

# Cours de mise à niveau de statistique

## Chapitre 1

### Population – Variables

# Préalables à une étude statistique

## Quelle(s) méthode(s) statistique(s) utiliser ?

**Choix d'une méthode** : orienté par le type de problèmes abordés, la nature des données recueillies sur les individus.

**Essentiel** : réflexion préalable sur les objectifs de l'étude et le type de mesures à effectuer.

- ▶ *Objectifs de l'étude* : formuler précisément les questions auxquelles on veut répondre.
- ▶ *Population étudiée* : définir les individus sur lesquels porte l'étude.
- ▶ *Mesures à effectuer* : préciser la nature et les échelles de mesure des variables prises en compte dans l'étude.
- ▶ *Echantillons* : préciser le nombre d'échantillons à analyser, leur taille respective. Préciser s'ils sont indépendants ou appariés.

# I Contexte d'une étude

## Exemple 1– Traumatisme

Un chercheur en psychologie clinique<sup>1</sup> s'est intéressé à l'efficacité de 3 thérapies cognitivo-comportementales à court-terme (*TCC*) chez des enfants de 6 à 16 ans manifestant un Trouble de Stress Post Traumatique (*TSPT*) suite à un accident domestique ou de la circulation.

La sévérité du traumatisme est mesurée par le **score TSPT** : score total obtenu à l'aide d'un questionnaire en 10 questions, chaque question étant cotée de 0 à 5. Le score TSPT est d'autant plus élevé que la sévérité est plus grande.

---

1. Source : [www.stat.ucl.ac.be/cours/stat2430](http://www.stat.ucl.ac.be/cours/stat2430) ; *Etude de cas : traumatisme*.

## Exemple 1– Traumatisme

Au départ de cette étude, il a constitué un échantillon de 50 enfants de 6 à 16 ans souffrant d'un *TSPT* à la suite d'un accident domestique ou de la circulation.

Il a ensuite réparti de façon aléatoire les enfants dans 4 sous-échantillons.

Tous les enfants d'un même groupe ont bénéficié du même traitement, le groupe n° 4 étant un groupe de contrôle :

- ▶ Groupe n° 1 (taille 12) : TCC n° 1 ; intervention uniquement avec l'enfant ;
- ▶ Groupe n° 2, (taille 12) : TCC n° 2 ; intervention uniquement avec les parents ;
- ▶ Groupe n° 3, (taille 12) : TCC n° 3 intervention avec l'enfant et les parents ;
- ▶ Groupe n° 4, (taille 14) : soutien de la famille et des proches uniquement.

# 1) Population ciblée par l'étude

## Exemple 1–Traumatisme

- ▶ **Population  $\mathcal{P}$**  : l'ensemble de tous les enfants de 6 à 16 ans souffrant d'un *TSPT* à la suite d'un accident domestique ou de la circulation.  
Sa **taille  $N$**  est inconnue.  
Un **individu** de la population : un enfant de 6 à 16 ans...
- ▶ **Etude de la population** : on ne peut pas étudier l'efficacité des *TCC* sur la population toute entière.
  - ☛ On prélève un échantillon d'enfants dans cette population.
- ▶ **Echantillon** d'enfants issu de la population.  
Sa taille  $n = 50$  est fixée par le chercheur.
  - ☛ Les observations sont faites sur cet échantillon.

## 2) Variables

### Exemple 1. Traumatisme

On étudie différents caractères communs à tous les enfants de la population :

- ▶ sexe, âge au moment de l'accident ; type d'accident ;
- ▶ niveaux de stress avant traitement, après traitement ;
- ▶ ...

Chaque caractère doit être *évalué* sur les enfants à l'aide d'une **variable**.

#### Notations générales utilisées pour une variable :

$X$  (majuscule) : la variable.

$x$  (minuscule) : la *valeur* de la variable sur un individu.

$X$  : « ensemble de toutes les *valeurs individuelles* ».

$x_i$  : la *valeur* de l'individu n°  $i$ .

## Exemple 1. Variables introduites dans l'étude

- ① *Sexe* : 1 pour fille et 2 pour garçon ;
- ② *Type d'accident* : 1 pour « *circulation* », 2 pour « *domestique* »
- ③ *Type de thérapie* : n° 1, n° 2, n° 3, n° 4 ;
- ④ *Age* : âge de l'enfant au moment du traumatisme ; [6; 16]
- ⑤ *Délai* : nombre de mois jusqu'au traitement, de 0 à 12 ;
- ⑥  $TSPT_1$  : score TSPT avant le début du traitement, de 0 à 50 ;
- ⑦  $TSPT_2$  : score TSPT à la fin du traitement, de 0 à 50 ;
- ⑧  $TSPT_1 - TSPT_2$  : mesure l'évolution de l'état de stress.

### 3) Type et échelle de mesure des variables à préciser

- ▶ **Variable qualitative** : définit des catégories d'individus. Ses valeurs, appelées *modalités*, codent les catégories.
  - ☛ **Pour un codage numérique** : aucune opération sur les nombres (+, -, x, /) n'est possible.
- ▶ **Variable quantitative** : prend des valeurs numériques.
  - **Variable discrète** : valeurs « isolées », généralement entières et en petit nombre.
  - **Variable continue** : valeurs dans un intervalle de nombres.

#### ☛ Traitement d'une variable discrète

Une variable discrète pourra être traitée comme une variable continue si son nombre de valeurs possibles est important.

Score entre 0 et 10 : traité comme une variable discrète.

Score entre 0 et 50 : traité comme une variable continue.

## Exercice – Exemple 1. Type des variables ?

- |   |   |                                      |                                       |
|---|---|--------------------------------------|---------------------------------------|
| ① | <i>Sexe</i>   | qualitative <input type="checkbox"/> | quantitative <input type="checkbox"/> |
| ② | <i>Type d'accident</i>                                | qualitative <input type="checkbox"/> | quantitative <input type="checkbox"/> |
| ③ | <i>Type de thérapie</i>                               | qualitative <input type="checkbox"/> | quantitative <input type="checkbox"/> |
| ④ | <i>Age</i>  | qualitative <input type="checkbox"/> | quantitative <input type="checkbox"/> |
| ⑤ | <i>Délai</i>  | qualitative <input type="checkbox"/> | quantitative <input type="checkbox"/> |
| ⑥ | <i>Score TSPT<sub>1</sub></i>                         | qualitative <input type="checkbox"/> | quantitative <input type="checkbox"/> |
| ⑦ | <i>Score TSPT<sub>2</sub></i>                         | qualitative <input type="checkbox"/> | quantitative <input type="checkbox"/> |
| ⑧ | <i>Différence TSPT<sub>1</sub> – TSPT<sub>2</sub></i> | qualitative <input type="checkbox"/> | quantitative <input type="checkbox"/> |

## Exercice – Exemple 1. Variables quantitatives traitées comme des variables discrètes ou continues ?

- |   |  |                                   |                                   |
|---|--|-----------------------------------|-----------------------------------|
| ④ | <i>Age</i>                                     | discrète <input type="checkbox"/> | continue <input type="checkbox"/> |
| ⑤ | <i>Délai</i>                                   | discrète <input type="checkbox"/> | continue <input type="checkbox"/> |
| ⑥ | <i>Score <math>TSPT_1</math></i>               | discrète <input type="checkbox"/> | continue <input type="checkbox"/> |
| ⑦ | <i>Score <math>TSPT_2</math></i>               | discrète <input type="checkbox"/> | continue <input type="checkbox"/> |
| ⑧ | <i>Différence <math>TSPT_1 - TSPT_2</math></i> | discrète <input type="checkbox"/> | continue <input type="checkbox"/> |

## II Propriétés globales d'une population

### 1) Loi et paramètres d'une variable

On étudie un caractère commun aux individus de  $\mathcal{P}$ .  
Le caractère est « mesuré » par une variable  $X$ .

#### 👉 Objectif de la statistique :

Etablir les propriétés générales de l'ensemble des individus pour le caractère étudié.

- ▶ La loi (ou distribution) de la variable donne une description globale de  $\mathcal{P}$ .  
Elle décrit la **variabilité** des « mesures » individuelles dans  $\mathcal{P}$ .
  - ▶ On la résume par des paramètres : caractéristiques de  $\mathcal{P}$ .
- ① **Distribution** : Répartition en pourcentages (*proportions*) des individus selon les valeurs possibles de la variable.
  - ② **Paramètres** : résumant différents aspects de la distribution.

## 2) Description globale d'une population.

Illustration sur un exemple.

### Exemple 2. Notes au baccalauréat

On dispose<sup>a</sup> des notes trimestrielles et des notes au bac, en mathématique, physique et philosophie, de l'ensemble des 909 élèves ayant passé les épreuves écrites du bac C en 1989 dans un même centre d'examen.

- ▶ Population  $\mathcal{P}$  étudiée : l'ensemble de ces élèves.
- ▶ Population de taille  $N = 909$ .
- ▶ Plusieurs variables : notes trimestrielles et au bac pour 3 matières différentes.

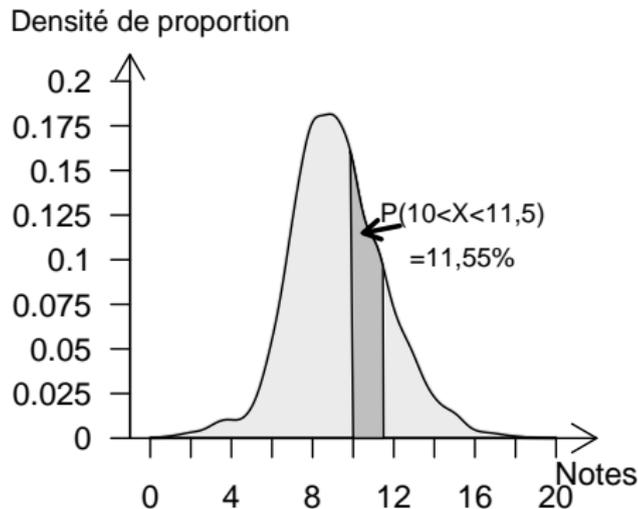
Variables quantitatives, traitées comme des variables continues.

---

a. B. Escoffié et J.L. Pages. *Initiation aux traitements statistiques, méthodes, méthodologie*. Presses universitaires de Rennes. 1997

## a) Distributions. Représentations graphiques.

### Variable ① Note de philosophie du trimestre 3

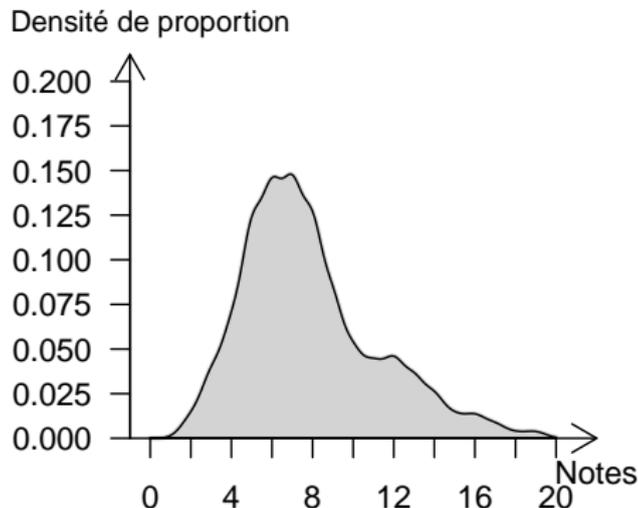


- ▶ Distribution continue
- ▶ Proportion :  
aire d'une surface  
sous la courbe.

Caractéristiques de forme :

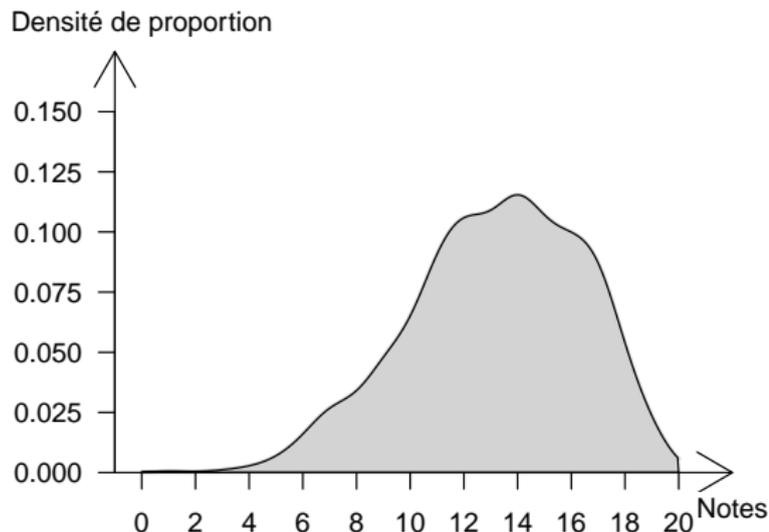
- ▶ Distribution **unimodale** : un seul « pic », prononcé ;
- ▶ Assez **symétrique**, léger étalement à droite.

## Variable ② Note de philosophie du bac



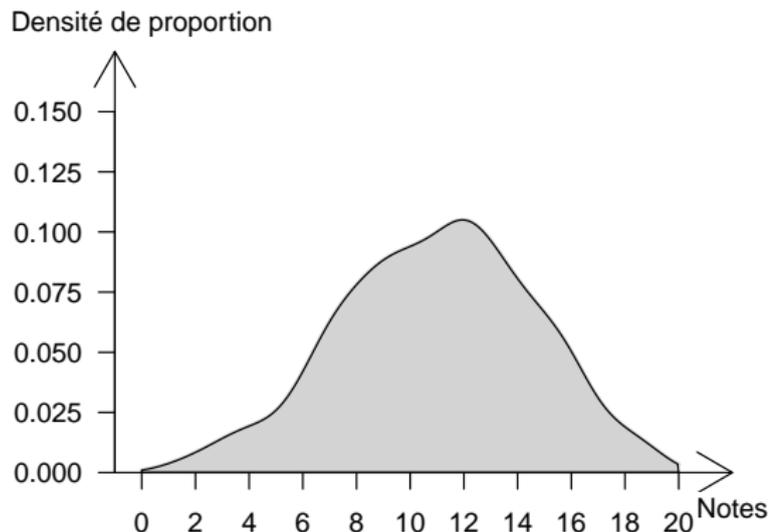
- ▶ Distribution **unimodale** ; pic moins haut ;
- ▶ **Asymétrique**, avec un étalement à droite prononcé.

## Variable ③ Note de mathématique du bac



- ▶ Distribution **unimodale**, plus aplatie que les deux premières.
- ▶ **Asymétrique**, avec un étalement à gauche.

## Variable ④ Note de physique du bac



- ▶ Distribution **unimodale** ; aplatie.
- ▶ Allure **symétrique**.

## b) Paramètres standard d'une variable quantitative

### Moyenne, variance et écart-type

Paramètres usuels résumant divers aspects de la distribution des notes.

#### Note de philosophie du bac

- ▶ **Paramètre de position** : la moyenne  $\mu = 7,84$ .
    - ☛ la moyenne  $\mu$  définit une valeur « centrale » de la population.
  - ▶ **Paramètres de dispersion** : la variance  $\sigma^2 = 10,80$  et l'écart-type  $\sigma = \sqrt{\sigma^2} = 3,28$ .
    - ☛ Ils mesurent la dispersion des notes autour de  $\mu$ .
    - ☛ Indices de « variabilité moyenne » des notes.
- La moyenne et l'écart-type s'expriment dans la même unité de mesure que la variable.

## Moyennes et écart-types des variables notes

- ▶ **Variable** ① Note de philosophie du trimestre 3 :  
moyenne **9,26** et écart-type **2,33**.
- ▶ **Variable** ② Note de philosophie du bac :  
moyenne **7,84** et écart-type **3,28**
- ▶ **Variable** ③ Note de physique du bac :  
moyenne **11** et écart-type **3,61**
- ▶ **Variable** ④ Note de mathématique du bac :  
moyenne **13,21** et écart-type **3,19**.

Quelle variable a :

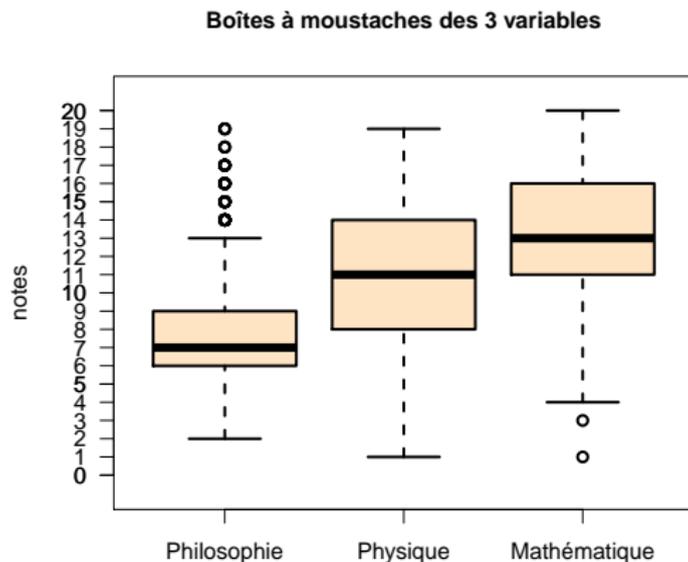
les notes globalement les plus élevées ?

les notes les plus dispersées ?

les notes les plus concentrées ?

## c) Autres paramètres standard : médiane, 1<sup>er</sup> et 3<sup>e</sup> quartiles

### Boîtes à moustaches des trois notes au bac



### Valeur centrale

Médiane : **—**

Quantile d'ordre 50%.

### Mesure de dispersion

Intervalle interquartile :  
limites de la boîte.

Quantiles d'ordre 25% et  
75%.

Contient 50% des notes.

### Valeurs atypiques ○

Valeurs extrêmes, au delà  
des moustaches.

## d) Evaluer la performance d'un élève en particulier

### Calcul de la note centrée réduite

Un élève a eu 9 à l'épreuve de philosophie du bac.

Par rapport à l'ensemble des élèves quelle est sa performance ?

**Connaissant  $\mu$  et  $\sigma$ , on peut calculer sa note centrée réduite.**

- ▶ **Note centrée** : écart entre la note 9 et la moyenne 7,84.

L'élève a une note centrée égale à  $9 - 7,84 = +1,16$ .

Ecart à la moyenne :  $+1,16$ .

Note au dessus de 7,84 : l'écart est positif.

**Cet écart est-il important compte tenu de la dispersion des notes ?**

- ▶ **Note centrée réduite** : on compare son écart à l'écart-type (écart « moyen ») 3,28 :

L'élève a une note centrée réduite égale à  $+\frac{1,16}{3,28} = +0,35$ .

☛ Sa note 9 est proche de la moyenne : son écart représente à peine un peu plus d'un tiers de l'écart-type.

- ▶ **Formule inverse** :  $9 = \mu + (0,35 \times \sigma)$

## d) Evaluer la performance d'un élève en particulier

### Calcul de proportion

On dispose de la totalité des notes de la population.

☛ On peut donc calculer la proportion d'élèves qui ont eu une note meilleure.

- ▶ Après calcul, on obtient  $P(X > 9) = 24,75\%$ .  
Environ un quart des élèves ont eu une meilleure note à l'épreuve.
- ▶ On peut exprimer ce résultat en terme de quantile :  
Sa note correspond au **3<sup>e</sup> quartile** de la distribution.

☛ Sa note n'a rien d'exceptionnel par rapport à l'ensemble des élèves.

### 3) Normalité d'une distribution

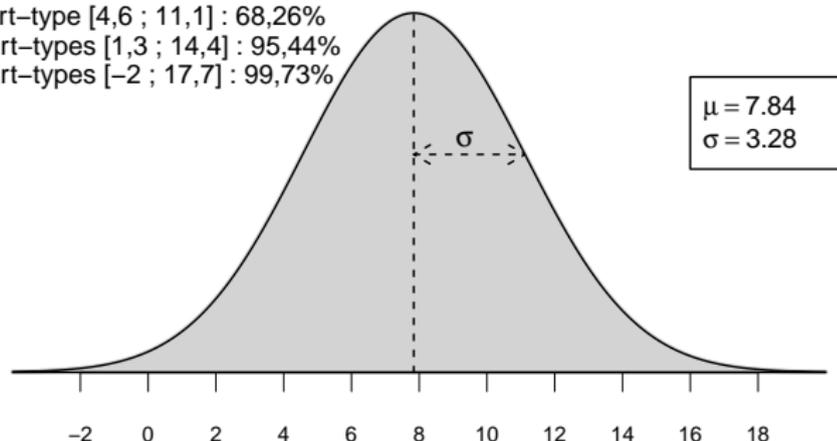
Loi continue de référence : loi normale de moyenne  $\mu$  et d'écart-type  $\sigma$

#### Références de forme :

symétrique, en forme de cloche ; aplatissement « standard ».

Loi  $\mathcal{N}(7,84 ; 3,28)$

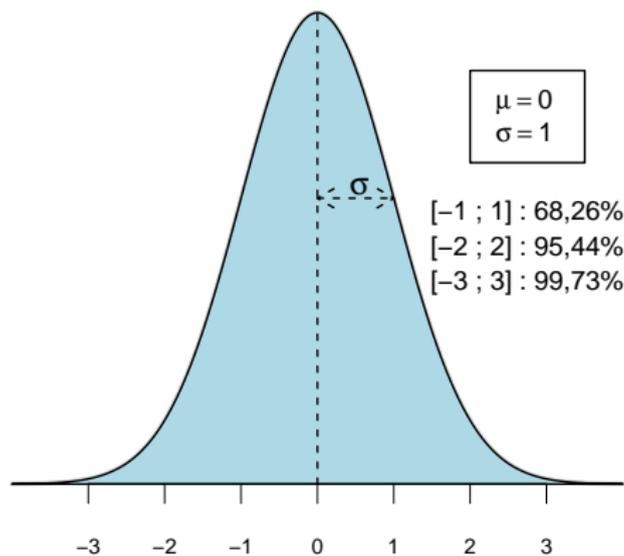
1 écart-type [4,6 ; 11,1] : 68,26%  
2 écart-types [1,3 ; 14,4] : 95,44%  
3 écart-types [-2 ; 17,7] : 99,73%



**Calcul de proportions et de quantiles** : à partir de la loi normale centrée réduite (moyenne 0 et écart-type 1).

# Loi normale centrée réduite : moyenne 0, écart-type 1

Loi  $\mathcal{N}(0; 1)$



# Note de philosophie du bac

Loi continue de référence : la loi normale de moyenne 7,84 et d'écart-type 3,28

La distribution des notes n'est pas normale.

