

# Cours de Statistiques

R. Baltensperger et M.-A. Schnetzer

2017–2018

# Chapitre 1

## Statistique descriptive

En introduction de ce cours, il semble intéressant de tenter une définition de la statistique. Ceci n'est pas une chose facile. Selon Wikipedia, il existe entre 100 et 120 définitions de ce terme, au singulier ou au pluriel. De manière générale, nous pouvons tout de même dire que la statistique est un outil d'aide à la décision pour tirer des conclusions en présence de variabilité. Après la décision, l'ingénieur pourra résoudre des problèmes et concevoir des produits et processus. En statistique, les méthodes servent à décrire et comprendre cette variabilité. On parle aussi de science des données. Il y a variabilité lorsque plusieurs jeux de données ne donnent pas le même résultat (jet d'un dé ou sondage, par exemple).

Tentons une définition tout de même: la statistique est l'ensemble des méthodes qui ont pour objet la collecte, le traitement et l'interprétation des données observées relatives à un groupe d'individus ou d'unités, avec comme but de reconnaître des structures dans les observations faites.

### 1.1 Concepts et vocabulaire

En présence de quantités de données, il est important de savoir quel état nous avons atteint aujourd'hui, de le comparer à hier, d'estimer les états futurs et de faire des énoncés sur les effets de mesures mises en place. Tout ceci doit pouvoir être réalisé sur la base d'ensembles de données importants. Le rôle de la statistique descriptive est justement de résumer et présenter des données que l'on n'arrive pas à synthétiser sans outil graphique. Par exemple, une moyenne générale sur un bulletin de notes est issue de nombreuses évaluations et prestations. La statistique descriptive explique comment l'on passe des résultats individuels à la moyenne finale.

Un grand ensemble de données utilisé pour calculer des caractéristiques (une moyenne, par exemple) ne parle que pour lui-même. On ne peut a priori pas tirer de conclusions sur le cas général. Historiquement, ce problème s'est posé dès que l'on a dû travailler sur un échantillon plutôt que sur une enquête exhaustive. On parle dans ce cas d'analyse sur un échantillon et nous passons de valeurs exactes à des estimations. C'est là que le concept de hasard entre en jeu (avec la théorie mathématique du calcul des probabilités): deux personnes tirent chacune un échantillon du même

ensemble de données et en calculent la moyenne. Elles arriveront à des résultats différents. L'inférence statistique nous permettra de faire des affirmations sur le tout à partir de l'échantillon (exemples des ordinateurs défectueux ou d'un sondage des votations).

- **Population** : ensemble des individus, objets ou faits sur lesquels porte une étude. Taille  $N$ .
- **Echantillon** : sous-ensemble de la population. Taille  $n (\leq N)$ .
- **Unité statistique** : élément (individu, objet ou fait) de la population.
- **Recensement** : étude statistique portant sur toute la population.
- **Sondage** : étude statistique portant sur un échantillon.
- **Variable statistique** : caractéristique étudiée sur une unité statistique.

Une **variable statistique** peut être **qualitative nominale** ou **ordinale** ou **quantitative discrète** ou **continue**.

## 1.2 Distribution des fréquences uni-dimensionnelle

Les données (cas discret) sont représentées selon les fréquences d'apparition, absolues ou relatives. On obtient ainsi un diagramme en bâtons. Pour illustrer cela, un exemple sera pris en classe.

Dans le cas d'une variable statistique continue, avec une précision suffisamment grande des mesures, chaque observation peut être différente. Dans ce cas, on construit des classes du type "de... à ...". On gagne ainsi en visibilité mais il y a une perte d'information puisqu'à l'intérieur d'une classe, plus rien n'est connu. Un exemple sera donné dans le cadre du cours (durée des appels téléphoniques). La représentation graphique de la distribution des fréquences par classe est un histogramme.

## 1.3 Fréquences cumulées et fonction de répartition empirique

Souvent, on s'intéresse à des questions du type: quelle proportion des observations si situe en-dessous ou en-dessus d'une certaine valeur, ou entre deux valeurs? Pour quelle valeur  $x_p$  peut-on dire: la proportion  $p$  des mesures est inférieure (ou égale) à  $x_p$ , les autres étant supérieure à  $x_p$ ? Pour répondre à ce genre de questions, on peut construire le tableau des fréquences cumulées (absolues ou relatives). Pour cela, il faut que les observations soient des nombres (variable qualitative ordinale ou quantitative). Cette façon de faire justifie la définition de  $F(x)$ , la fonction de répartition empirique d'une variable statistique. Le terme 'empirique' signifie: obtenu par les observations, par opposition à "théorique", obtenu par calcul, réflexion ou modélisation.

## 1.4 Mesures de tendance centrale et de dispersion

### 1.4.1 La moyenne

La **moyenne** est la plus connue des mesures de tendance centrale. Pour la **population**, on la calcule ainsi:

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

et pour un **échantillon**:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

En présence d'un tableau de fréquences avec des classes (histogramme), on remplace l'ensemble des valeurs d'une classe par le milieu de la classe. La moyenne est donc dans ce cas une approximation:  $\bar{x} \approx \frac{1}{n} \sum_{j=1}^m m_j h_j$ , où  $m$  désigne le nombre de classe et  $m_j$  le milieu de la classe  $j$ .

La moyenne  $\bar{x}$  d'un ensemble de valeurs  $x_1, \dots, x_n$  est la valeur de  $\lambda$  qui minimise la fonction  $f(\lambda) = \sum_{i=1}^n (x_i - \lambda)^2$ .

**Exemple 1** Calculer la moyenne de 6 salaires mensuels donnés par  $x_1 = 3500.-$ ,  $x_2 = 4200.-$ ,  $x_3 = 4600.-$ ,  $x_4 = 5000.-$ ,  $x_5 = 6200.-$  et  $x_6 = 36500.-$  (10000.-).

### 1.4.2 La médiane

La **médiane** est une valeur telle que 50% des données lui sont inférieures lorsque les données sont triées en ordre croissant. La médiane n'est pas affectée par les valeurs extrêmes de la population. Pour la **population**, on la calcule ainsi:

$$\tilde{\mu} = \begin{cases} x_{\frac{N+1}{2}}, & N \text{ impair}, \\ \frac{1}{2}(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}), & N \text{ pair}, \end{cases}$$

et pour un **échantillon**:

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & n \text{ impair}, \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & n \text{ pair}. \end{cases}$$

La médiane  $\tilde{x}$  d'un ensemble de valeurs  $x_1, \dots, x_n$  est la valeur de  $\lambda$  qui minimise la fonction  $f(\lambda) = \sum_{i=1}^n |x_i - \lambda|$ .

En présence d'un tableau de fréquences avec des classes (histogramme), on utilise la fonction de répartition empirique pour calculer la médiane. Un exemple sera pris en classe (taux des personnes professionnellement actives en fonction de l'âge en Bavière).

**Exemple 2** Calculer la médiane des 6 salaires de l'exemple précédent (4800.-).

### 1.4.3 Variance empirique( $s^2$ ) et écart-type ( $s$ )

On aimerait savoir comment les valeurs d'un échantillon s'écartent de la moyenne  $\bar{x}$ . Pour ce faire, on calcule la moyenne des écarts quadratiques. Pour un **échantillon**:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right), \quad s = \sqrt{s^2}.$$

En présence d'un tableau de fréquences avec des classes (histogramme), chaque valeur d'une classe est estimée par la valeur située au milieu de la classe.

**Exemple 3** Calculer l'écart-type de l'exemple précédent ( $\sigma \approx 11879$ ,  $s \approx 13013$ ).

## 1.5 La boîte à moustaches (Boxplot)

La représentation graphique de la boîte à moustaches, pour la lire et l'interpréter, il est nécessaire d'en connaître la construction.

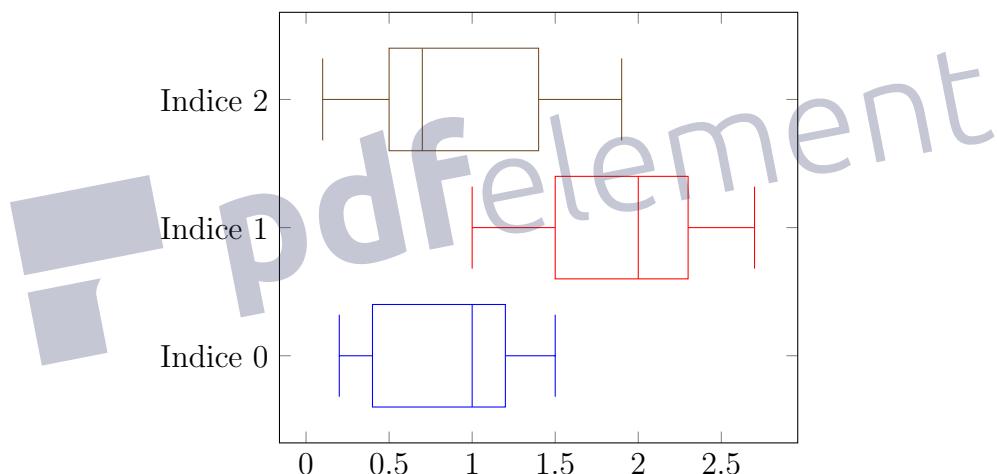


Figure 1.1: Exemple de boîte à moustaches (de Tukey)

La boîte à moustaches utilise cinq valeurs qui résument des données

1. le minimum;
2. les trois quartiles  $Q_1$ ,  $Q_2 = \tilde{x}$  (la médiane),  $Q_3$ ;
3. le maximum.

### 1.5.1 Les quartiles et l'écart interquartile

#### A) Les quartiles

**Exemple 4** On a 9 valeurs ordonnées

$$1, 3, 4, 5, 6, 7, 9, 10, 15, \quad n = 9$$

1. La **médiane**  $\tilde{x} = Q_2$  partage la suite (ordonnée) de nombres en deux groupes d'effectifs égaux. Ici,  $\tilde{x} = Q_2 = 6$ .
2. Le **quartile**  $Q_1$  repartage le groupe **du bas** en deux groupes d'effectifs égaux. Cela donne dans notre exemple  $Q_1 = 4$ .
3. Le **quartile**  $Q_3$  repartage le groupe **du haut** en deux groupes d'effectifs égaux. Cela donne dans notre exemple  $Q_3 = 9$ .

Selon que  $n$  est pair ou impair, on procédera différemment pour évaluer les quartiles.

### Procédure

1. Classer les données dans l'ordre croissant.
2. Diviser les données en deux groupes des tailles égales. On obtient alors le groupe du bas et le groupe du haut. Chacun des groupes contient 50% des observations.
  - Si  $n$  est **pair**, on calcule la médiane  $Q_2$  comme la moyenne des deux points du milieu.
  - Si  $n$  est **impair**, il faut reproduire la valeur du point milieu dans les deux groupes.
3. Calculer à nouveau la médiane du groupe du bas. On obtient alors le quartile  $Q_1$  qui correspond à 25% des observations.
4. Calculer à nouveau la médiane du groupe du haut. On obtient alors le quartile  $Q_3$  qui correspond à 75% des observations.

### Remarques

1. Les quartiles sont des **indicateurs de positionnement**.
2. Les quartiles divisent les données en 4 parts égales: 25%, 50%, 75% et 100%.
3. Pour  $0 < \alpha < 1$ , on définit plus généralement le **quantile**  $q_\alpha$ : le **quantile** d'ordre  $\alpha$  est la valeur  $q_\alpha$  qui partage l'ensemble des données (observations) en deux parties
  - (a) l'une formée des  $\alpha \cdot 100\%$  valeurs plus petites que  $q_\alpha$ .
  - (b) l'autre formée des  $(1 - \alpha) \cdot 100\%$  valeurs plus grandes que  $q_\alpha$ .

Si la fonction de répartition empirique  $F(x)$  est continue et monotone croissante (strictement),  $q_\alpha$  peut être obtenu par la relation  $F(q_\alpha) = \alpha$ .

### Calcul d'un quantile à partir d'un échantillon ordonné

1. Ordonner les données en ordre croissant ( $n$  données). On obtient alors la suite

$$x_1, \quad x_2, \quad x_3, \dots, \quad x_n$$

2. Calculer le rang du quantile, c'est-à-dire la position qu'occupe le quantile dans la liste ordonnée

$$\text{rang} = r_\alpha = \alpha(n - 1) + 1.$$

- Si  $r_\alpha$  est entier, alors  $q_\alpha = x_{r_\alpha}$
- Si  $r_\alpha$  n'est pas entier, alors

$$q_\alpha = x_{\lfloor r_\alpha \rfloor} + (r_\alpha - \lfloor r_\alpha \rfloor)(x_{\lceil r_\alpha \rceil} - x_{\lfloor r_\alpha \rfloor})$$

$\lfloor r_\alpha \rfloor$ : partie entière inférieure de  $r_\alpha$

$\lceil r_\alpha \rceil$ : partie entière supérieure de  $r_\alpha$

**Exemple 5** Ci-dessous, on trouve le temps d'attente à un guichet de poste

$$14, 4, 10, 7, 12, 15, 1, 3, 17, 8, 12 \quad (n = 11)$$

On veut calculer  $q_\alpha$  pour  $\alpha = 20\%$  et  $\alpha = 56\%$ .

### B) Ecart interquartile

C'est un indicateur de dispersion (comme l'est l'écart-type). Cela indique dans quelle mesure les mesures sont groupées autour du centre ou, au contraire, elles s'en écartent.

1. L'**étendue** est la différence entre la valeur maximum et la valeur minimum des données (observations) ordonnées.
2. L'**étendue interquartile** ou **écart interquartile** (EIQ) correspond à 50% des effectifs situés dans la partie centrale de la distribution.

$$EIQ = Q_3 - Q_1.$$

### C) Lecture d'une boîte à moustaches

On repère sur la boîte à moustaches d'une variable

1. l'échelle des valeurs de la variable, située sur l'axe vertical (ou horizontal);
2. la valeur du premier quartile  $Q_1$  (25% des effectifs) correspondant au trait inférieur de la boîte;
3. la valeur du deuxième quartile  $Q_2$  (50% des effectifs) représentée par un trait horizontal à l'intérieur de la boîte;
4. la valeur du troisième quartile  $Q_3$  (75% des effectifs) correspondant au trait supérieur de la boîte;
5. les deux "moustaches" inférieure et supérieure représentées par des traits ou des crochets de part et d'autre de la boîte. Ce deux moustaches délimitent les valeurs dites **adjacentes** qui sont déterminées à partir de l'EIQ;

6. Les valeurs dites extrêmes, atypiques, exceptionnelles (outliers) situées au-delà des valeurs adjacentes sont individualisées. Elles sont représentées par des marqueurs (carré, étoile, triangle, . . . ).

#### D) Délimitation des longueurs des moustaches (valeurs adjacentes)

1. L'extrémité de la moustache **inférieure** est la valeur minimum dans les données qui est supérieure à la valeur frontière basse

$$Q_1 - 1.5 \cdot EIQ.$$

2. L'extrémité de la moustache **supérieure** est la valeur maximum dans les données qui est inférieure à la valeur frontière haute

$$Q_3 + 1.5 \cdot EIQ.$$

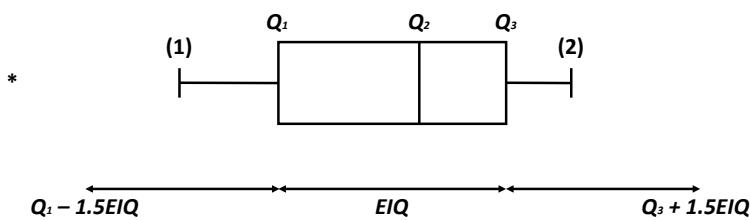


Figure 1.2: Boxplot en général. (1): valeur adjacente de la moustache inférieure. (2): valeur adjacente de la moustache supérieure. \*: valeur atypique.

#### Remarque

L'intervalle  $[Q_1 - 1.5 \cdot EIQ; Q_3 + 1.5 \cdot EIQ]$  contient à peu près 99% des données.

**Exemple 6** Ci-dessous, on trouve le nombre de tasses de café bues en une journée à la terrasse d'un bistro ( $n = 20$ )

11, 13, 18, 20, 21, 23, 25, 25, 27, 28, 31, 34, 35, 41, 42, 43, 44, 46, 54, 93

Faire le boxplot correspondant.

#### Interprétation du boxplot

1. La **médiane** nous renseigne sur le milieu des données.
2. Les **largeurs** des deux parties de la boîte rendent compte de la dispersion des valeurs situées au centre des observations (la boîte contient 50% de l'ensemble des observations; 25% à gauche et 25% à droite de la médiane).
3. La **longueur** des moustaches renseigne sur la dispersion des valeurs situées au début de la série ordonnée (les valeurs les plus petites correspondant à 25% des observations) ou à la fin de celle-ci (les valeurs les plus grandes correspondant aussi à 25% des observations).
4. de façon générale, la boîte et les moustaches seront d'autant plus étendues que la dispersion de la série est grande.

## 1.6 Distribution des fréquences à deux dimensions

## 1.7 Données multivariées: régression et corrélation

Lorsqu'une ou plusieurs variables sont examinées pour un échantillon ou une population, chacune de ces variables est d'abord analysée individuellement avec des tableaux, des graphes et des mesures comme vu dans les sections précédentes.

Par la suite, il se peut que l'on cherche à déterminer s'il existe un lien entre deux (ou plusieurs) de ces variables et s'il est possible de l'exprimer mathématiquement. Le lien peut être linéaire, quadratique exponentiel ... On va chercher à déterminer le modèle "au plus près".

**Exemple 7** *Dans le tableau ci-dessous, on trouve le nombre de personnes vivant dans un logement ( $x$ ) et les dépenses effectuées dans des épiceries par semaine ( $y$ ). On peut faire une représentation graphique des données. Il semble qu'il existe une*

$x$	2	2	3	4	1	5
$y$	95	120	136	201	72	261

Tableau 1.1: Personnes vivant dans un logement et les dépenses par semaine.

*relation linéaire entre les données.*

Il s'agit ici de trouver une éventuelle relation entre les deux variables (aléatoires)  $X$  et  $Y$ . Une première approche est la recherche d'une relation affine entre  $x$  et  $y$ , c'est-à-dire une relation de la forme  $Y = aX + b$ . Supposons que l'on ait observé  $n$  fois le couple  $(X, Y)$  et que les valeurs obtenues soient

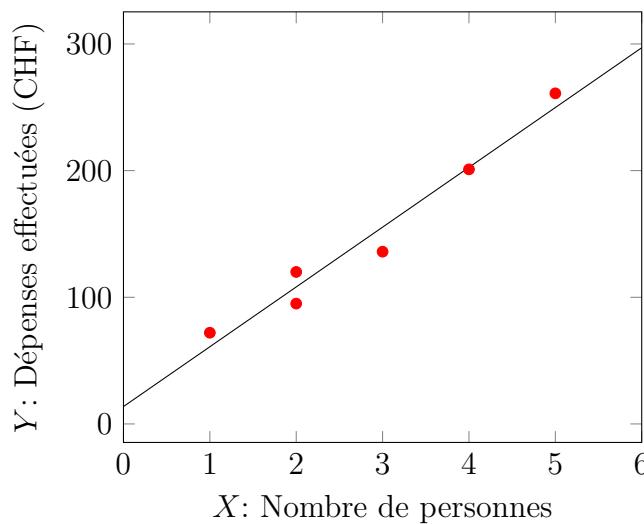


Figure 1.3: Couples de points  $(X, Y)$  et droite optimale.

Les couples précédents représentés dans le plan  $(X, Y)$  ne seront en général pas alignés sur une droite. Nous allons donc écrire la relation en corrigeant avec une erreur stochastique  $\varepsilon_i$

$$Y_i = aX_i + b + \varepsilon_i, \quad 1 \leq i \leq n.$$

Nous allons chercher  $a$  et  $b$  qui approchent “au mieux” le nuage de points définis par les couples observés. Comme définir “au mieux”?

D’habitude on travaille avec l'**erreur quadratique totale** définie par

$$E(a, b) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (aX_i + b - Y_i)^2.$$

Pour trouver les valeurs de  $a$  et de  $b$  qui minimisent  $E(a, b)$ , on impose

$$\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0.$$

La seconde condition nous fournit

$$\bar{Y} = a\bar{X} + b.$$

Ainsi, la droite optimale passe par le point (moyen)  $(\bar{X}, \bar{Y})$ . Cela implique que

$$b_{opt} = \bar{Y} - a\bar{X}.$$

En remplaçant  $b_{opt}$  dans  $E(a, b)$  et en dérivant l’expression obtenue par rapport à  $a$ , on trouve

$$a_{opt} = \frac{C_{xy}}{S_x^2}$$

avec

$S_x^2$  : variance empirique de  $X$ .

$C_{xy}$  : covariance empirique de l’échantillon.

La covariance empirique se calcule ainsi

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) = \frac{1}{n-1} \left[ \sum_{i=1}^n X_i Y_i - n \cdot \bar{X} \bar{Y} \right].$$

L’équation de la droite optimale (celle qui minimise l’erreur quadratique) est donnée par

$$y - \bar{Y} = \frac{C_{xy}}{S_x^2}(x - \bar{X}).$$

On peut montrer que l’erreur minimale commise si l’on remplace le nuage de points observés par la droite optimale est donnée par

$$E_{\min} = (n-1)S_y^2 \left( 1 - \left( \frac{C_{xy}}{S_x S_y} \right)^2 \right).$$

Comme  $E_{\min} \geq 0$ , on a

$$-1 \leq \frac{C_{xy}}{S_x S_y} \leq 1.$$

On définit  $r_{xy}$  comme étant le **coefficient de corrélation empirique de l’échantillon**

$$r_{xy} = \frac{C_{xy}}{S_x S_y}.$$

$r_{xy}$  donne une information sur l’intensité de la relation entre les deux variables.

### 1.7.1 Constats

1.  $E_{\min} = 0$  implique que  $r_{xy} = \pm 1$ . Ce qui veut dire que les points sont alignés sur la droite. Si  $r_{xy} = 1$ , la pente est positive, si  $r_{xy} = -1$ , la pente est négative.
2. Plus  $|r_{xy}|$  est proche de 1 plus  $E_{\min}$  est proche de 0.
3. Plus  $|r_{xy}|$  est proche de 0 plus  $E_{\min}$  est grande.
4.  $E_{\min}$  est maximale lorsque  $r_{xy} = 0$ . La droite ne représente pas les données de manière satisfaisante. On ne peut par contre pas dire que les variables  $X$  et  $Y$  ne sont liées par aucune relation.
5. Lorsque  $|r_{xy}|$  est proche de 1 cela n'implique pas forcément un lien causal entre les variables  $X$  et  $Y$ .

**Exemple 8** Dans le tableau ci-dessous, on trouve l'évolution de la population humaine et du nombre de couples de cigognes de la ville d'Oldenburg entre 1930 et 1936.

	1930	1931	1932	1933	1934	1935	1936
# couples de cigognes	132	142	166	188	240	250	252
Habitants	55'400	55'400	65'000	67'700	69'800	72'300	76'000

Tableau 1.2: Population et couple de cigognes de la ville d'Oldenburg dans les années 1930.

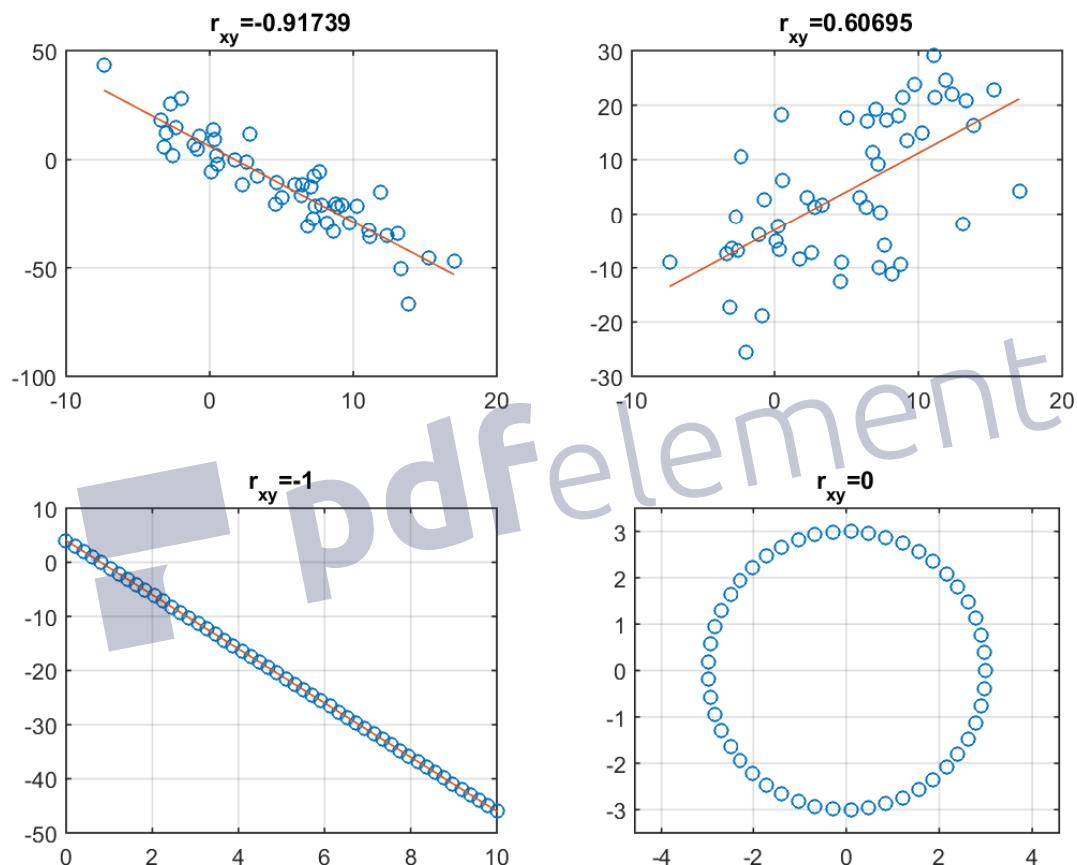
#### Remarque

Dans Excel, c'est le **coefficient de détermination**  $r^2$  qui est calculé. Celui-ci s'obtient par

$$r^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \quad \text{avec} \quad \hat{Y}_i = f(X_i).$$

On peut montrer que dans le cas linéaire

$$r^2 = r_{xy}^2.$$

Figure 1.4: Nuages de points et  $r_{xy}$  correspondants.

# Chapitre 2

## Probabilités

Pour comprendre les idées et concepts de la statistique, il est important de bien comprendre les idées et les modèles de la théorie des probabilités. Ce cours va donc aborder cette branche des mathématiques, née dans un échange de lettres entre Fermat et Pascal (1654).

Le calcul des probabilités s'occupe de phénomènes aléatoires, c'est-à-dire des phénomènes qui, lorsqu'ils sont observés dans des conditions déterminées ne mènent pas toujours à la même issue. Même si ces phénomènes ont des issues variées dépendant du hasard, on observe une certaine régularité statistique. En ingénierie, le bruit qui perturbe un signal est modélisé par des probabilités. En physique statistique, on pense que la probabilité est partie intégrante de ses fondements.

### 2.1 Fondamentaux de la combinatoire

Notation:  $n! = n \cdot (n - 1) \cdot (n - 2) \cdot (n - 3) \cdots 3 \cdot 2 \cdot 1$ ,  $0! = 1$  (par définition)

#### Arrangements

**Exemple 9** Une urne contient 5 boules numérotées de 1 à 5. On en tire deux successivement sans remise. Combien y a-t-il de tirages possibles, sachant que l'ordre des tirages compte?

Plus généralement: Une urne contient  $n$  boules numérotées de 1 à  $n$ . On en tire  $r$  successivement sans remise. Combien y a-t-il de tirages possibles, sachant que l'ordre des tirages compte?

Arrangement de  $r$  objets parmi  $n$  objets (l'ordre est important):

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

**Exemple 10** Une classe est composée de 18 personnes (2 filles et 16 garçons).

On veut faire un conseil de classe qui contient un président, un vice-président et un caissier. Combien de conseils de classe différents peut-on former ?

$$A_3^{18} = \frac{18!}{(18-3)!} = 4896 \quad (18 \cdot 17 \cdot 16 = 4896).$$

## Permutations

On appelle une permutation de  $n$  objets, une disposition ordonnée de  $n$  objets distincts choisis parmi  $n$  objets.

$$P_n = n!$$

**Exemple 11** Lors d'une exposition, un peintre dispose de 5 emplacements sur un mur pour accrocher 5 des tableaux qu'il a avec lui. De combien de façons différentes peut-il accrocher les tableaux ?

$$P_5 = 5! = 120 \quad (5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120)$$

Et s'il avait pris 8 tableaux ?

$$A_5^8 = \frac{8!}{(8-5)!} = 6720$$

## Combinaisons

On appelle combinaison de  $r$  parmi  $n$ , un sous-ensemble de  $r$  objets choisis parmi  $n$  objets distincts (arrangement où l'ordre n'est plus important).

$$C_r^n = \frac{n!}{(n-r)!r!} = \binom{n}{r}.$$

**Exemple 12** Un département de professeur est composé de 15 personnes. 3 personnes doivent être choisies pour représenter le département à l'assemblée des professeurs de l'école.

1. Combien de trios différents peut-on former si les rôles des personnes dans le trio sont différents ?

2. Combien de trios différents peut-on former si les rôles des personnes dans le trio sont identiques (indifférents) ?

$$C_3^{15} = \frac{15!}{(15-3)!3!} = 455$$

3. Marc fait partie du département.

Quelle est la probabilité que Marc fasse partie du trio dans la situation où les rôles sont identiques ?

4. Claude fait aussi partie du département.

Quelle est la probabilité que Claude fasse partie du trio et que Marc ne fasse pas partie du trio, dans la situation où les rôles sont différents (Claude veut un rôle précis dans le trio) ?

## 2.2 Hasard et événements

Par “hasard”, on comprend l’ensemble de tous les facteurs qui ont une influence sur le résultat d’une expérience, mais que nous ne connaissons pas et que nous ne pouvons pas contrôler. S’il y a de telles influences, on parle d’une **expérience aléatoire**. Elle est menée selon des règles précises, fixées à l’avance et est reproduisible aussi souvent que nécessaire.

**Exemple 13** *La distance de freinage d’urgence d’un véhicule, comme dépendant de la vitesse de celui-ci, est mesurée à plusieurs reprises. Pourquoi s’agit-il ici d’une expérience aléatoire?*

La réalisation d’une seule expérience aléatoire est appelé un **essai**, son résultat est appelé issue ou événement élémentaire. L’ensemble de toutes les issues possibles s’appelle espace des événements élémentaires ou **univers**. On désigne souvent cet ensemble par la lettre grecque  $\Omega$  (oméga). Ces événements élémentaires ne sont pas décomposables et jouent en quelque sorte le rôle d’“atomes”.

**Exemple 14** *On lance un dé à 6 faces. Décrire l’univers dans cette expérience, i.e. l’ensemble contenant les événements élémentaires.*

Tout sous-ensemble  $A$  de l’ensemble fondamental  $\Omega$  des résultats d’une expérience aléatoire porte le nom d’**événement**. On dit qu’un événement  $A$  se réalise lorsque le résultat observé de l’expérience est un des résultats appartenant au sous-ensemble  $A$  de  $\Omega$ .

Les événements peuvent être décrits par des ensembles ou par des mots. Les événements  $\emptyset$  et  $\Omega$  s’appellent respectivement événement **impossible** et événement **certain**. Question: combien d’événements existe-t-il dans l’exemple du jet d’un dé? Plus généralement, combien d’événements existe-t-il étant donné un espace des événements élémentaires de  $n$  éléments?

Comme avec les nombres, où l’on génère d’autres nombres par les opérations  $+$ ,  $-$ ,  $\cdot$  et  $/$ , on connaît des opérations de “calcul” avec les événements (qui ne sont finalement rien d’autre que des ensembles) qui forment les objets de **l’algèbre des événements**. Les opérations utilisées sont intersection ( $\cap$ ), réunion ( $\cup$ ), différence ( $\backslash$ ) et complément ( $\bar{A}$  est le complément de l’événement  $A$ ).

Reprendre l’exemple précédent et décrire les événements  $A$ : “le nombre obtenu est divisible par 3” et  $B$ : “le nombre obtenu est au moins 3”.

## 2.3 Probabilité et théorème de Laplace

Lorsqu’on jette une pièce de monnaie, l’issue de l’expérience, c’est-à-dire l’apparition de *pile* ou de *face* n’est pas prévisible. Si on répète cette épreuve (ou expérience aléatoire)  $n$  fois et si l’on compte le nombre de fois  $k$  où *face* apparaît, on constate que la fréquence (relative)  $\frac{k}{n}$  du nombre d’apparitions de face dans la série d’épreuves est voisine de 0.5. On suppose bien sûr que la pièce est équilibrée, c’est-à-dire que chaque résultat a une chance sur deux de se produire.

Buffon, au 18ème siècle, à réalisé 4040 jets et a obtenu 2048 fois face (fréquence des faces : 0.5069). Pearson au 20ème siècle a réalisé 24000 jets et à obtenu 12012 fois face (fréquence des faces : 0.5005).

On dit que lorsqu'on jette une pièce de monnaie, on a une probabilité de  $\frac{1}{2}$  de voir l'événement “face” se produire au cours de cette épreuve. La probabilité d'un événement est donc la limite vers laquelle tend la fréquence de cet événement lorsqu'on répète  $n$  fois l'épreuve et lorsque  $n \rightarrow \infty$ . C'est un modèle mathématique qui repose sur des axiomes, par exemple sur le fait que la pièce ne va pas s'arrêter sur sa tranche. Notons qu'un modèle de peut pas être prouvé comme étant correct; au mieux, on peut le trouver raisonnablement consistant et reconnaître qu'il ne contredit pas nos croyances de la réalité.

Nous formulons une hypothèse du modèle qui stipule que chaque événement  $A$  possède un nombre qui lui est attaché et qui est sa probabilité  $p$  d'apparition, avec  $p \in [0; 1]$ . Ce  $p$  est exactement la valeur vers laquelle la probabilité relative  $f_n(A) = \frac{k}{n}$  converge lorsque  $n \rightarrow \infty$ .

**Exemple 15** Considérons l'expérience aléatoire qui consiste à lancer un dé. L'ensemble fondamental des résultats  $\Omega$  de cette expérience aléatoire peut être donné par  $\Omega = \{1, 2, 3, 4, 5, 6\}$ . L'événement  $A$  défini par “le résultat du lancer est un nombre pair” correspond au sous-ensemble  $A = \{2, 4, 6\} \subset \Omega$ . Pour que cet événement se réalise, il faut que le résultat du lancer soit un des trois nombres 2, 4 ou 6.

### Compatibilité de deux événements

Deux événements  $A$  et  $B$  sont **compatibles** si les deux peuvent se réaliser en même temps, c'est-à-dire si  $A \cap B \neq \emptyset$ . Deux événements  $A$  et  $B$  sont **incompatibles** si  $A \cap B = \emptyset$ .

Soit  $A$  un événement possible ( $A \neq \emptyset$ ), sans être certain ( $A \neq \Omega$ ). L'événement  $A$  ne se réalise pas si et seulement si sa négation (ou son complément)  $\bar{A}$  se réalise.

Pour pouvoir calculer avec les probabilités, on a besoin d'un ensemble d'hypothèses de base, non discutables, appelés “axiomes”, que l'on ne prouve pas, mais que l'on suppose comme valables par expérience. Ils ont été proposés par Kolmogorov en 1933 et leur plausibilité est facile à comprendre lorsque l'on remplace “probabilité” par “fréquence relatives”. Ces axiomes sont au nombre de trois:

1. Pour chaque événement  $A$ , on a :  $0 \leq P(A) \leq 1$ ;
2. L'univers a comme probabilité 1:  $P(\Omega) = 1$ ;
3. Si  $A_1, A_2, \dots$  sont des événements disjoints deux à deux, alors on a  $P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$

### Probabilité de réalisation d'un événement lorsque $\Omega$ contient $n$ résultats

Considérons une expérience aléatoire dont l'univers contient un nombre fini  $n$  de résultats, c'est-à-dire  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ . À chacun des  $n$  événements simples (événements ne correspondant qu'à un seul résultat) pour cette expérience aléatoire correspond une probabilité de réalisation; c'est-à-dire à chacun des événements  $\{\omega_i\}$

correspond une probabilité de réalisation  $P(\{\omega_i\})$ , ( $i = 1, 2, \dots, n$ ). Ces  $n$  probabilités sont telles que : pour  $i = 1, 2, \dots, n$ :

$$0 \leq P(\{\omega_i\}) \leq 1 \quad \text{et} \quad \sum_i P(\{\omega_i\}) = 1.$$

Pour un événement  $A \subset \Omega$ , la probabilité de réalisation de  $A$  se note  $P(A)$  et, de façon générale,

$$P(A) = \sum_{i=1}^n P(\{\omega_i\}). \quad (2.1)$$

### Simplification du calcul lorsqu'il y a équiprobabilité

Si  $\Omega$  contient un nombre fini  $n$  d'événements élémentaires et si chacun de ceux-ci a même probabilité d'apparition, alors

$$P(A) = \frac{\#A}{\#\Omega} = \frac{\#A}{n}, \quad (2.2)$$

où le symbole  $\#$  désigne la **cardinalité** (le nombre d'éléments) de l'ensemble qui le suit.

### Exemple 16 Situation où les éléments de $\Omega$ ne sont pas équiprobables

*Une urne contient 12 boules de forme et de masse identiques. Une de celles-ci est numérotée 2, six sont numérotées 3, trois sont numérotées 4 et les deux dernières sont numérotées 5. Considérons l'expérience aléatoire qui consiste à prélever au hasard une boule de l'urne et à noter son numéro. Définissons l'événement  $A$  par “le numéro de la boule prélevée est un nombre pair”.*

Soit  $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\} = \{2, 3, 4, 5\}$  et  $A = \{2, 4\}$ . Pour établir les probabilités de réalisation des 4 événements simples associables à  $\Omega$ , on peut supposer que chacune des 12 boules de l'urne a la même probabilité d'être prélevée. On obtient alors que :  $P(\{\omega_1\}) = P(\{2\}) = 1/12$ ,  $P(\{\omega_2\}) = P(\{3\}) = 6/12$ ,  $P(\{\omega_3\}) = P(\{4\}) = 3/12$ ,  $P(\{\omega_4\}) = P(\{5\}) = 2/12$ . Pour calculer la probabilité de réalisation de l'événement  $A$ , on doit utiliser la formule générale (2.1), car il n'y a pas équiprobabilité des éléments de  $\Omega$ . On obtient :

$$P(A) = P(\{2, 4\}) = P(\{2\}) + P(\{4\}) = 1/12 + 3/12 = 4/12.$$

### Exemple 17 Situation où les éléments de $\Omega$ sont équiprobables

*Considérons la même expérience aléatoire que dans l'exemple précédent pour une urne qui contient trois boules numérotées 2, trois boules numérotées 3, trois boules numérotées 4 et trois boules numérotées 5.*

$$P(A) = \frac{\#A}{\#\Omega} = \frac{2}{4} = \frac{1}{2}.$$

## 2.4 Evénements indépendants et probabilité conditionnelle

### 2.4.1 Evénements indépendants

Deux événements  $A$  et  $B$  d'un univers  $U$  sont **indépendants** ( $B$  n'a aucune influence sur  $A$  et vice-versa) si :

- $P(A \cap B) = P(A) \cdot P(B)$ .

Trois événements  $A$ ,  $B$  et  $C$  d'un univers  $U$  sont indépendants si :

- $P(A \cap B) = P(A) \cdot P(B)$
- $P(A \cap C) = P(A) \cdot P(C)$
- $P(B \cap C) = P(B) \cdot P(C)$
- $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$

**Exemple 18** On jette une pièce de monnaie 2 fois de suite.

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

$A$  : face apparaît au premier jet

$B$  : pile apparaît au deuxième jet

$C$  : le même côté sort deux fois

$D$  : le nombre de face est inférieur à 2

On peut vérifier que  $A$  et  $B$  sont indépendants, que  $C$  et  $D$  ne sont pas indépendants.

**Exemple 19** On lance un dé.

$$U = \{1, 2, 3, 4, 5, 6\}$$

$$A = \{1, 6\}$$

$$B = \{2, 4, 6\}$$

$$A \cap B = \{6\}$$

On peut calculer  $P(A)$ ,  $P(B)$  et  $P(A \cap B)$ .

On nous précise **après avoir jeté le dé** qu'un nombre pair est sorti ( $B$  s'est produit).

Quelle est la probabilité que  $A$  se réalise sachant que  $B$  s'est réalisé ?

### 2.4.2 Probabilité conditionnelle

Si  $A$  et  $B$  sont deux événements d'un univers  $U$  et si  $P(B) > 0$  alors la **probabilité conditionnelle de  $A$  par  $B$**  est le nombre noté :

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

**Exemple 20** Sur 100000 garçons qui naissent, 92659 sont encore en vie à 50 ans et 71040 à 70 ans.

Quelle est la probabilité qu'un homme en vie à 50 ans le soit encore à 70 ans ?

On définit les événements :  $A$  : "être en vie à 70 ans" et  $B$  : "être en vie à 50 ans"

**Exemple 21** On choisit au hasard une famille parmi celles qui ont deux enfants.

- Quelle probabilité y-a-t-il que ce soit deux filles si l'on sait que l'aînée est une fille ?

Hypothèse :  $(G, G)$ ,  $(G, F)$ ,  $(F, G)$ ,  $(F, F)$  sont équiprobables

$$A = \{(F, F)\}$$

$$B = \{(F, G), (F, F)\}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{\frac{1}{4}}{\frac{2}{4}} = \frac{1}{2}$$

- Quelle est la probabilité d'avoir deux filles sachant qu'au moins un des deux enfants est une fille ?

$$A = \{(F, F)\}$$

$$B = \{(G, F), (F, G), (F, F)\}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

Si  $A, B, C, \dots$  sont des événements d'un univers  $U$ , on a :

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

$$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B)$$

$$P(A \cap B \cap C \cap D) = P(A) \cdot P(B|A) \cdot P(C|A \cap B) \cdot P(D|A \cap B \cap C)$$

...

La définition de la probabilité conditionnelle nous livre directement le théorème de multiplication pour des événements quelconques:

$$P(A \cap B) = P(B) \cdot P(A|B) = P(A) \cdot P(B|A).$$

Un moyen pratique de représenter graphiquement les probabilités conditionnelles est l'arbre de probabilités dont des exemples seront donnés en classe.

Le théorème des probabilités totales et le théorème de Bayes sont les étapes suivantes de cette section.

### 2.4.3 Epreuves successives dépendantes

Lors d'épreuves successives, la probabilité d'un chemin est donnée par la probabilité des branches qui forment ce chemin.

**Exemple 22** Une urne contient 3 boules blanches et 4 boules noires. On tire successivement 3 boules de l'urne sans remettre les boules dans l'urne (tirage sans remise).

Calculer la probabilité de tirer  $(n, b, b)$ .

**Exemple 23** Deux urnes identiques  $U_1$  et  $U_2$ .  $U_1$  contient 2 boules noires et 1 boule blanche.  $U_2$  contient 1 boule noire et 3 boules blanches.

On choisit une urne au hasard puis on extrait 2 boules (sans remise). On peut construire l'arbre des configurations possibles.

**Exemple 24** Il existe un test qui permet de diagnostiquer une certaine maladie et ayant les proportions suivantes: 95% des personnes malades réagissent positivement au test et 95% des personnes en bonne santé réagissent négativement au test. On admet que 5 personnes sur mille sont atteints de cette maladie.

Vous faites un test et il se révèle positif. Quelle est la probabilité que vous soyez réellement malade?

**Exemple 25** Le fameux paradoxe des anniversaires: Quel est le plus petit nombre de personnes qui peuvent être réunies dans une salle de telle manière à ce qu'il y ait plus de 50% de chance d'un anniversaire commun? (c'est-à-dire: deux anniversaires le même jour, mais pas forcément la même année). Indication: calculer la probabilité que tous les anniversaires soient différents pour 2, 3, ...,  $k$  personnes.

### 2.4.4 Epreuves successives indépendantes

Nous présentons ce concept à l'aide de deux exemples:

**Exemple 26** On lance un dé 3 fois de suite.

$$P((2, 4, 6)) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{216}$$

$$U = \{216 \text{ triplets}\}$$

**Exemple 27** On lance un dé 3 fois de suite. On cherche la probabilité de ne voir apparaître ni 3 ni 5.

## 2.5 Variables aléatoires (v. a.)

Souvent, les résultats d'une expérience aléatoire se laissent exprimer avec des nombres (exemples?)

Une fonction  $X : \Omega \rightarrow \mathbb{R}$  qui associe à chaque événement élémentaire de  $\Omega$  un nombre réel s'appelle une **variable aléatoire**, si tous les événements de la forme

$X = x$  et  $X \in I$  pour tout réel  $x$  et intervalle  $I$  possèdent une probabilité qui vérifie le système d'axiomes de Kolmogorov.

**Exemple 28** On jette une pièce de monnaie deux fois de suite. L'univers  $U$  est donné par  $U = \{(p, p), (p, f), (f, p), (f, f)\}$ .  $X$  indique le nombre de face qui se présentent lors de l'épreuve.

$$\begin{array}{ll} X : & \begin{array}{l} (p, p) \rightarrow 0 \\ (p, f) \rightarrow 1 \\ (f, p) \rightarrow 1 \\ (f, f) \rightarrow 2 \end{array} \end{array}$$

**Champ** de la variable aléatoire (ensemble des valeurs possibles)  $CH(X) = \{0, 1, 2\}$ .

**Exemple 29** Lors d'une campagne de publicité à la TV, M. Laprise a besoin de 5 personnes pour représenter son entreprise. Il choisit parmi les 3 femmes et les 6 hommes qui travaillent pour lui.

La compagnie de publicité s'intéresse à deux aspects 1. à l'âge des représentants et 2. à la proportion de femmes dans la délégation.

Variables aléatoires

1.  $X$  âge des représentants
2.  $Y$  proportion de femmes dans la délégation

$X : U \rightarrow [16, 65] = CH(X)$ . C'est une v. a. continue.

$Y : U \rightarrow \left\{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}\right\} = CH(Y)$ . C'est une v. a. discrète

### 2.5.1 Variable aléatoire discrète et sa distribution

On considère ici que  $X$  est une v. a. discrète.

La **loi de probabilité** de  $X$  (ou distribution de  $X$ ) est la fonction  $p$  (ou  $p_X$ ) qui associe à chaque élément  $x_i$  de  $CH(X)$  la probabilité de l'événement élémentaire  $\{x_i\}$ .

$$\begin{aligned} X : U &\rightarrow \{x_1, x_2, \dots, x_n\}, \quad CH(X) = \{x_1, x_2, \dots, x_n\} \\ p : CH(X) &\rightarrow [0, 1] \quad (\text{C'est une probabilité}) \\ x_i &\rightarrow p(x_i) = p(\{x_i\}) = P(X = x_i) \end{aligned}$$

On dira que l'image  $p(x_i)$  correspond à la probabilité que la v. a.  $X$  prenne la valeur  $x_i$ :

$$p(x_i) = p(\{x_i\}) = P(X = x_i).$$

Voici la distribution de probabilité de  $X$  si  $CH(X) = \{x_1, x_2, \dots, x_n\}$

$x_i$	$p(x_i)$
$x_1$	$P(X = x_1)$
$x_2$	$P(X = x_2)$
$\vdots$	$\vdots$
$x_n$	$P(X = x_n)$

On a les propriétés suivantes

1.  $0 \leq P(X = x_i) \leq 1$ ;
2.  $\sum_{i=1}^n P(X = x_i) = P(X = x_1) + P(X = x_2) + \dots + P(X = x_n) = 1$ .

La fonction  $F : \mathbb{R} \rightarrow [0; 1]$  avec  $F(x) = P(X \leq x)$  est appelée fonction de répartition de  $X$ .

**Exemple 30** On jette une pièce de monnaie deux fois de suite. L'univers est donné par  $U = \{(p, p), (p, f), (f, p), (f, f)\}$ .  $X$  indique le nombre de face qui se présentent lors de l'épreuve.  $CH(X) = \{0, 1, 2\}$ . On a  $X : U \rightarrow CH(X)$  et  $p : CH(X) \rightarrow [0, 1]$ . Distribution de probabilité de  $X$

$x_i$	$p(x_i)$
0	$P(X = 0) = 1/4$
1	$P(X = 1) = 2/4 = 1/2$
2	$P(X = 2) = 1/4$

**Exemple 31** Une boîte contient 12 billets numérotés de 1 à 12. On tire 3 billets de cette boîte à la suite, sans remise.  $X$  v. a. représentant le numéro le plus élevé des 3 billets tirés.  $CH(X) = \{3, 4, 5, \dots, 12\}$ . On cherche ici la distribution de probabilité de  $X$  et  $P(X > 7)$ .

### 2.5.2 Variable aléatoire continue et sa densité

Pour passer de la distribution d'une v.a. discrète à celle d'une v.a. continue, imaginons que les mesures (ou poids) de probabilité ne sont plus concentrés sur des points  $x_1, x_2, \dots$  mais sur tout l'axe réel (ou une partie de l'axe). Faisons l'analogie avec la physique en considérant un bâton de 1 kg posé le long d'un axe  $x$  et de densité variable. La densité au point  $x$  est alors la limite

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{\Delta M}{\Delta x},$$

où  $\Delta m$  est la masse de la tranche du bâton allant de  $x$  à  $x + \Delta x$ . Pour de petits  $\Delta x$ , la masse  $\Delta m$  est proche de  $\delta m \approx f(x)\Delta x$  ou, avec les différentielles:  $dm = f(x)dx$ . Pour obtenir la masse totale de 1, il faut additionner ou plus précisément intégrer tous les éléments  $dm$  pour  $x$  de  $-\infty$  à  $\infty$  (en supposant que le bâton est infini). Par analogie à la physique, on obtient ainsi la définition suivante:

Une fonction  $f : \mathbb{R} \rightarrow \mathbb{R}$  s'appelle densité ou densité de probabilité si les trois conditions suivantes sont remplies:

1.  $f(x) \geq 0$  pour tout  $x \in \mathbb{R}$ ,
2.  $f$  est continue sauf éventuellement en un nombre fini de points,
3.  $\int_{-\infty}^{\infty} f(x) dx = 1$ .

Nous posons  $F(x) = P(X \leq x)$  et appelons  $F(x)$  comme dans le cas discret la fonction de répartition de  $X$ . Nous pouvons maintenant préciser ce que l'on entend par variable aléatoire continue:

Une **variable aléatoire**  $X$  est appelée **continue** si sa fonction de répartition  $F(x) = P(X \leq x)$  peut s'écrire comme une intégrale de la forme

$$F(x) = \int_{-\infty}^x f(u) du,$$

avec une densité de probabilité  $f(x)$ .

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx.$$

La distribution **uniforme** continue, la distribution **exponentielle** et la distribution **normale** sont trois familles de distributions importantes pour les variables aléatoires continues.

## 2.6 Espérance et variance d'une distribution

Soit  $X$  une v. a. discrète ( $CH(X) = \{x_1, x_2, x_3, \dots, x_n\}$ ) et  $p$  sa loi de probabilité.

1. L'**espérance (mathématique)** d'une variable aléatoire discrète  $X$ , notée  $E(X)$  ( $= \mu = \mu_X$ ) est donnée par

$$E(X) = \sum_{x_i} x_i \cdot P(X = x_i).$$

2. La **variance** de  $X$  notée  $Var(X)$  ( $= \sigma^2$ ) est donnée par

$$Var(X) = \sum_{x_i} (x_i - E(X))^2 \cdot P(X = x_i).$$

3. L'**écart-type** de  $X$ , noté  $\sigma$  ( $= \sigma_X$ ) est donné par

$$\sigma = \sqrt{Var(X)}$$

**Exemple 32** On reprend l'exemple 31 et on calcule  $E(X)$ ,  $Var(X)$  et  $\sigma$ .

Soit maintenant une v.a. continue  $X$  de densité  $f(x)$ .

1. L'**espérance** notée  $E(X)$  ( $= \mu = \mu_X$ ) d'une variable aléatoire **continue** dont la densité de probabilité est donnée par  $f(x)$  s'obtient par

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

2. La **variance** notée  $Var(X)$  ( $= \sigma^2$ ) d'une variable aléatoire **continue** dont la densité de probabilité est donnée par  $f(x)$  s'obtient par

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)dx.$$

$$\begin{aligned} E(x+y) &= E(x)+E(y) \\ E(a^*X) &= a^*E(X) \\ Var(x+y) &= var(x)+var(y) \end{aligned}$$

$$var(a^*X) = a^2Var(x)$$

3. L'**écart-type** noté  $\sigma (= \sigma_X)$  d'une variable aléatoire **continue** s'obtient par

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

L'espérance de la somme de deux variables aléatoires est la somme des espérances de chacune des variables aléatoires. Si  $X$  et  $Y$  sont indépendantes, on a en plus:  $E(X + Y) = E(X) + E(Y)$ . D'autre part, si les deux variables aléatoires  $X$  et  $Y$  sont indépendantes, alors  $Var(X + Y) = Var(X) + Var(Y)$ .

Soit  $X$  une variable aléatoire d'espérance  $\mu$  et de variance  $\sigma^2$ . Alors la variable aléatoire

$$Z := \frac{X - \mu}{\sigma}$$

s'appelle la **variable standardisée** de  $X$ . Pour  $Z$  on a:  $E(Z) = 0$ ,  $Var(Z) = 1$ .

Finalement, il est important de connaître l'espérance et la variance de la moyenne de  $n$  variables aléatoires indépendantes et de même distribution. Ceci sera fait en classe.

#### RAJOUT

On considère  $X_1, X_2, X_3, \dots, X_n$  des variables aléatoires indépendantes et identiquement distribuées

on considère la variable aléatoire moyenne

$$\begin{aligned} /X_n &:= (1/n) * (X_1 + X_2 + \dots + X_n) \\ E(/X_n) &= (1/N) * (E(X_1) + E(X_2) + \dots + E(X_n)) = m \\ &= m = m = m = m \\ m \text{ ici est l'espérance d'une variable aléatoire} \\ \text{Var}(/X_n) &= (1/(N^2)) * \text{Var}(X_1 + \dots + X_n) \\ &\quad * (n * \text{Var}(X_1)) \\ &\quad (1/(N^2)) * \text{Var}(X_1) \end{aligned}$$

on utilise souvent  $/X_n$  pour estimer  $E(X_1)$   
qui est souvent inconnue.

**Exemple 33** 1. Un arrêt de bus est desservi toutes les 10 minutes.  $X$  v. a.

indiquant le temps d'attente (en minute) jusqu'au passage du prochain bus lorsqu'on se rend à cet arrêt sans tenir compte de l'horaire. La densité est donnée par

$$f(x) = \begin{cases} 0, & \text{si } x < 0 \text{ ou } x > 10, \\ \frac{1}{10}, & \text{si } 0 \leq x \leq 10. \end{cases}$$

On peut calculer un certain nombre de probabilités.

2.  $X$  variable aléatoire représentant la «durée de vie» d'un modèle de voiture (exprimée en années) dont voici la densité

$$f(x) = \begin{cases} 0, & \text{si } x < 0, \\ 0.2e^{-0.2x}, & \text{si } x \geq 0. \end{cases}$$

Quelle est la probabilité que la voiture ait une durée de vie comprise entre 2 et 6 ans ? (Environ 0.369)

Quelle est la probabilité que la voiture ait une durée de vie inférieure à 2 ans ? (Environ 0.329)

Quelle est la probabilité que la voiture ait une durée de vie supérieure à 10 ans ? (Environ 0.135)

## 2.7 Distributions discrètes importantes

### Loi uniforme discrète

Soit  $X$  une v. a. discrète avec  $CH(X) = \{x_1, x_2, \dots, x_n\}$ .

On dira que  $X$  suit une **loi uniforme** notée  $X \sim \mathcal{U}(n)$  si pour tout  $x_i \in CH(X)$  on a

$$P(X = x_i) = \frac{1}{n}$$

**Exemple 34** 1. Jet d'une pièce de monnaie.  $X$  est une v. a. (de loi uniforme) qui représente le côté obtenu.  $X \sim \mathcal{U}(2)$ . On a  $CH(X) = \{0, 1\} = \{\text{pile, face}\}$

2. Jet d'un dé.  $X$  représente la face obtenue.  $X \sim \mathcal{U}(6)$ . On a  $CH(X) = \{1, 2, 3, 4, 5, 6\}$ . On calcule  $E(X)$ ,  $Var(X)$  et  $\sigma$ .

3. Le choix au hasard d'une personne dans une classe de 10 personnes est une v. a.  $X$  uniforme :  $X \sim \mathcal{U}(10)$

### Loi binomiale

Cette loi apparaît par exemple dans les situations suivantes

- Connaître le nombre de filles d'un groupe de personnes choisies au hasard.
- Identifier le nombre de pièces défectueuses d'un échantillon d'un grand lot dont la proportion de pièces défectueuses est connue.

Soit une expérience aléatoire dont le processus consiste en une répétition  $n$  fois dans les mêmes conditions d'une même **épreuve de Bernoulli** avec  $P(\text{succès}) = p$  et  $P(\text{échec}) = 1 - p = q$ . Soit  $X$  la v. a. qui comptabilise le nombre de succès obtenus à la suite de ces  $n$  épreuves. On a  $CH(X) = \{0, 1, 2, \dots, n\}$ .

On dira que  $X$  suit une **loi binomiale de paramètre  $n$  et  $p$**  notée  $X \sim \mathcal{B}(n; p)$  avec

$$P(X = k) = C_k^n \cdot p^k \cdot q^{n-k}, \quad k = 0, 1, \dots, n$$

On peut montrer que l'espérance de  $X$  est donnée par  $E(X) = n \cdot p$  et la variance de  $X$  par  $Var(X) = n \cdot p \cdot q$ .

### Propriété

Soient  $X_1 \sim \mathcal{B}(n_1; p)$  et  $X_2 \sim \mathcal{B}(n_2; p)$ , deux variables aléatoires indépendantes. On peut montrer que  $X_1 + X_2 \sim \mathcal{B}(n_1 + n_2; p)$ ,  $E(X_1 + X_2) = (n_1 + n_2) \cdot p$  et  $Var(X_1 + X_2) = (n_1 + n_2) \cdot p \cdot q$ .

**Exemple 35** Dans la région de Montréal, on a évalué que si le nom d'une personne est sur une liste d'attente d'un centre hospitalier pour faire du bénévolat sur demande, la probabilité d'être appelé(e) durant le mois est de 0.3. La liste d'attente d'un hôpital (Maisonneuve-Rosemont) est de 30 noms. On a  $p = P(\text{succès}) = P(\text{être appelé}) = 0.3$  et  $q = P(\text{échec}) = 1 - p = 0.7$ . On définit  $X \sim \mathcal{B}(30; 0.3)$

nombre de personnes appelées durant le mois. On veut répondre aux trois questions suivantes:

1. Quelle est la probabilité que plus du tiers des personnes de cette liste soient appelées durant le mois ?
2. Sachant que 7 personnes ont déjà étées appelées, que devient la probabilité que plus du tiers des personnes de la liste soient appelées ?
3. Si la liste est de 60 personnes, combien peut-on espérer appeler de personnes durant 1 mois ?

### Loi de Poisson

On veut déterminer le nombre de fois qu'un événement se produit mais non plus dans un **ensemble dénombrable fini** (loi binomiale) mais dans un **ensemble continu** (temps, longueur, aire, espace,...). Cette loi apparaît par exemple dans les situations suivantes:

- Nombre de tremblements de terre se produisant dans les Caraïbes par an.
- Nombre de désintégrations radioactives d'une substance donnée, dans un intervalle de temps donné.

Soit  $X$  une variable aléatoire qui **comptabilise** le nombre de réalisation d'un événement dans les conditions suivantes

1. La réalisation de l'événement se vérifie par l'examen d'un ensemble de la forme  $]0; t]$ .
2. L'accomplissement d'un événement dans un sous-intervalle de  $]0; t]$  n'influence pas la réalisation de l'événement dans un autre sous-intervalle. (L'événement se passe dans deux sous-intervalles avec la même probabilité si les sous-intervalles sont de même tailles.)
3. La probabilité que l'événement se produise dans un intervalle très petit est presque nulle (rareté).
4. Le **nombre moyen** de réalisation de l'événement dans  $]0; t]$  est égal à  $\lambda$ .

On dira que  $X$  suit une **loi de Poisson de paramètre  $\lambda$**  notée  $X \sim \mathcal{P}(\lambda; t)$  (ou  $X \sim \mathcal{P}(\lambda)$ ) si

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

On peut montrer que l'espérance de  $X$  est donnée par  $E(X) = \lambda$  et la variance de  $X$  par  $Var(X) = \lambda$ .

### Propriété

Soient  $X_1 \sim \mathcal{P}(\lambda_1; t)$  et  $X_2 \sim \mathcal{P}(\lambda_2; t)$ , deux variables aléatoires indépendantes. On peut montrer que  $X_1 + X_2 \sim \mathcal{P}(\lambda_1 + \lambda_2; t)$ ,  $E(X_1 + X_2) = \lambda_1 + \lambda_2$  et la variance  $Var(X_1 + X_2) = \lambda_1 + \lambda_2$

**Exemple 36** Un chien dépitEUR de drogue a été entraîné selon une nouvelle méthode. A l'aéroport de Genève, on rapporte qu'il détecte en moyenne 1.7 cas de passation de drogue par semaine, et à Zürich, sa moyenne hebdomadaire passe à 2.3 cas. On modélise le nombre de détections en  $t$  semaines par une variable aléatoire de Poisson de paramètre  $\lambda t$  et on choisit  $\lambda$  de sorte que la moyenne observée coïncide avec la moyenne théorique, c'est-à-dire l'espérance.

1. Quelle est la probabilité que le chien détecte plus de 5 cas en 2 semaines à Genève ?

Réponse: on définit  $X \sim \mathcal{P}(3.4)$  la variable aléatoire qui donne le nombre de cas détectés en deux semaines à Genève. On a  $E(X) = 3.4$ , ce qui justifie le choix du paramètre  $\lambda$ .

On a  $P(X > 5) = 1 - P(X \leq 5) \cong 0.1295$  avec l'aide de la machine.

2. On raconte que l'an dernier à un des deux aéroports le chien a déjà détecté 5 cas en 2 semaines. Sachant cela quelle est la probabilité que ce soit à Genève sachant que le chien a été utilisé 2 semaines sur 3 à Zürich ?

Réponse: on définit les événements  $G$ : le chien est utilisé à Genève et  $Z$ : le chien est utilisé à Zürich. par un arbre de probabilités, on peut résoudre ce problème de probabilité conditionnelle:

Ainsi

$$P(T = 5) = P(T = 5 \cap G) + P(T = 5 \cap Z)$$

et

$$P(G|T = 5) = \dots \cong 0.268.$$

### Approximation d'une loi binomiale par une loi de Poisson

$X \sim \mathcal{B}(n; p)$  est approchée par  $Y \sim \mathcal{P}(n \cdot p)$  si  $n \geq 30$  et  $n \cdot p < 5$  ou si  $p < 0.1$  et  $n \cdot p < 10$ .

**Exemple 37** Soit  $X \sim \mathcal{B}(30; 1/6)$ . Alors  $Y \sim \mathcal{P}(30 \cdot 1/6) = \mathcal{P}(5)$ . On a  $P(X = 5) = C_5^{30} \cdot (1/6)^5 \cdot (5/6)^{25} \cong 0.192$  et  $P(Y = 5) \cong 0.175$

## 2.8 Distributions continues importantes

### Loi uniforme (continue)

Soit  $X$  variable aléatoire continue. On dira que  $X$  suit une **loi uniforme** sur  $[a, b]$ , notée  $X \sim \mathcal{U}(a; b)$  ou  $X$  est uniformément distribuée sur  $[a, b]$ , si sa densité est donnée par

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

Son **espérance** est donnée par  $E(X) = \frac{a+b}{2}$  et sa **variance** par  $Var(X) = \frac{(b-a)^2}{12}$ .

## Loi exponentielle

Dans de nombreuses situations, nous sommes intéressés à la durée de vie  $T$  définie comme le temps entre l'instant de mise en fonction d'un appareil et l'instant de la première panne. La plupart des phénomènes naturels sont soumis au processus de vieillissement. Mais il existe des phénomènes où il n'y a pas de vieillissement ou d'usure. Il s'agit en général de phénomènes accidentels. Pour ces phénomènes, la probabilité d'être encore en vie ou de ne pas tomber en panne avant un délai donné, sachant que l'objet est en bon état à un instant  $t$ , ne dépend pas de  $t$ . Par exemple, pour un verre en cristal, la probabilité d'être cassé dans les cinq ans ne dépend pas de sa date de fabrication ni de son âge.

Nous voulons définir une distribution de probabilité pour la variable aléatoire continue  $T$  qui prend ses valeurs dans  $[0; \infty[$  en se donnant pour hypothèse que  $P(T \geq 0) = 1$  et que  $P(T \geq t + s | T \geq t) = P(T \geq s)$ .

Les mathématiques nous montrent que la variable ainsi définie a la densité suivante, pour un  $\lambda > 0$ :

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

Cette distribution est appelée **loi exponentielle** de paramètre  $\lambda$ . On écrit notée  $T \sim Exp(\lambda)$ .

On peut montrer que dans ce cas :  $E(T) = \frac{1}{\lambda}$  et  $Var(T) = \frac{1}{\lambda^2}$ .

**Exemple 38** Il s'écoule en moyenne 120 jours entre deux cas d'une certaine maladie due à une bactérie en Suisse. Un cas a été déclaré dans un hôpital A le 10 juin 2006. Quelle est la probabilité qu'il s'écoule moins de 180 jours avant le prochain cas en Suisse ? (Environ 0.7769)

Le 10 juin 2006, on a annoncé dans un hôpital B (suite à des démissions) qu'il n'y aurait pas de spécialiste des urgences durant une certaine période de transition, avant l'arrivée de nouveaux médecins spécialisés. La direction de l'hôpital a même mentionné qu'elle était sûre à 80% qu'il n'y aurait pas de cas de cette bactérie durant cette période. De combien de jours parlait-elle ? (26 jours)

Le 30 juin 2006 aucun nouveau cas de cette bactérie n'avait été déclaré. Quelle est la probabilité qu'il s'écoule encore au moins 180 jours avant le prochain cas ? (Environ 0.223)

## Loi normale (Loi de probabilité la plus importante)

Lorsqu'on examine des phénomènes tels que le poids des hommes, la taille des femmes, le diamètre des boulons fabriqués par une machine, les erreurs de mesures, ... on retrouve souvent une forme de cloche pour la fonction représentant ces phénomènes.

$f(x) = e^{-\frac{x^2}{2}}$  est une fonction dont le graphe est en forme de cloche. Etonnamment, la primitive de  $e^{-\frac{x^2}{2}}$  n'existe pas sous forme analytique!

### Loi normale centrée réduite

Une v. a.  $X$  suit une **loi normale centrée et réduite** notée  $X \sim \mathcal{N}(0; 1)$ , si sa fonction densité est donnée par

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

C'est une fonction paire :  $f(-x) = f(x)$  et on peut montrer que c'est une densité  $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) dx = 1$ .

Son **espérance** est  $E(X) = 0$  et sa **variance** est  $Var(X) = 1$ .

### Loi normale

Soit  $X$  une variable aléatoire continue,  $\mu, \sigma \in \mathbb{R}$ ,  $\sigma > 0$ . On dit que  $X$  suit une **loi normale** notée  $X \sim \mathcal{N}(\mu; \sigma^2)$  si sa fonction de densité est

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Pour une variable aléatoire normale d'espérance  $\mu = 1$  et de variance  $\sigma^2 = 1$ , on note

$$\Phi(t) := P(X \leq t) = \int_{-\infty}^t \phi(x) dx.$$

Son **espérance** est  $E(X) = a$  et sa **variance** est  $Var(X) = b^2$ .

**Exemple 39** Si on a  $X \sim \mathcal{N}(1; 9)$ . Comment peut-on calculer  $P(-2 \leq X \leq 1)$  ?

$$P(-2 \leq X \leq 1) = \frac{1}{3\sqrt{2\pi}} \int_{-2}^1 e^{-\frac{(x-1)^2}{18}} dx.$$

Cette intégrale est impossible à calculer sans machine ou table. On peut montrer que si  $X \sim \mathcal{N}(a; b^2)$ , alors en définissant  $Z = \frac{X-a}{b}$ , on a  $Z \sim \mathcal{N}(0; 1)$  (centrée réduite). On obtient alors pour notre problème

$$\begin{aligned} P\left(\frac{-2-1}{3} \leq \frac{X-1}{3} \leq \frac{1-1}{3}\right) &= P(-1 \leq Z \leq 0) \\ &= P(Z \leq 0) - P(Z \leq -1) = \frac{1}{2} - P(Z \leq -1) \\ &= \frac{1}{2} - P(Z \geq 1) = \frac{1}{2} - (1 - P(Z < 1)) \\ &= -\frac{1}{2} + P(Z < 1) \end{aligned}$$

et on trouve la réponse en consultant les tables!

**Exemple 40** L'âge des citoyens d'une ville suit une loi normale de moyenne 32 et d'écart-type 18.  $X$  sera la v. a. représentant l'âge des citoyens de cette ville,  $X \sim \mathcal{N}(32; 18^2)$ . Une personne est choisie au hasard dans cette ville. On va trouver la probabilité que cette personne soit

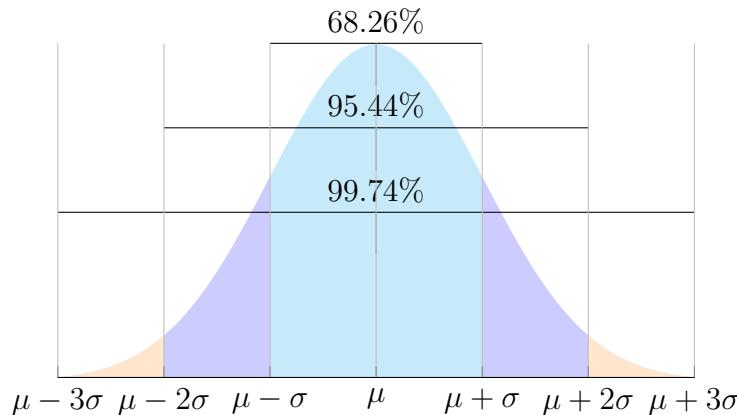


Figure 2.1: Illustration de la propriété 2.

1. Agée entre 32 et 59 ans ?
2. Agée de plus de 50 ans ?
3. Agée de moins de 18 ans ?
4. Agée entre 20 et 40 ans?

Le propriétaire d'une salle de cinéma désire offrir une réduction aux 5% des plus âgés de la ville.

1. A partir de quel âge offrira-t-il cette réduction ?
2. Et si la réduction s'appliquait aux 15% les plus âgés ?

## Propriétés

1. Soient  $X_1 \sim \mathcal{N}(\mu_1; \sigma_1^2)$  et  $X_2 \sim \mathcal{N}(\mu_2; \sigma_2^2)$ , deux variables aléatoires indépendantes. On peut montrer que  $X_1 \pm X_2 \sim \mathcal{N}(\mu_1 \pm \mu_2; \sigma_1^2 + \sigma_2^2)$ ,  $E(X_1 + X_2) = \mu_1 + \mu_2$  et  $Var(X_1 \pm X_2) = \sigma_1^2 + \sigma_2^2$ .
2. Si  $X \sim \mathcal{N}(\mu; \sigma^2)$  alors

$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &\cong 0.6826 \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\cong 0.9544 \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\cong 0.9974 \end{aligned}$$

## Théorème limite central

Soient  $X_1, X_2, \dots, X_n$ , des variables aléatoires indépendantes identiquement distribuées (même fonction de densité, même espérance  $\mu$  et même variance  $\sigma^2$ ) mais pas forcément de loi normale. La somme  $S_n := X_1 + \dots + X_n$  a une espérance de  $n\mu$  et une variance de  $n\sigma^2$ , donc un écart-type de  $\sqrt{n}\sigma$ . La variable standardisée

correspondante  $Z_n := \frac{S_n - n\mu}{\sqrt{n}\sigma}$  a la propriété suivante: Pour  $n \rightarrow \infty$ , la distribution de  $Z_n$  converge vers la distribution normale standard, i.e.

$$\lim_{n \rightarrow \infty} P(Z_n \leq z) = \Phi(z).$$

### Remarques

1. Les variables aléatoires peuvent être discrètes ou continues.
2. Pour  $n \geq 30$ , on obtient de bons résultats.

### Approximation de la loi binomiale par la loi normale (De Moivre La-place)

On peut approximer  $X \sim \mathcal{B}(n; p)$  par  $Y \sim \mathcal{N}(np; npq)$  lorsque  $np(1-p) > 9$  (règle d'usage), i.e.  $n$  doit être suffisamment grand, respectivement  $p$  ne doit pas être trop proche de 0 ou 1.

**Exemple 41** Prenons  $X \sim \mathcal{B}(50; 0.75)$  et calculons  $P(X = 40) = 0.098$ . On peut approximer  $X$  et ce résultat par  $Y \sim \mathcal{N}(37.5; 9.375)$ . On a  $P(Y = 40) = 0$ . Mais  $P(39.5 \leq Y \leq 40.5) = \dots \cong 0.094$ .

On pourrait, par exemple, également calculer  $P(X \geq 40)$  et comparer la réponse obtenue avec l'approximation  $P(Y \geq 40)$ .

## 2.9 Inférence statistique

Lorsqu'on veut, pour une population donnée, examiner le comportement d'une certaine caractéristique dans les unités statistiques, on peut effectuer une étude sur **toutes** les unités statistiques c'est ce que l'on appelle un **recensement**.

On a une alternative qui est d'étudier **un sous-groupe** de la population c'est un **sondage** sur un échantillon.

Recensement	Sondage
Population de taille $N$	Echantillon de taille $n$
<b>Série statistique</b>	<b>Série statistique</b>
$x_1, x_2, \dots, x_N$	$x_1, x_2, \dots, x_n$
<b>Calcul des paramètres</b>	<b>Calcul des paramètres</b>
$\mu = \frac{\sum_{i=1}^N x_i}{N}$ moyenne de la population totale	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ moyenne empirique
$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$ variance de la population totale	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ variance empirique

Tableau 2.1: Illustration

On aimerait savoir si le sondage est une information suffisante pour avoir une information sur la population totale ?

La variable  $X$  est une statistique de moyenne **exacte**  $\mu$  et de variance **exacte**  $\sigma^2$ .  $\bar{x}$  est un **estimateur ponctuel** de  $\mu$  et  $s^2$  est un **estimateur ponctuel** de  $\sigma^2$ .

Peut-on estimer le comportement de **toute** la population à l'aide du comportement d'un échantillon ? Quel est le meilleur estimateur pour un paramètre donné ( $\bar{x}$  ou  $s^2$ ) ? On ne peut pas faire d'estimation ponctuelle «sans erreur» car

1. L'échantillon ne représente pas **exactement** la population.
2. Le nombre d'échantillons possibles est très grand (les répétitions sont admises).
3. Pour chaque échantillon différent alors  $\bar{x}$  et  $s^2$  seront différents.

Pour chaque série statistique  $x_1^j, x_2^j, \dots, x_n^j$ , on peut calculer la moyenne empirique et la variance empirique

$$\bar{x}_j = \frac{\sum_{i=1}^n x_i^j}{n}, \quad s_j^2 = \frac{\sum_{i=1}^n (x_i^j - \mu)^2}{n-1}$$

On définit  $X_1$  comme étant la v. a. représentant toutes les valeurs possibles pour  $x_1$ .  
On définit  $X_2$  comme étant la v. a. représentant toutes les valeurs possibles pour  $x_2$ .

On définit  $X_n$  comme étant la v. a. représentant toutes les valeurs possibles pour  $x_n$ .

Enfin  $\bar{X}$  est la v. a. représentant la moyenne des variables aléatoires  $X_1, \dots, X_n$ .

C'est-à-dire  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ .

On peut montrer le résultat suivant (échantillon avec remise). Soit  $X$ , une variable statistique de moyenne  $\mu$  et de variance  $\sigma^2$  pour une population de taille  $N$ . Soit  $\bar{X}$ , la variable moyenne d'échantillonnage formée à partir de tous les échantillons aléatoires de taille  $n$  formés avec remise à partir de la population.

On a alors  $E(\bar{X}) = E(X) = \mu$  et  $Var(\bar{X}) = \frac{Var(X)}{n} = \frac{\sigma^2}{n}$  (plus on a d'échantillons, plus l'écart à la moyenne sera petit).

De plus, si  $X$  suit une loi normale alors  $\bar{X}$  suit une loi normale. Si  $X$  suit une loi quelconque alors  $\bar{X}$  suit approximativement une loi normale (si  $n \geq 30$ ,  $\bar{X}$  s'approche déjà d'une loi normale).

Ainsi la valeur moyenne d'un échantillon de taille  $n$  ne devrait pas être trop éloigné de la valeur moyenne  $\mu$  de la population car  $E(\bar{X}) = E(X) = \mu$  et la dispersion des valeurs de toutes les moyennes d'échantillons est assez petite  $\frac{1}{n}\sigma^2$ .

### En résumé

1. Si  $X \sim \mathcal{N}(\mu; \sigma^2)$  alors  $\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$
2. Si  $X \sim$  loi quelconque avec  $n \geq 30$  alors  $\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$

**Exemple 42** On tire 25 hommes au hasard d'une population dont la taille est distribuée normalement avec une moyenne de 172 [cm] et un écart-type de 6 [cm].

Calculer la probabilité que la moyenne des tailles de l'échantillons soit inférieure à 170 [cm].

On sait que  $X \sim \mathcal{N}(172; 36) = \mathcal{N}(\mu; \sigma^2)$  ainsi  $\bar{X} \sim \mathcal{N}(172; \frac{36}{25}) = \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$  avec  $n = 25$ .

$$\begin{aligned} P(\bar{X} < 170) &= P\left(\frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} < \frac{170 - \mu}{\sigma} \cdot \sqrt{n}\right) = P\left(Z < -\frac{5}{3}\right) = P\left(Z > \frac{5}{3}\right) \\ &= 1 - P\left(Z \leq \frac{5}{3}\right) \cong 1 - P(Z \leq 1.67) \cong 1 - 0.9525 = 0.0475 \end{aligned}$$

### 2.9.1 Intervalle de confiance pour la moyenne

On peut estimer la moyenne et la variance d'une **population** à l'aide des valeurs correspondantes d'un **échantillon**. On obtient alors des valeurs **ponctuelles** qui ne coïncident pas exactement, le plus souvent, avec le paramètre cherché. On préfère souvent donner un **intervalle de confiance** pour le paramètre en précisant qu'il a telle ou telle probabilité (**niveau de confiance**) de recouvrir le paramètre cherché.

**Exemple 43** Une entreprise fabrique des écrans pour ordinateur. Elle sait que leur durée de vie est une distribution (approximativement) normale d'écart-type  $\sigma = 50$  heures. Désirant connaître la durée de vie moyenne  $\mu$  des écrans fabriqués, elle détermine cette durée de vie moyenne pour un échantillon de  $n = 16$  écrans.  $\bar{X} = 2200$  heures.

On va construire un intervalle de confiance pour  $\mu$  au niveau de 95%. On sait  $\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$ . Ainsi l'écart-type est de  $\frac{\sigma}{\sqrt{n}}$ . On aimerait que  $P(\mu - a \leq \bar{X} \leq \mu + a) = 0.95$ . En normalisant, on trouve (avec  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ )

$$P\left(\frac{-a}{\sigma/\sqrt{n}} \leq Z \leq \frac{a}{\sigma/\sqrt{n}}\right) = 0.95$$

En regardant dans la table, on trouve  $\frac{a}{\sigma/\sqrt{n}} \cong 1.96$ , et donc  $a = \frac{1.96}{\sqrt{n}} \cdot \sigma$ . On peut montrer que  $\mu - a \leq \bar{X} \leq \mu + a \Leftrightarrow \bar{X} - a \leq \mu \leq \bar{X} + a$ . Ainsi l'intervalle de confiance à 95% pour la moyenne est donné par

$$\bar{X} - \frac{1.96}{\sqrt{n}} \sigma \leq \mu \leq \bar{X} + \frac{1.96}{\sqrt{n}} \sigma$$

Dans notre problème, on trouve  $2175.5 \leq \mu \leq 2224.5$ . C'est l'intervalle de confiance pour  $\mu$  à 95%.

De manière similaire, on dérive l'intervalle de confiance à 90% pour la moyenne. Il est donné par

$$\bar{X} - \frac{1.6449}{\sqrt{n}} \sigma \leq \mu \leq \bar{X} + \frac{1.6449}{\sqrt{n}} \sigma$$

### En résumé ( $\sigma^2$ connue)

Soit  $X$  une variable aléatoire de moyenne  $\mu$  et de variance  $\sigma^2$  **connue**. Soit  $\bar{x}$  la moyenne d'un échantillon aléatoire de cette population. Si  $X \sim \mathcal{N}(\mu; \sigma^2)$  ou  $X \sim$  loi quelconque avec  $n \geq 30$ , alors

1.

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

Rappel:  $\bar{X}$  est la variable moyenne d'échantillonnage formée à partir de tous les échantillons aléatoires de taille  $n$  formés avec remise à partir de la population.

2. l'intervalle de confiance pour la moyenne  $\mu$  de la population au niveau  $(1 - \alpha) \cdot 100\%$  est donné par

$$\left[ \bar{x} - z_{\frac{\alpha}{2}} \left( \frac{\sigma}{\sqrt{n}} \right), \bar{x} + z_{\frac{\alpha}{2}} \left( \frac{\sigma}{\sqrt{n}} \right) \right]$$

avec  $z_c$  la valeur prise par  $Z$  ( $\sim \mathcal{N}(0; 1)$ ) pour laquelle  $P(Z > z_c) = c$  (voir la table des quantiles de la loi normale centrée réduite).

Jusqu'ici, on a supposé que  $\sigma^2$  est **connue**. En réalité, ceci n'est pas le cas et on doit se contenter de  $s$  (l'écart-type de l'échantillon).

On peut remplacer  $\sigma$  (écart-type de la population) par  $s$  (écart-type de l'échantillon) lorsque  $n$  (taille de l'échantillon) est suffisamment grand, c'est-à-dire que  $n \geq 30$ .

Sinon (lorsque  $n < 30$ ), les valeurs issues de la loi normale (par exemple 1.96 et 1.6449) doivent être remplacées par celles obtenues par la distribution **T de Student** à  $n - 1$  **degrés de liberté lorsque l'on sait que la variable étudiée suit une loi normale**.

### En résumé ( $\sigma^2$ inconnue)

Soit  $X$  une variable aléatoire de moyenne  $\mu$  et de variance  $\sigma^2$  **inconnue**. Soit un échantillon aléatoire de cette population de moyenne  $\bar{x}$  et d'écart-type  $s$ .

1. Si  $n \geq 30$ , alors

(a)

$$Z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim \mathcal{N}(0, 1);$$

Rappel:  $\bar{X}$  est la variable moyenne d'échantillonnage formée à partir de tous les échantillons aléatoires de taille  $n$  formés avec remise à partir de la population.

(b) l'intervalle de confiance pour la moyenne  $\mu$  de la population au niveau  $(1 - \alpha) \cdot 100\%$  est donné par

$$\left[ \bar{x} - z_{\frac{\alpha}{2}} \left( \frac{s}{\sqrt{n}} \right), \bar{x} + z_{\frac{\alpha}{2}} \left( \frac{s}{\sqrt{n}} \right) \right]$$

avec  $z_c$  la valeur prise par  $Z$  ( $\sim \mathcal{N}(0; 1)$ ) pour laquelle  $P(Z > z_c) = c$  (voir la table des quantiles de la loi normale centrée réduite en annexe).

2. Si  $n < 30$  et  $X \sim$  loi normale, alors

(a)

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n-1}}} \sim T_{n-1}$$

où  $T_k$  est la loi  $T$  de **Student-Fisher** (distribution de Student-Fisher) à  $k$  degrés de liberté (aussi notée  $T_k$ ) (voir la table de la loi en annexe);

(b) l'intervalle de confiance pour la moyenne  $\mu$  de la population au niveau  $(1 - \alpha) \cdot 100\%$  est donné par

$$\left[ \bar{x} - t_{\frac{\alpha}{2}, n-1} \left( \frac{s}{\sqrt{n-1}} \right), \bar{x} + t_{\frac{\alpha}{2}, n-1} \left( \frac{s}{\sqrt{n-1}} \right) \right]$$

avec  $t_{c,k}$  la valeur prise par  $T$  ( $\sim T_k$ ) pour laquelle  $P(T \geq t_{c,k}) = c$  (voir la table de la loi en annexe).

**Exemple 44** Afin de tester la durabilité d'une nouvelle peinture pour des lignes de marquage, un département d'autoroute a peint des lignes de test sur des routes très utilisées dans 8 villes différentes. Des compteurs ont été placés pour déterminer le nombre de passage de véhicules sur les lignes jusqu'à déterioration de celles-ci. Voici les résultats obtenus. On suppose ici que la durabilité de la nouvelle peinture suit approximativement une loi normale.

Ville	1	2	3	4	5	6	7	8
Nb passages ( $\cdot 10^6$ )	14.26	16.78	13.65	10.83	12.64	13.37	16.20	14.94

On va construire un intervalle de confiance pour la moyenne  $\mu$  à 95%.

# Chapitre 3

## Quelques tests d'hypothèses

### 3.1 Introduction

Il arrive fréquemment que l'on se pose des questions concernant une population. Comme par exemple: ce dé est-il pipé? Ce remède est-il efficace? Cette lotion capillaire tient-elle ses promesses? On est alors amené à faire des **hypothèses** sur la population parente. On extrait ensuite un **échantillon** de la population et on confronte les valeurs de l'échantillon avec ce que l'hypothèse nous permettait de prévoir. Si ces valeur sont trop improbables, (couramment on introduit des seuils critiques de 5% et de 1%), on **rejette** l'hypothèse.

**Exemple 45** *On lance une pièce de monnaie 100 fois et on observe 58 fois face. La pièce est-elle symétrique?*

#### Formellement

1. On formule deux hypothèses  $H_0$  et  $H_1$  (**hypothèse nulle** et **hypothèse alternative**) concernant un paramètre de la population ou de la distribution.
2. On fixe **un seuil de signification**  $\alpha$  (entre 1% et 5% en général)
3.  $X$  étant une statistique associée à un échantillon de taille  $n$ , on détermine une **zone critique**  $I$  telle que la probabilité que  $X$  prenne une valeur de  $I$  soit inférieur à  $\alpha$ , sous l'hypothèse  $H_0$ .
4. On **rejette**  $H_0$  (et on accepte  $H_1$ ) si l'échantillon que l'on a prélevé livre une valeur de la statistique  $X$  située dans la zone critique. Dans le cas contraire, on ne rejette pas  $H_0$  (on accepte  $H_0$  et on rejette  $H_1$ ).

#### Remarques:

1. Dans l'exemple précédent, on a fait un **test unilatéral**, car la zone critique est située à droite de 58.73. On a procédé ainsi car on a suspecté que la pièce présente plus souvent le côté face que le côté pile. Ainsi,  $H_1 : p > 0.5$ .
2. Si on avait seulement suspecté que la pièce n'est pas symétrique ( $H_1 : p \neq 0.5$ ), on aurait fait un **test bilatéral**. Dans ce cas, on cherche  $c$  tel que  $P(|X - 50|) \geq$

$c) \leq 0.05$  et on aurait trouvé que la zone critique est donnée par  $X \geq 61$  ou  $X \leq 39$ .

**Exemple 46** Une fabrique achète régulièrement à un fournisseur un certain type de transistors. Elle sait, par les nombreux sondages qu'elle a effectués, que le coefficient d'amplification de ces transistors a une distribution de moyenne  $\mu = 155$  et d'écart-type  $\sigma = 42$ . Lors d'un nouveau contrôle portant sur 200 transistors, elle observe un coefficient d'amplification moyen de 146. Peut-elle admettre, au seuil de 1%, que la qualité n'a pas varié?

**Exemple 47** On jette une pièce de monnaie 100 fois et on observe 62 faces. Peut-on admettre que la pièce est symétrique?

Par l'exemple précédent, on a introduit la notion d'**erreur de 1<sup>e</sup> espèce et de 2<sup>e</sup> espèce**.

## 3.2 Logistique d'un test d'hypothèses de comparaison

Pour comprendre le déroulement d'un test d'hypothèses de comparaison, examinons ce qui se passe dans un test de comparaison d'une moyenne.

**Exemple 48** Tout au long de cette section, nous traiterons des deux exemples suivants:

1. Les étudiants des HES travaillant l'été gagnaient CHF 8000.– par année selon une étude vieille de deux ans. Est-ce toujours valable?
2. L'espérance de vie des femmes au Québec était de 77 ans il y a cinq ans. Compte tenu de l'avancement médical en gériatrie, on croit que cette espérance a augmenté. Est-ce exact?

Dans les deux exemples ci-dessus, on a affaire à une population  $X$  pour laquelle, théoriquement, la moyenne  $\mu$  de la variable  $X$  a comme valeur (connue)  $\mu_0$

**Exemple 49** Reprenons nos exemples:

1.  $X$ : variable aléatoire représentant le gain annuel des étudiants HES travaillant l'été.  $\mu = \mu_0 = 8000.–$
2.  $X$ : variable aléatoire représentant l'espérance de vie des femmes au Québec.  $\mu = \mu_0 = 77$ .

Pour diverses raisons, nous sommes portés à croire que la valeur de  $\mu$  a subi des modifications. On est donc en présence de deux hypothèses plausibles

- **hypothèse du statu quo ou hypothèse nulle**, notée  $H_0$ , signifiant qu'aucun changement n'est survenu et que la valeur moyenne  $\mu$  est toujours égale à  $\mu_0$ ;
- **hypothèse alternative ou contre-hypothèse**, notée  $H_1$ , signifiant qu'un changement est survenu sous une des trois formes suivantes
  1.  $\mu$  est **diffrérente** de  $\mu_0$ ;

2.  $\mu$  est **plus grande** que  $\mu_0$ ;
3.  $\mu$  est **plus petite** que  $\mu_0$ .

Il y a donc trois paires d'hypothèses plausibles relativement à une moyenne

Test bilatéral

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &\neq \mu_0 \end{aligned}$$

Test unilatéral à droite

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &> \mu_0 \end{aligned}$$

Test unilatéral à gauche

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &< \mu_0 \end{aligned}$$

**Exemple 50** Reprenons nos exemples:

1.  $H_0 : \mu = 8000.$  – (*Hypothèse nulle, statu quo*)  
 $H_1 : \mu \neq 8000.$  – (*Hypothèse alternative, contre-hyothèse*)
2.  $H_0 : \mu = 77$  (*Hypothèse nulle, statu quo*)  
 $H_1 : \mu > 77$  (*Hypothèse alternative, contre-hyothèse*)

### Que fait-on pour la suite?

Pour confronter l'hypothèse de statu quo ( $H_0$ )  $\mu = \mu_0$ , on choisira un échantillon de façon aléatoire, on calculera la moyenne  $\bar{x}$  de cet échantillon que l'on confrontera à  $\mu_0$  pour prendre une décision.

#### Règle de décision

- Si la valeur de  $\bar{x}$  est “près de la valeur  $\mu_0$ ”, on optera pour l'hypothèse  $H_0$  (statu quo);
- Si  $\bar{x}$  est “loin de la valeur  $\mu_0$ ”, on décidera de rejeter l'hypothèse  $H_0$ .

**Que signifient  $\bar{x}$  est “près de la valeur  $\mu_0$ ” ou  $\bar{x}$  est “loin de la valeur  $\mu_0$ ”?**

On examine les règles de décisions des différents cas pour nous aider à nous faire une idée sur ces expressions.

Cas 1: Test bilatéral avec

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &\neq \mu_0 \end{aligned}$$

Règle de décision:

- Si  $\bar{x} \in [\mu_0 - d; \mu_0 + d]$  on accepte  $H_0$
- Si  $\bar{x} \notin [\mu_0 - d; \mu_0 + d]$  on rejette  $H_0$

Cas 2: Test unilatéral à droite avec

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &> \mu_0 \end{aligned}$$

Règle de décision:

- Si  $\bar{x} \in ] -\infty; \mu_0 + d]$  on accepte  $H_0$

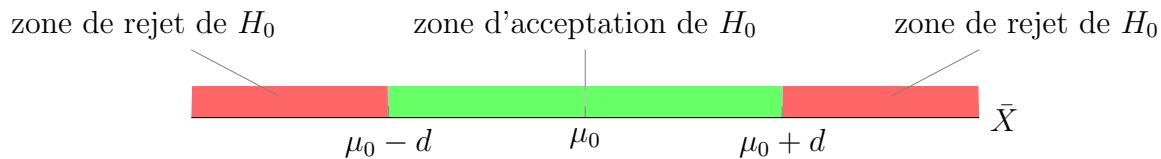


Figure 3.1: Illustration de la règle de décision. Test bilatéral.

- Si  $\bar{x} \in ]\mu_0 + d; +\infty[$  on rejette  $H_0$

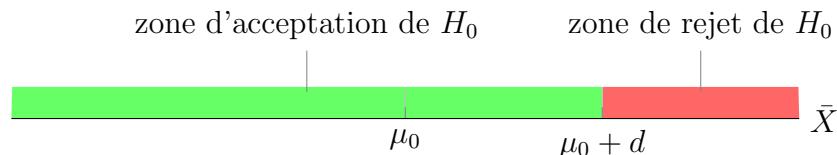


Figure 3.2: Illustration de la règle de décision. Test unilatéral à droite.

Cas 3: Test unilatéral **à gauche** avec

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0$$

Règle de décision:

- Si  $\bar{x} \in [\mu_0 - d; +\infty[$  on accepte  $H_0$
- Si  $\bar{x} \in ]-\infty; \mu_0 - d]$  on rejette  $H_0$

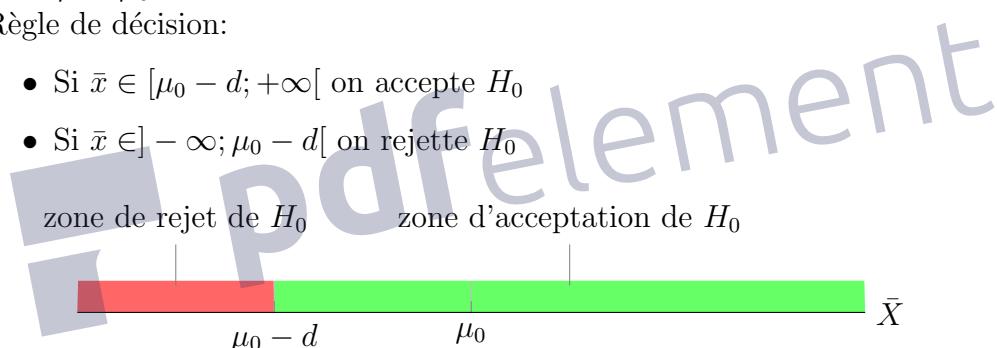


Figure 3.3: Illustration de la règle de décision. Test unilatéral à droite.

Pour nous assurer de l'objectivité de notre décision, nous établirons notre “règle de décision” avant même de connaître la valeur  $\bar{x}$  et nous l'établirons en cherchant à minimiser les erreurs de décision possibles

Examinons l'organigramme des décisions possibles et quantifions les probabilités conditionnelles qui s'y trouvent.

### Types d'erreurs dans un test d'hypothèses de comparaison

Lors d'un test d'hypothèses, on commet une **erreur de première espèce** lorsqu'on décide de rejeter l'hypothèse  $H_0$  alors qu'en réalité elle est vraie. **On rejette une hypothèse correcte.** La probabilité de commettre une erreur de première espèce, notée  $\alpha$ , est donnée par

$$\alpha = P(\text{rejeter } H_0 | H_0 \text{ est vraie}).$$

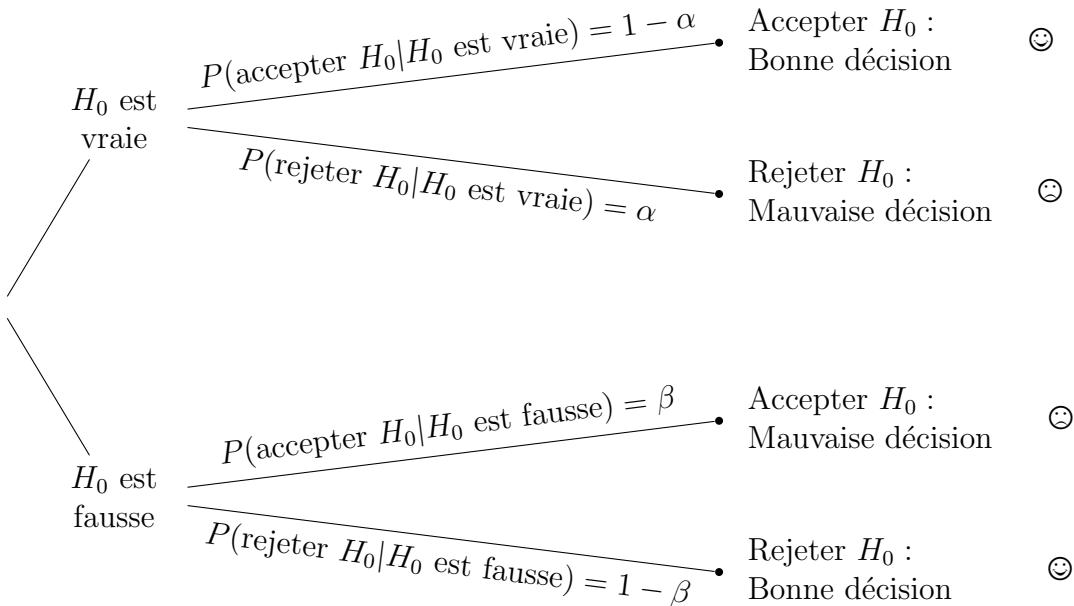


Figure 3.4: Organigramme des décisions possibles.

Lors d'un test d'hypothèses, on commet une **erreur de deuxième espèce** lorsqu'on décide d'accepter l'hypothèse  $H_0$  alors qu'en réalité est elle est fausse. **On accepte une hypothèse fausse.** La probabilité de commettre une erreur de deuxième espèce, notée  $\beta$ , est donnée par

$$\beta = P(\text{accepter } H_0 | H_0 \text{ est fausse}).$$

On se concentre pour l'instant sur l'erreur de première espèce. On reviendra plus tard sur l'erreur de deuxième espèce.

On sait

“étant donné que  $H_0$  est vraie”  $\Leftrightarrow$  “étant donné que  $\mu = \mu_0$ ” et puisque (par le théorème limite central)  $\mu_{\bar{X}} = \mu_X = \mu$ , cela signifie également “étant donné que  $\bar{X}$  est centré en  $\mu_0$ ”. Ainsi

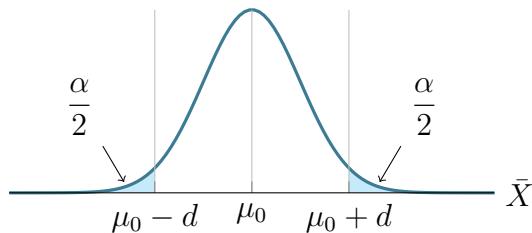
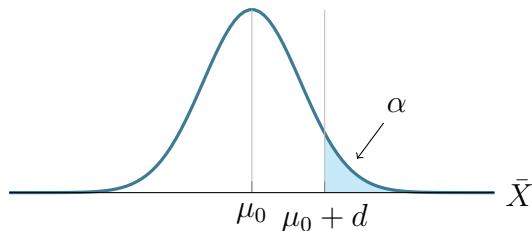
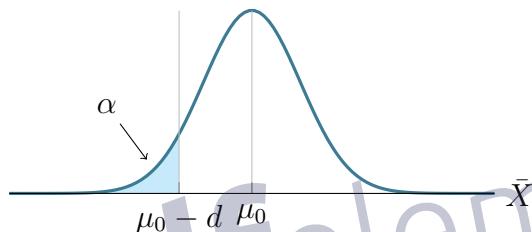
$$\begin{aligned} \alpha &= P(\text{rejeter } H_0 | H_0 \text{ est vraie}) \\ &= P(\bar{X} \text{ est dans la zone de rejet de } H_0 | \text{la courbe de } \bar{X} \text{ est centrée en } \mu_0). \end{aligned}$$

Ainsi  $\alpha$  est la surface sous la courbe de  $\bar{X}$  (centrée en  $\mu_0$ ) correspondant à la zone de rejet de  $H_0$ .

### Graphiquement

Fixer  $\alpha$ , c'est déterminer ce que l'on est prêt à accepter comme probabilité de commettre l'erreur de première espèce et c'est également déterminer les zones d'acceptation et de rejet de  $H_0$ .

C'est une situation qui est semblable aux intervalles de confiance! En utilisant un cheminement de même nature, on trouve que les valeurs de  $d$  sont des multiples de l'écart-type  $\sigma_{\bar{X}}$  du style

Figure 3.5: Test bilatéral avec  $H_0 : \mu = \mu_0$ ,  $H_1 : \mu \neq \mu_0$ .Figure 3.6: Test unilatéral avec  $H_0 : \mu = \mu_0$ ,  $H_1 : \mu \geq \mu_0$ .Figure 3.7: Test unilatéral avec  $H_0 : \mu = \mu_0$ ,  $H_1 : \mu \leq \mu_0$ .

bilatéral	unilatéral	$\sigma_{\bar{X}}$	$\sigma$
$z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{X}}$	$z_{\alpha} \cdot \sigma_{\bar{X}}$	$\frac{\sigma}{\sqrt{n}}$	<b>connu</b>
$t_{\frac{\alpha}{2}, n-1} \cdot \sigma_{\bar{X}}$	$t_{\alpha, n-1} \cdot \sigma_{\bar{X}}$	$\frac{s}{\sqrt{n-1}}$	<b>inconnu</b>

Tableau 3.1: Valeurs de  $d$  (lorsque  $X$  suit une loi normale).

Puisque  $\alpha$  est important dans cette démarche, nous l'appelons **significatif**.

La probabilité de commettre l'erreur de première espèce  $\alpha$  se nomme aussi **le seuil de signification** du test.

### 3.3 Test de comparaison d'une moyenne $\mu$ à une valeur $\mu_0$ (test de Student)

**Exemple 51** Un employé responsable du contrôle de qualité des lampes électriques doit tester avec un seuil de signification de 5% la durée moyenne théorique de 2500 heures d'une certaine marque de lampes de 60 watts. Il sait que la durée de vie de ces lampes suit une loi normale d'écart-type de 55 h. La moyenne obtenue pour un échantillon de 20 lampes est de 2479 h. Que décidera-t-il?

### **Solution**

1. Examen de la situation
2. Vers la règle de décision
3. Prise de décision

De manière générale, la mise en place d'un test d'hypothèses (de comparaison d'une moyenne  $\mu$  à une valeur  $\mu_0$ ) comprend les phases suivantes

1. Examen de la situation
  - (a) La variable aléatoire  $X$  est à définir à laquelle  $\mu$  est rattachée
  - (b) Identifier  $\mu_0$  et poser  $H_0$  et  $H_1$
  - (c) Définir le seuil de signification  $\alpha$  choisi
2. Vers la règle de décision
  - (a) La variable aléatoire  $\bar{X}$  est à définir et préciser ce que l'on connaît sur  $\bar{X}$
  - (b) Si  $H_0$  est vraie, dessiner les zones "ok/nok"
  - (c) Calculer les frontières: trouver la valeur de  $d$
  - (d) Formuler la règle de décision
3. Prise de décision: utiliser la valeur  $\bar{x}$  prise par la variable aléatoire  $\bar{X}$  pour prendre la décision.

**Exemple 52** L'entraîneur d'une équipe de football affirme que cette année, la stratégie de jeu qu'il préconise en deuxième mi-temps est meilleure qu'avant. On veut vérifier cette bonne nouvelle. On examine les statistiques des matchs des années passées. La moyenne des points gagnés en deuxième mi-temps est de 35. Au seuil de signification de 10% et en supposant que le nombre de points accumulés suit une loi normale est-ce que l'on peut affirmer que l'entraîneur a raison de vanter sa stratégie si un échantillon de 25 matchs de cette année révèle une moyenne de 38.1 points gagnés avec un écart-type de 6.6 points?

### **Solution**

1. Examen de la situation
2. Vers la règle de décision
3. Prise de décision

#### **3.3.1 Puissance d'un test d'hypothèses**

$$\begin{aligned}\beta &= P(\text{commettre une erreur de deuxième espèce}) \\ &= P(\text{accepter } H_0 | H_0 \text{ est fausse}) \\ &= P(\bar{X} \text{ est dans "ok" calculée à priori avec } \mu = \mu_0 | \\ &\quad \bar{X} \text{ suit une loi centrée en } \mu = \mu_1 \neq \mu_0)\end{aligned}$$

Ainsi pour **chaque**  $\mu_1$  suggéré une valeur  $\beta$  est calculée et notée  $\beta_{\mu_1}$ .

### Graphiquement

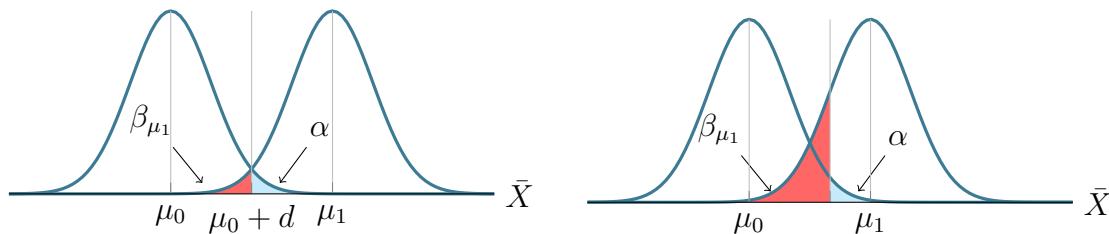


Figure 3.8: Puissance d'un test d'hypothèses pour deux valeurs de  $\mu_1$ .

Plus  $\mu_1$  est proche de  $\mu_0$  plus la valeur de  $\beta_{\mu_1}$  sera grande. La **puissance** (force) d'un test se mesure donc à sa capacité à détecter le fait que la moyenne  $\mu$  vaut maintenant une nouvelle valeur  $\mu_1$ , même si cette valeur est proche de l'ancienne valeur  $\mu_0$ . La **puissance d'un test d'hypothèses** se mesure en évaluant

$$\begin{aligned} 1 - \beta_{\mu_1} &= P(\text{rejeter } H_0 | H_0 \text{ est fausse}) \\ &= P(\text{rejeter } H_0 | \bar{X} \text{ est centré en } \mu = \mu_1) \end{aligned}$$

pour différentes valeur de  $\mu_1$ .

**Exemple 53** On reprend l'exemple 51 des lampes présentés plus haut.

La durée de vie du lot de 60 lampes suit théoriquement une loi normale de moyenne 2500 (heures) et d'écart-type 55 (heures). Pour un échantillon de 20 lampes, avec un seuil de 5%, on a trouvé comme règle de décision  
accepter  $H_0$  si  $2475.9 \text{ h} < \bar{x} < 2524.1 \text{ h}$   
rejeter  $H_0$  sinon.

Notre objectif ici est de déterminer l'erreur de deuxième espèce pour les valeurs moyennes

$$\mu = 2505, 2510, 2495, 2485.$$

On trouve

$$\beta_{2505} = \beta_{2495} = 0.9305, \quad \beta_{2510} = 0.8721, \quad \beta_{2485} = 0.7967$$

et ainsi la puissance du test est donnée par

$$1 - \beta_{2505} = 1 - \beta_{2495} = 0.0695, \quad 1 - \beta_{2510} = 0.1279, \quad 1 - \beta_{2485} = 0.2033$$

### Facteurs qui affectent la puissance d'un test

1. Distance de  $\bar{X}$  à  $\mu_0$ .
2. Taille de l'échantillon  $n$ . Lorsque  $n$  augmente, il y a moins de chevauchement car  $s_{\bar{X}} = \frac{s_X}{\sqrt{n}}$ .

3. Si  $\alpha$  augmente alors  $1 - \beta$  augmente. L'erreur de type 1 augmente lorsque l'erreur de type 2 diminue et vice-versa.
4. Un test unilatéral est plus puissant qu'un test bilatéral.

### 3.3.2 Test de comparaison des moyennes $\mu_1$ et $\mu_2$ de deux populations

- On aimerait savoir si le temps moyen de réaction d'un composé chimique est amélioré avec l'introduction d'une certaine substance.
- On aimerait évaluer si la résistance moyenne du béton est altérée par la température de séchage.
- On aimerait déterminer si un nouveau procédé de fabrication pour une puce électronique modifiera de façon significative leur durée de vie.

Dans chacun des exemples présentés ci-dessus, on va obtenir deux échantillons dont on va extraire deux moyennes et on va s'intéresser à la différence des moyennes.

#### Eléments nécessaires pour faire un test d'hypothèses de comparaison de moyennes

1.  $X_1$  et  $X_2$  sont deux variables aléatoires représentant les variables mesurées sur deux populations dont les moyennes et les variances sont données par  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1^2$  et  $\sigma_2^2$ .
2.  $\bar{X}_1$ : variable d'échantillonnage fabriquée à partir des échantillons de taille  $n_1$  provenant de la première population.
3.  $\bar{X}_2$ : variable d'échantillonnage fabriquée à partir des échantillons de taille  $n_2$  provenant de la deuxième population.

Hypothèses possibles pour le test de comparaison:

Test bilatéral	Test unilatéral à droite	Test unilatéral à gauche
----------------	--------------------------	--------------------------

$$H_0 : \mu_1 = \mu_2 \quad H_0 : \mu_1 = \mu_2 \quad H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2 \quad H_1 : \mu_1 > \mu_2 \quad H_1 : \mu_1 < \mu_2$$

On peut montrer que

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

et

$$Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

et si  $X_1$  et  $X_2$  sont indépendantes et suivent des lois normales, alors  $\bar{X}_1 - \bar{X}_2$  suivra également une loi normale. Dans ce cas, la valeur de  $d$  sera donnée par

bilatéral	unilatéral	$\sigma_{\bar{X}_1 - \bar{X}_2}$	$k$	$s_p^2$	$\sigma_1^2$ et $\sigma_2^2$
$z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{X}_1 - \bar{X}_2}$	$z_\alpha \cdot \sigma_{\bar{X}_1 - \bar{X}_2}$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	—	—	connues
$t_{\frac{\alpha}{2}, k} \cdot \sigma_{\bar{X}_1 - \bar{X}_2}$	$t_{\alpha, k} \cdot \sigma_{\bar{X}_1 - \bar{X}_2}$	$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$n_1 + n_2 - 2$	(*)	inconnues, égales
$t_{\frac{\alpha}{2}, k} \cdot \sigma_{\bar{X}_1 - \bar{X}_2}$	$t_{\alpha, k} \cdot \sigma_{\bar{X}_1 - \bar{X}_2}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	(**)	—	inconnues, différentes

Tableau 3.2: Valeurs de  $d$ 

avec

$$(*) : \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \quad (**) : \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}.$$

**Exemple 54** Un manufacturier de rouleaux de tapisserie effectue des essais afin de savoir si l'additif d'un certain produit réduit le temps de séchage de la colle qu'il applique au dos de ses rouleaux prêt-à-poser. La même semaine dans des conditions semblables on fait l'essai de 22 pièces de tapisserie du produit original et de 18 du produit modifié. On sait que le temps de séchage du produit original suit une loi normale d'écart-type 32 minutes et on suppose que l'additif n'a pas eu d'effet sur la dispersion des données. En utilisant un seuil de 5% peut-on penser que le temps de séchage est réduit significativement si le temps moyen de séchage du produit original est de 143 minutes tandis que le temps moyen des autres a été de 129 minutes.

### Solution

1. Examen de la situation
2. Vers la règle de décision
3. Prise de décision

## 3.4 Analyse de la variance (ANOVA)

**Exemple 55** On souhaite évaluer l'effet de cinq traitements différents sur le comportement de patients dépressifs âgés de 18 à 55 ans. On mesure le niveau de dépression (au travers d'un score). Cinq échantillons de neuf patients chacun ont été considérés.

On a ici une variable **quantitative** (score) et une variable **qualitative** (traitement) dont les **modalités** sont les cinq différents traitements. On se pose la question de savoir si ces cinq traitements diffèrent. On désire comparer l'effet des traitements et voir s'il y a un lien entre la variable quantitative et la variable qualitative.

On peut utiliser des tests de comparaison de deux moyennes pour deux échantillons indépendants vus dans la section précédente. Dans ce cas, on doit comparer le traitement 1 au traitement 2, le traitement 1 au traitement 3 et ainsi de suite... Cela fait au total

$$C_2^5 = 10$$

tests de comparaison de deux moyennes. Cela implique un nombre relativement élevé de calculs. De plus, les comparaisons ne sont pas indépendantes puisque l'on réutilise le même ensemble de données plusieurs fois. Ainsi le nombre d'erreur de première espèce augmente. Supposons que  $P(\text{erreur de première espèce}) = 5\%$  pour une comparaison. Si on effectue  $N$  comparaisons, la probabilité de commettre **au moins** une erreur de première espèce est donnée par

$$P(\text{commettre au moins une erreur de première espèce}) = 1 - (1 - 0.05)^N.$$

Lorsque  $N = 10$  (comme dans notre exemple), cette probabilité vaut environ 0.4! C'est énorme et inadmissible. Il faut donc trouver un autre moyen pour comparer les différents traitements. C'est l'ANOVA qui a été développé par Fisher (sous l'hypothèse de normalité).

### 3.4.1 ANOVA à un facteur

On utilise l'analyse de la variance (ANOVA) à un facteur lorsque l'on dispose

- d'une variable quantitative  $Y$  (variable dépendante) VD;
- d'une variable qualitative  $X$  à  $k$  modalités (variable indépendante également appelée facteur) VI;
- de  $k$  échantillons indépendants ( $E_1, E_2, \dots, E_k$ ) de taille  $n_1, n_2, \dots, n_k$ .

On cherche un lien entre la VI et la VD. On veut étudier l'influence des différentes modalités de la VI sur la VD.

**Exemple 56**    1. *Etude de la réussite scolaire pour des élèves de troisième secondaire de différents pays. La VI a trois niveaux: Pays 1, 2 et 3 et la VD est la performance à l'examen de différents élèves.*

2. *On s'intéresse au taux de cholestérol en fonction de la catégorie socio-professionnelle. La VI est la catégorie socio-professionnelle (dont les modalités sont retraité, étudiant, agriculteur, cadre, ouvrier) et la VD est le taux de cholestérol dans le sang.*
3. *On veut mettre en évidence l'effet d'un substrat nutritif sur la croissance de plantes. La VI est le type de substrat à plusieurs niveaux (concentration de substrat) et la VD est la longueur de la tige.*

#### Remarques

1. Il existe différents types d'ANOVA qui se distinguent par le nombre de facteurs étudiés. Ici on ne traitera que de l'**analyse à un facteur** (une seule VI). S'il y a plusieurs variables indépendantes, on parle alors d'**analyse factorielle** ou de **plan factoriel**.
2. Dans ce qui suit (à part tout à la fin du chapitre lors de la conclusion), on supposera que les  $k$  échantillons sont tous de même taille, c'est-à-dire que  $n_1 = n_2 = \dots = n_k = n$ .

### 3.4.2 Hypothèses de l'ANOVA

- Hypothèse nulle:  $H_0$

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu.$$

Les échantillons **aléatoires** et **indépendants** proviennent de  $k$  populations statistiques **normales** de **même** moyenne et de **même** variance  $\sigma^2$  (homogénéité des variances ou homoscédasticité) ou d'une population unique **normale**  $\mathcal{N}(\mu, \sigma^2)$ .

- Hypothèse alternative:  $H_1$

$$H_1 : \mu_i \neq \mu_j \text{ pour au moins un } i \text{ différent de } j.$$

Certaines moyennes (ou toutes) diffèrent les unes des autres. Autrement dit, le facteur testé a un effet **significatif** sur la variable mesurée.

**Exemple 57** On a trois échantillons d'élèves de troisième secondaire qui font leurs études dans trois pays différents (Pays 1, Pays 2 et Pays 3). Chaque échantillon est composé de cinq élèves choisis aléatoirement parmi la population des élèves du pays. On fait passer le même test de logique (noté sur 100) aux trois échantillons d'élèves. Voici les résultats obtenus: On aimerait déterminer si les élèves des pays ont des

Elève	Pays 1	Pays 2	Pays 3
1	30	40	50
2	35	45	55
3	40	50	60
4	45	55	65
5	50	60	70

Tableau 3.3: Résultats au test de logique pour 15 élèves issus de 3 pays.

performances différentes ou non. L'ANOVA va répondre à la question suivante: Y a-t-il une influence du pays sur la performance à l'examen de logique?

En général, le tableau présenté dans l'exemple 57 sera de la forme

Individu	Niveau 1	Niveau 2	...	Niveau $k$
1	$Y_{11}$	$Y_{12}$	...	$Y_{1k}$
2	$Y_{21}$	$Y_{22}$	...	$Y_{2k}$
:	:	:	:	:
$r$	$Y_{r1}$	$Y_{r2}$	...	$Y_{rk}$

Tableau 3.4: Tableau de données pour l'ANOVA.

#### Remarque

On peut voir le cadre de l'ANOVA ainsi

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, k$$

ou

$$Y_{ij} = \mu + a_j + \varepsilon_{ij}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, k$$

$\mu$  étant la moyenne générale ( $\mu = \frac{1}{k}(\mu_1 + \mu_2 + \dots + \mu_k)$ ), les  $a_j$  étant l'effet du facteur  $f$  et  $\varepsilon_{ij}$  l'effet aléatoire ( $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ ).

Les hypothèses de l'ANOVA se traduisent alors ainsi:

- **Hypothèse nulle:**  $H_0$

$$H_0 : a_1 = a_2 = \dots = a_k = 0.$$

- **Hypothèse alternative:**  $H_1$

$$H_1 : \text{il existe au moins deux } a_j \neq 0.$$

## A) Etude descriptive des données

### 1. Moyenne des $k$ échantillons et moyenne globale

On calcule les moyennes des  $k$  échantillons  $\bar{Y}_j$  et la moyenne des moyennes  $\bar{Y}$

$$\bar{Y}_j = \frac{1}{r} \sum_{i=1}^r Y_{ij}, \quad \bar{Y} = \frac{1}{k} \sum_{j=1}^k \bar{Y}_j = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^k Y_{ij}, \quad n = k \cdot r.$$

### 2. Variabilité intergroupe

Les  $\bar{Y}_j$  ne sont pas identiques. On quantifie la variabilité entre les différents groupes (variabilité intergroupe). Pour ce faire, on calcule le **carré moyen intergroupe**  $CM_{inter}$

$$CM_{inter} = \frac{SC_{inter}}{k - 1},$$

avec  $SC_{inter}$  la somme des carrés (SC) des écarts intergroupe

$$SC_{inter} = r \cdot \sum_{j=1}^k (\bar{Y}_j - \bar{Y})^2.$$

### 3. Variabilité intragroupe

Il y a également une variabilité à l'intérieur de chaque groupe (qui est différente pour chaque groupe). On quantifie la variabilité à l'intérieur des différents groupes (variabilité intragroupe). Pour ce faire, on calcule le **carré moyen intragroupe**  $CM_{intra}$

$$CM_{intra} = \frac{SC_{intra}}{n - k},$$

avec  $SC_{intra}$  la somme des carrés (SC) des écarts intragroupe

$$SC_{intra} = \sum_{j=1}^k \sum_{i=1}^r (Y_{ij} - \bar{Y}_j)^2.$$

## Remarques

1. On a la décomposition de la somme des carrés totale

$$SC_{tot} = \sum_{i=1}^r \sum_j j = 1^k (Y_{ij} - \bar{Y})^2 = SC_{inter} + SC_{intra}.$$

C'est la **relation fondamentale** de l'ANOVA.

2. Cette relation ne s'applique pas aux variabilités

$$CM_{tot} \neq CM_{inter} + CM_{intra}.$$

## B) Test d'hypothèses de l'ANOVA

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu \\ H_1 : \mu_\ell \neq \mu_j \text{ pour au moins un couple } (\ell, j) \\ \text{Niveau } \alpha. \end{cases}$$

- (a) **Statistique du test:** La statistique du test, notée  $F$  est définie par le rapport

$$F = \frac{CM_{inter}}{CM_{intra}}.$$

Sous  $H_0$  on peut montrer que  $F$  suit une loi de Fisher (-Snedecor) à  $(k-1, n-k)$  degrés de liberté, notée  $\mathcal{F}(k-1, n-k)$ .

- (c) **Critère de décision**

$$P(F \geq f_\alpha) = \alpha.$$

- (d) **Règle de décision:** Accepter  $H_0$  si  $f_{obs} < f_\alpha$ . Rejeter  $H_0$  sinon. On peut également appliquer la règle de décision ainsi:

$$\alpha_{obs} = P(F \geq f_{obs}).$$

Si  $\alpha_{obs} < \alpha$ , on rejette  $H_0$ .

## Remarque

En cas de rejet de  $H_0$ , on ne sait pas quelles sont les moyennes qui sont significativement différentes.

## En résumé

Table d'ANOVA pour  $k$  échantillons indépendants

Sources des variations	Somme des carrés	Degrés de liberté	Variabilité	$F$
Entre les groupes	$SC_{inter}$	$k - 1$	$CM_{inter}$	
A l'intérieur des groupes	$SC_{intra}$	$n - k$	$CM_{intra}$	$\frac{CM_{inter}}{CM_{intra}}$

Tableau 3.5: Table d'ANOVA.

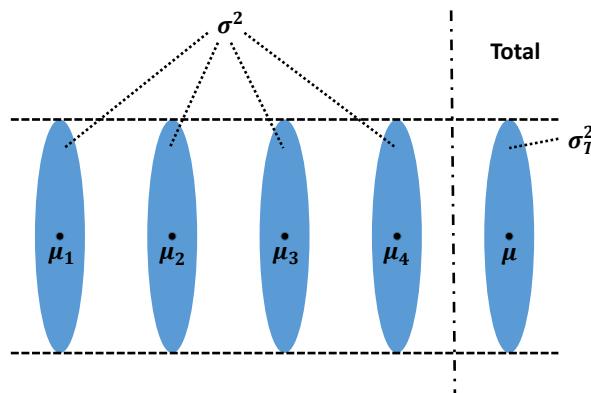
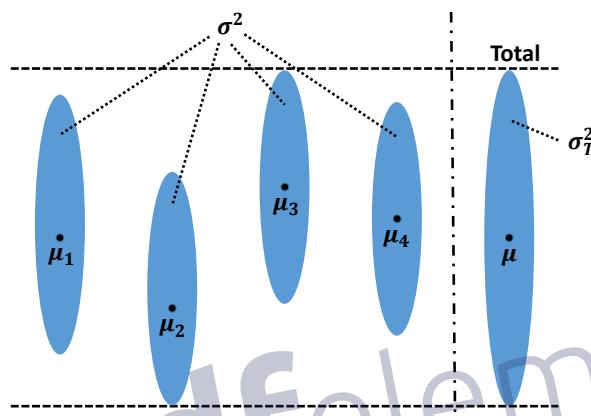
## Calculs

	Taille des groupes <b>identique</b>	Taille des groupes <b>différente</b>
	$n_1 = n_2 = \dots = n_k = r$	$n_1 \neq n_2 \neq \dots \neq n_k, n = n_1 + n_2 + \dots + n_k$
1.	$\bar{Y}_j = \frac{1}{r} \sum_{i=1}^r Y_{ij}$	$T_j = \sum_{i=1}^{n_j} Y_{ij}$
2.	$s_j^2 = \frac{1}{r-1} \sum_{i=1}^r (Y_{ij} - \bar{Y}_j)^2$	$CM = \frac{1}{n} \left( \sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij} \right)^2 = \frac{1}{n} \left( \sum_{j=1}^k T_j \right)^2$
3.	$\bar{Y} = \frac{1}{k} \sum_{j=1}^k \bar{Y}_j$	$SC_{tot} = \sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}^2 - CM$
4.	$SC_{inter} = r \sum_{j=1}^k (\bar{Y}_j - \bar{Y})^2$	$SC_{inter} = \sum_{j=1}^k \frac{T_j^2}{n_j} - CM$
5.	$CM_{inter} = \frac{SC_{inter}}{k-1}$	$CM_{inter} = \frac{SC_{inter}}{k-1}$
6.	$SC_{intra} = \sum_{j=1}^k \sum_{i=1}^r (Y_{ij} - \bar{Y}_j)^2 = (r-1) \sum_{j=1}^k s_j^2$	$SC_{intra} = SC_{tot} - SC_{inter}$
7.	$CM_{intra} = \frac{SC_{intra}}{n-k}$	$CM_{intra} = \frac{SC_{intra}}{n-k}$
8.	$F = \frac{CM_{inter}}{CM_{intra}}$	$F = \frac{CM_{inter}}{CM_{intra}}$

## Principe de l'analyse de variance

Considérons le cas où il y a quatres groupes

- Si  $H_0$  est vraie, les moyennes  $\mu_1, \mu_2, \mu_3, \mu_4$  sont **égales**. La variance totale  $\sigma_T^2$  est **égale** à la variance  $\sigma^2$  de chaque population.
- Si  $H_1$  est vraie, les moyennes  $\mu_1, \mu_2, \mu_3, \mu_4$  sont **differentes**. La variance totale  $\sigma_T^2$  n'est pas **égale** à la variance  $\sigma^2$  de chaque population.

Figure 3.9: Hypothèse  $H_0$  de l'ANOVA.Figure 3.10: Contre-hypothèse  $H_1$  de l'ANOVA.

### 3.5 Introduction au test du $\chi^2$

Il arrive que l'on ne connaisse pas la loi de probabilité d'un phénomène aléatoire (d'une variable aléatoire) ou alors, croyant la connaître, que l'on ait soudain des doutes sur la capacité de cette loi de rendre compte valablement de ce phénomène.

On va introduire une **mesure** pour évaluer l'accord plus ou moins bon existant entre la réalité et la loi présumée. On va le faire en comparant des valeurs théoriques espérées d'occurrence de certains événements avec celles effectivement observées dans un échantillon.

On procédera comme dans un test d'hypothèses en formulant une hypothèse  $H_0$  concernant la loi de probabilité du phénomène aléatoire ( $H_1$ : alternative) et en fixant un seuil de signification.

Pour obtenir une statistique sensible à l'écart existant entre les effectifs théoriques moyens et les effectifs de l'échantillon, on procède alors comme suit:

On répartit les valeurs de la **population** en de l'**échantillon** que l'on va en tirer en  $k$  classes disjointes. Supposons que l'échantillon (de taille  $n$ ) livre pour ces classes  $o_1, o_2, \dots, o_k$  éléments (observés) alors que la distribution de la population (sous  $H_0$ ) permettait d'en prévoir  $e_1, e_2, \dots, e_k$ .

Quel est l'écart entre les effectifs théoriques attendus et ceux observés dans l'échantillon?

$$Q = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

est une mesure de l'écart existant entre les effectifs théoriques attendus et ceux observés dans l'échantillon.  $Q$  sera d'autant plus grand que ce désaccord sera important.

On peut montrer que  $Q$  suit la loi du  $\chi^2$  à  $k - 1$  degrés de liberté lorsque la taille de l'échantillon tend vers l'infini. On peut considérer que tel est le cas si chaque classe comporte au moins 5 valeurs espérées (théoriques).

**Exemple 58** *On a lancé un dé 90 fois et obtenu pour les issues 1 à 6 les effectifs suivants:*

Issue	Effectifs
1	12
2	16
3	20
4	11
5	13
6	18

*Si le dé est régulier ( $H_0$ ), les effectifs moyens attendus sont tous de 15. Ainsi*

$$Q = \chi^2 = \frac{(12 - 15)^2}{15} + \frac{(16 - 15)^2}{15} + \dots + \frac{(18 - 15)^2}{15} = \frac{64}{15} = 4.266.$$

*Or  $P(\chi^2 \geq 11.1) = 0.05$  pour  $k - 1 = 5$  degrés de liberté. La zone critique, ou seuil de signification de 5% est donc formée des  $\chi^2$  tels que  $\chi^2 \geq 11.1$ .*

*La valeur de 4.266 obtenue pour  $\chi^2$  ne se trouve pas dans cette zone, on ne peut donc pas rejeter  $H_0$  au seuil de 5%. On n'a donc aucune raison de soupçonner la régularité du dé.*

# Chapitre 4

## Annexes

### 4.1 Moyenne et variance de quelques lois

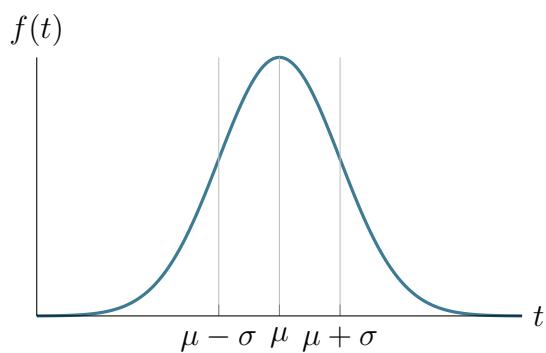
Loi	moyenne	variance
Binomiale $\mathcal{B}(n; p)$	$np$	$np(1 - p)$
Poisson $\mathcal{P}(\lambda)$	$\lambda$	$\lambda$
Uniforme $\mathcal{U}(a; b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentielle $\mathcal{E}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normale $\mathcal{N}(\mu; \sigma^2)$	$\mu$	$\sigma^2$
Normale centrée réduite $\mathcal{N}(0; 1)$	0	1

### 4.2 Loi normale de Laplace-Gauss

On dit que  $X$  suit une loi normale de moyenne  $\mu$  et d'écart type  $\sigma$ , notée  $\mathcal{N}(\mu; \sigma^2)$  si sa densité est

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}.$$

Il n'existe pas de forme analytique pour sa fonction de répartition  $F(x) = \int_{-\infty}^x f(t)dt$



### 4.2.1 Loi normale centrée réduite

Toute loi normale peut être ramenée à une loi normale de moyenne 0 et d'écart type 1, notée  $\mathcal{N}(0; 1)$ , moyennant le changement de variable

$$Z = \frac{X - \mu}{\sigma}.$$

Il n'existe pas de forme analytique pour sa fonction de répartition

$$\Phi(u) = P(Z \leq u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{t^2}{2}} dt.$$

Propriétés

1.

$$P(a < X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

2.

$$P(Z \leq -u) = \Phi(-u) = 1 - \Phi(u)$$

3.

$$P(-u < Z \leq u) = 2\Phi(u) - 1$$

Pour les calculs, on peut utiliser une table si on ne possède pas d'ordinateur ou de machine à calculer programmable.

### 4.2.2 Fonction de répartition $\Phi$ de la loi normale $\mathcal{N}(0; 1)$



#### 4.2.3 Quantiles de la loi normale $\mathcal{N}(0; 1)$

$P(Z > u)$	$P(Z \leq u)$	$P(-u < Z \leq u)$	$u$
0.400	0.600	0.200	0.2534
0.300	0.700	0.400	0.5244
0.250	0.750	0.500	0.6745
0.200	0.800	0.600	0.8416
0.100	0.900	0.800	1.2816
0.050	0.950	0.900	1.6449
0.025	0.975	0.950	1.9600
0.010	0.990	0.980	2.3263
0.0050	0.9950	0.9900	2.5758
0.0025	0.9975	0.9950	2.8070
0.0010	0.9990	0.9980	3.0902
0.0005	0.9995	0.9990	3.2906
0.0001	0.9999	0.9998	3.7191

### 4.3 Loi du $T$ de Student-Fisher

Avec  $k$  degrés de liberté, valeur de  $t$  pour lesquelles

$$P(T_k \geq t) = \alpha \quad \text{ou} \quad P(T_k < t) = 1 - \alpha.$$

$k \backslash \alpha$	0.25	0.2	0.15	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
$\infty$	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291
$1 - \alpha$	0.75	0.8	0.85	0.90	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995

## 4.4 Loi du $\chi^2$ de Pearson

Avec  $k$  degrés de liberté, valeur de  $x$  pour lesquelles

$$P(\chi_k^2 \geq x) = \alpha \quad \text{ou} \quad P(\chi_k^2 < x) = 1 - \alpha.$$

$k \backslash \alpha$	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.7055	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.6052	5.9915	7.3778	9.2103	10.597
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.2514	7.8147	9.3484	11.345	12.838
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.7794	9.4877	11.143	13.277	14.860
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.2364	11.071	12.833	15.086	16.750
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.812	18.548
7	0.9893	1.2390	1.6899	2.1673	2.8331	12.017	14.067	16.013	18.475	20.278
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.362	15.507	17.535	20.090	21.955
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.684	16.919	19.023	21.666	23.589
10	2.1559	2.5582	3.2470	3.9403	4.8652	15.987	18.307	20.483	23.209	25.188
11	2.6032	3.0535	3.8157	4.5748	5.5778	17.275	19.675	21.920	24.725	26.757
12	3.0738	3.5706	4.4038	5.2260	6.3038	18.549	21.026	23.337	26.217	28.300
13	3.5650	4.1069	5.0088	5.8919	7.0415	19.812	22.362	24.736	27.688	29.820
14	4.0747	4.6604	5.6287	6.5706	7.7895	21.064	23.685	26.119	29.141	31.319
15	4.6009	5.2293	6.2621	7.2609	8.5468	22.307	24.996	27.488	30.578	32.801
16	5.1422	5.8122	6.9077	7.9616	9.3122	23.542	26.296	28.845	32.000	34.267
17	5.6972	6.4078	7.5642	8.6718	10.085	24.769	27.587	30.191	33.409	35.719
18	6.2648	7.0149	8.2307	9.3905	10.865	25.989	28.869	31.526	34.805	37.157
19	6.8440	7.6327	8.9065	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.4338	8.2604	9.5908	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.0337	8.8972	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.6427	9.5425	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.2604	10.196	11.689	13.091	14.848	32.007	35.173	38.076	41.638	44.181
24	9.8862	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.653	40.647	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.088	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.759	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.535	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.392	64.278	96.578	101.880	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.930	82.358	118.498	124.342	129.561	135.807	140.170