte le nombre de tests positifs dans une telle expérience suit une loi binomiale $\mathcal{B}(n,p)$.

Ce théorème bien pratique ne repose pas sur des outils compliqués, c'est un « simple » résultat de combinatoire. Nous pouvons donc en donner une preuve élémentaire. Soit k un entier fixé entre 0 et n. On veut montrer, avec les hypothèses ci-dessus, que :

$$P(X=k) = C_n^k p^k (1-p)^{n-k}$$

Démonstration : Considérons pour chaque entier i entre 1 et n, la variable aléatoire X_i qui, dans une expérience de ce type, ne retient que le résultat du i-ème test c'est-à-dire que X_i vaut 0 si ce test est négatif, 1 s'il est positif. Alors clairement :

$$X = X_1 + X_2 + \cdots + X_n$$

Notons E_n l'ensemble des n-uplets $(\varepsilon_1, \ldots, \varepsilon_n)$ de 0 et de 1. On sait qu'il y en a 2^n. Chacun représente l'un des résultats possibles pour notre série de n test.

Parmi ceux-ci, quels sont ceux pour lesquels la variable X prendra la valeur k ? Ce sont les n-uplets qui contiennent exactement k fois la valeur 1 et $n-k$ fois la valeur 0. Choisir un tel n-uplet revient à choisir parmi les indices i entre 1 et n ceux pour lesquels on aura $\varepsilon_i = 1$. C'est donc choisir un ensemble de k indices parmi n, ce qui peut se faire de C_n^k façons possibles. L'ensemble $\Omega_{n,k}$ des n-uplets de ce type contient donc C_n^k éléments.

Pour chacun de ces éléments, c'est-à-dire pour chacun des n-uplets de 0 et de 1 qui contiennent exactement k fois la valeur 1, l'hypothèse que les différents tests sont indépendants se traduit mathématiquement par le fait que :

$$P(X_1 = \varepsilon_1 \text{ et } X_2 = \varepsilon_2 \text{ et } \dots \text{ et } X_n = \varepsilon_n) = P(X_1 = \varepsilon_1) \times P(X_2 = \varepsilon_2) \times \cdots \times P(X_n = \varepsilon_n)$$

Or pour chaque i on a $P(X_i = 1) = p$ et $P(X_i = 0) = 1 - p$. Comme il y a k indices i pour lesquels $\varepsilon_i = 1$ (resp. $n - k$ indices i pour lesquels $\varepsilon_i = 0$) on aura donc k facteurs p (resp. $n - k$ facteurs $1 - p$) dans le produit ci-dessus, soit :

$$P(X_1 = \varepsilon_1 \text{ et } X_2 = \varepsilon_2 \text{ et } \dots \text{ et } X_n = \varepsilon_n) = p^k(1 - p)^{n-k}$$

Comme $X = k$ si et seulement si le résultat $(\varepsilon_1, \dots, \varepsilon_n)$ est un élément de $\Omega_{n,k}$ on a donc :

$$P(X = k) = \sum_{(\varepsilon_1,\dots,\varepsilon_n)\in\Omega_{n,k}} P(X_1 = \varepsilon_1 \text{ et } X_2 = \varepsilon_2 \text{ et } \dots \text{ et } X_n = \varepsilon_n) = C_n^k p^k (1 - p)^{n-k}$$

C.Q.F.D.

A.2 L'approximation de Poisson

Nous avons vu au chapitre 1 le théorème 1.4.10 selon lequel la loi de Poisson de paramètre fournit une bonne approximation de la loi binomiale $\mathcal{B}(n, p)$ de même espérance, sous les hypothèses :

$$n \geq 30 \qquad p \leq 0,1 \qquad np \leq 10$$

Comment démontre-t-on un tel résultat ? Il s'agit en fait d'une formulation pratique du théorème suivant.

Théorème de Poisson.
Soit $(p_n)_{n\in\mathbf{N}}$ une suite de nombres pris dans $[0, 1]$. On suppose que np_n tend vers une limite $\lambda > 0$ quand n tend vers l'infini. Alors la suite $u_n = C_n^k p_n^k (1 - p_n)^{n-k}$ converge elle aussi, vers la limite suivante :

$$\lim_{n\to+\infty} C_n^k p_n^k (1 - p_n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}$$

Posons pour tout entier n, $\lambda_n = np_n$. Par hypothèse λ_n tend vers une limite λ qui est *finie* et *non nulle*. Ceci implique que $p_n = \lambda_n/n$ tend vers 0 quand n tend vers l'infini. Le théorème de Poisson affirme donc que si on prend un réel p dans $[0, 1]$ et un entier n, tels que np ne soit ni trop grand (ce qui nécessite que p soit petit) ni trop petit (il faut donc que n soit grand) alors :

$$C_n^k p^k (1 - p)^{n-k} \approx e^{-\lambda}\frac{\lambda^k}{k!}$$

où $\lambda = np$. C'est exactement ce que dit le théorème 1.4.10, en précisant simplement ce qu'on entend par « petit » ou « grand ». C'est une étude précise de l'erreur dans l'approximation ci-dessus qui a amené à retenir les conditions suivantes en pratique :
– « np pas trop grand » veut dire ici $np \leq 10$.

– « p assez petit » veut dire ici $p \leq 0,1$.
– « n assez grand » veut dire ici $n \geq 30$.
Bien entendu, si par np est entre 10 et 100, on trouvera d'autres valeurs seuils p_0 et n_0 telles que l'approximation ci-dessus de la loi $\mathcal{B}(n,p)$ par la loi $\mathcal{P}(np)$ reste suffisamment bonne dès que $p \leq p_0$ et $n \geq n_0$.

La preuve du théorème de Poisson ne requiert pas non plus d'outils trop sophistiqués, mais tout de même une bonne dextérité dans l'étude des limites de suites.

Démonstration : Nous voulons montrer, sous l'hypothèse que $\lambda_n = np_n$ tend vers $\lambda > 0$, que :

$$\lim_{n\to+\infty} C_n^k p_n^k (1-p_n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}$$

Rappelons que C_n^k vaut :

$$C_n^k = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}$$

Le terme dont nous cherchons la limite peut donc s'écrire :

$$\frac{n(n-1)\cdots(n-k+1)}{k!} p_n^k (1-p_n)^n (1-p_n)^{-k}$$

Comme p_n tend vers 0, le terme $1-p_n$ tend vers 1 et donc le dernier facteur $(1-p_n)^{-k}$ tend vers 1.

Le facteur p_n^k tend vers 0, mais il est multiplié par $n(n-1)\cdots(n-k+1)$ qui tend vers $+\infty$, il y a donc forme indéterminée. Pour la lever on ré-écrit astucieusement :

$$n(n-1)\cdots(n-k+1)p_n^k = \frac{n(n-1)\cdots(n-k+1)}{n\times n\times\cdots\times n} n^k p_n^k$$

Le facteur $n^k p_n^k = (np_n)^k$ tend vers λ^k. Les k autres facteurs sont de la forme :

$$\frac{n-i}{n} = 1 - \frac{i}{n}$$

avec i entre 0 et $k-1$, ils tendent donc tous vers 1. À ce stade nous avons donc montré que :

$$\lim_{n\to+\infty} \frac{n(n-1)\cdots(n-k+1)}{k!} p_n^k (1-p_n)^{-k} = \frac{\lambda^k}{k!}$$

Il ne reste donc plus qu'à vérifier que $(1-p_n)^n$ tend vers $e^{-\lambda}$. C'est la partie la plus délicate. On a :

$$(1-p_n)^n = e^{\ln(1-p_n)^n} = e^{n\ln(1-p_n)}$$

Comme ce qui se trouve dans le logarithme tend vers 1, $\ln(1-p_n)$ tend vers 0. Avec le facteur n qui tend vers l'infini, nous avons donc une nouvelle forme indéterminée. Pour la lever, nous allons revenir à la définition de la dérivée d'une fonction en un point comme limite du taux d'accroissement. Par exemple pour une fonction f dérivable au point 1 :

$$\lim_{x\to 1} \frac{f(x)-f(1)}{x-1} = f'(1)$$

La fonction $f(x) = \ln(x)$ est dérivable sur $]0,+\infty[$, on le sait, et sa dérivée est $1/x$. On a donc :

$$\lim_{x\to 1} \frac{\ln(x)-\ln(1)}{x-1} = \frac{1}{1} = 1$$

En particulier, comme $1 - p_n$ tend vers 1 on a :

$$\lim_{n\to+\infty} \frac{\ln(1-p_n) - \ln(1)}{(1-p_n) - 1} = 1$$

Comme $\ln 1 = 0$ ceci donne après simplification :

$$\lim_{n\to+\infty} \frac{\ln(1-p_n)}{-p_n} = 1$$

D'où aussi :

$$\lim_{n\to+\infty} \frac{n\ln(1-p_n)}{-np_n} = 1$$

Or par hypothèse le dénominateur tend vers $-\lambda$, donc :

$$\lim_{n\to+\infty} n\ln(1-p_n) = \lim_{n\to+\infty} -np_n \times \frac{n\ln(1-p_n)}{-np_n} = \lim_{n\to+\infty} -np_n = -\lambda$$

On en déduit finalement :

$$\lim_{n\to+\infty} (1-p_n)^n = \lim_{n\to+\infty} e^{n\ln(1-p_n)} = e^{-\lambda}$$

C.Q.F.D.

A.3 L'approximation par la loi normale

Le théorème 1.4.13 du chapitre 1 affirme que la loi binomiale $\mathcal{B}(n,p)$ peut être approchée par la loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$, autrement dit par une loi normale de même espérance $\mu_n = np$ et de même variance $\sigma_n^2 = np(1-p)$ que la loi $\mathcal{B}(n,p)$, au prix d'une petite « correction de continuité ».

Théorème.
Soit X une variable aléatoire suivant une loi $\mathcal{B}(n,p)$. Posons $q = 1-p$. Si $n \geq 30$ et si $\frac{5}{n} \leq p \leq 1 - \frac{5}{n}$ alors pour tous $k, l \in \mathbf{N}$:

$$P(k \leq X \leq l) \approx P(k - 0,5 \leq Z \leq l + 0,5)$$

où Z suit la loi $\mathcal{N}(np, \sqrt{npq})$.

Dans ce théorème, comme $X \sim \mathcal{B}(n,p)$ qui est une loi discrète à modalités dans $\mathbf{N}$, en posant $k' = k-1$ nous avons $P(X < k) = P(X \leq k-1) = P(X \leq k'$, donc :

$$P(k \leq X \leq l) = P(X \leq l) - P(X \leq k-1) = P(X \leq l) - P(X \leq k')$$

D'autre part comme $Z \sim \mathcal{N}(\mu_n, \sigma_n)$ qui est une loi continue et comme $k - 0,5 = k' + 0,5$, nous avons $P(Z < k - 0,5) = P(Z \leq k - 0,5) = P(Z \leq k' + 0,5)$ et donc :

$$P(k - 0,5 \leq Z \leq l + 0,5) = P(Z \leq l + 0,5) - P(Z < k - 0,5) = P(Z \leq l + 0,5) - P(Z \leq k' + 0,5)$$

Pour avoir le théorème d'approximation il suffit donc de savoir que pour tout entier k :

$$P(X \leq k) \approx P(Z \leq k + 0,5)$$

Ce dernier résultat est essentiellement une application du théorème suivant.

Théorème de De Moivre et Laplace.
Soit p un réel pris dans $[0,1]$. Pour tout entier $n \in \mathbf{N}^*$ soit X_n une variable aléatoire suivant une loi $\mathcal{B}(n,p)$. Posons $\mu_n = np = E(X_n)$ et $\sigma_n^2 = np(1-p) = V(X_n)$. Alors pour tout réel x fixé on a :

$$\lim_{n\to+\infty} P\left(\frac{X_n - \mu_n}{\sigma_n} \leq x\right) = \frac{1}{2\pi}\int_{-\infty}^{x} e^{-t^2}dt$$

Dans le membre de droite on reconnaît la fonction de répartition de la loi $\mathcal{N}(0,1)$, autrement dit ce théorème dit exactement que si N est une variable aléatoire suivant une loi $\mathcal{N}(0,1)$:

$$\lim_{n\to+\infty} P\left(\frac{X_n - \mu_n}{\sigma_n} \leq x\right) = P(N \leq x)$$

Par conséquent, si $X \sim \mathcal{B}(n,p)$ et si n est assez grand (ce qui se traduit par $n \geq 30$ dans le théorème d'approximation que nous avons donné au chapitre 1), alors pour tout réel x :

$$P\left(\frac{X - \mu_n}{\sigma_n} \leq x\right) \approx P(N \leq x)$$

Et donc, en posant $y = \mu_n + \sigma_n x$ et $Z = \mu_n + \sigma_n N$:

$$P(X \leq y) \approx P(Z \leq y)$$

Comme $N \sim \mathcal{N}(0,1)$ on sait que $Z \sim \mathcal{N}(\mu_n, \sigma_n)$. On retrouve donc presque ce qu'on voulait, en remplaçant y par un entier k quelconque. La seule différence est que la correction de continuité semble avoir disparu !

De plus rien dans le théorème de De Moivre et Laplace ne semble nécessiter l'hypothèse que p soit compris entre $5/n$ et $1 - 5/n$. En fait cette condition et la correction de continuité sont nécessaires seulement pour que l'approximation de la loi binomiale par la loi normale marche dès $n = 30$. Si on ne fait pas cette hypothèse sur p ou si on omet la correction de continuité alors l'approximation sera encore valable mais moins bonne, à moins d'aller chercher des valeurs de n plus grandes ce qui peut s'avérer gênant en pratique.

Argument graphique.

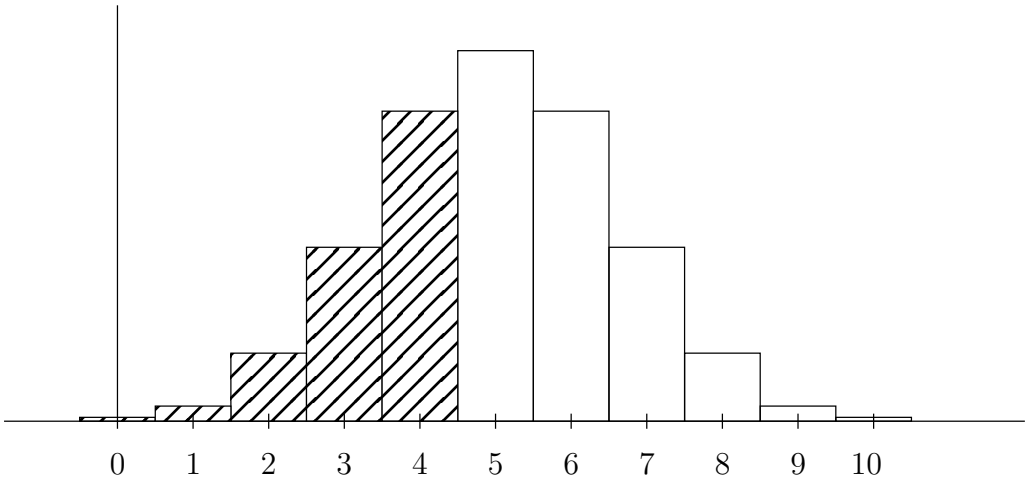
Il n'est pas envisageable de donner ici une preuve du théorème de De Moivre et Laplace : il requiert des outils beaucoup plus sophistiqués que celui de Poisson (fonction charactéristique, convergence en loi...). Signalons tout de même qu'il s'agit d'un cas particulier (le premier démontré historiquement) d'un résultat beaucoup plus général dont nous avons déjà parlé : le théorème de la limite centrale.

On peut cependant se convaincre empiriquement de la justesse du théorème de De Moivre et Laplace, et du même coup saisir la raison d'être de la correction de continuité, à l'aide d'une représentation graphique.

Que représente en effet, pour un entier k entre 0 et n, la probabilité $P(X_n \leq k)$ quand $X_n \sim \mathcal{B}(n,p)$? On sait que :

$$P(X_n \leq k) = \sum_{l=0}^{k} C_n^k p^k (1-p)^k$$

Dans cette somme, chacun des termes peut être vu comme l'aire d'un rectangle de base 1 et de hauteur $C_n^k p^k (1-p)^k$. Leur somme représentera donc l'aire de la zone hachurée dans le dessin ci-dessous (pour $n = 10$, $p = 1/2$ et $k = 4$).

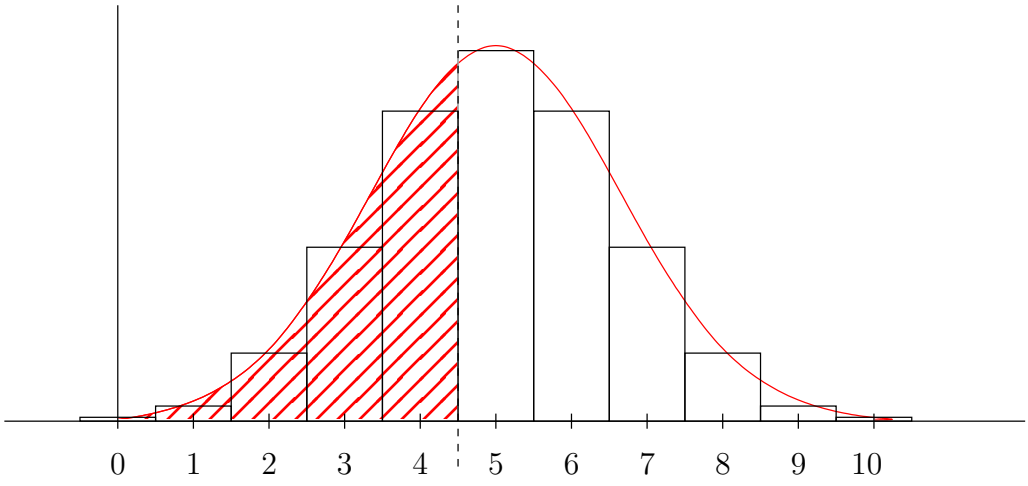


Ici l'espérance et l'écart-type de X valent respectivement $\mu = np$ et $\sigma = \sqrt{np(1-p)}$. Si on introduit une variable aléatoire $Z \sim \mathcal{N}(\mu, \sigma)$ on a par définition de la loi normale, pour tout réel x :

$$P(Z \leq x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

Autrement dit $P(Z \leq x)$ est l'intégrale sur $]-\infty, x]$ de la fonction $f_Z(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$.

Reportée sur le même graphique, cette fonction a l'allure suivante :



On sait que l'intégrale de f_Z sur $]-\infty, x]$ est aussi une mesure de l'aire sous la courbe de f_Z à gauche de x. Sur cette figure on voit que cette intégrale de $-\infty$ à $4,5$ fournit une meilleure approximation de $P(X \leq 4)$ que l'intégrale de $-\infty$ à 4.

Plus généralement l'intégrale de $k-0,5$ à $k+0,5$ fournit une bonne approximation de l'aire du rectangle dont la base est centrée sur k. Autrement dit :

$$P(X = k) \approx \int_{k-0,5}^{k+0,5} f_Z(t)dt = P(k - 0,5 \leq Z \leq k + 0,5)$$

Cela vaut aussi pour tous les entiers $l \leq k$, d'où en sommant sur tous ces entiers :

$$P(X \leq k) \approx \int_{-\infty}^{k+0,5} f_Z(t)dt = P(Z \leq k + 0,5)$$

On retrouve ainsi la formule utilisée pour la théorème d'approximation, avec la correction de continuité. L'approximation serait encore meilleure si X suivait une loi $\mathcal{B}(n, p)$ avec une plus grande valeur de n. Pour des valeurs de n vraiment grandes, la correction de continuité devient même superflue comme le prédit le théorème de De Moivre et Laplace.