# Mathématiques

# INFORMATIQUE DU BÂTIMENT

## Objectifs d'apprentissage:

- Compréhension des notions de base du calcul des probabilités et des statistiques
- Aptitude à calculer les probabilités d'occurrence
  Connaissance des distributions de probabilités les plus importantes et leurs grandeurs caractéristiques
- Capacité d'évaluer de manière critique des données statistiques
- Capacité de simuler des expériences de probabilités

# 1. 1 Table des matières

## 2. Introduction aux statistiques

Les statistiques représentent l'enseignement de l'utilisation scientifique des données.

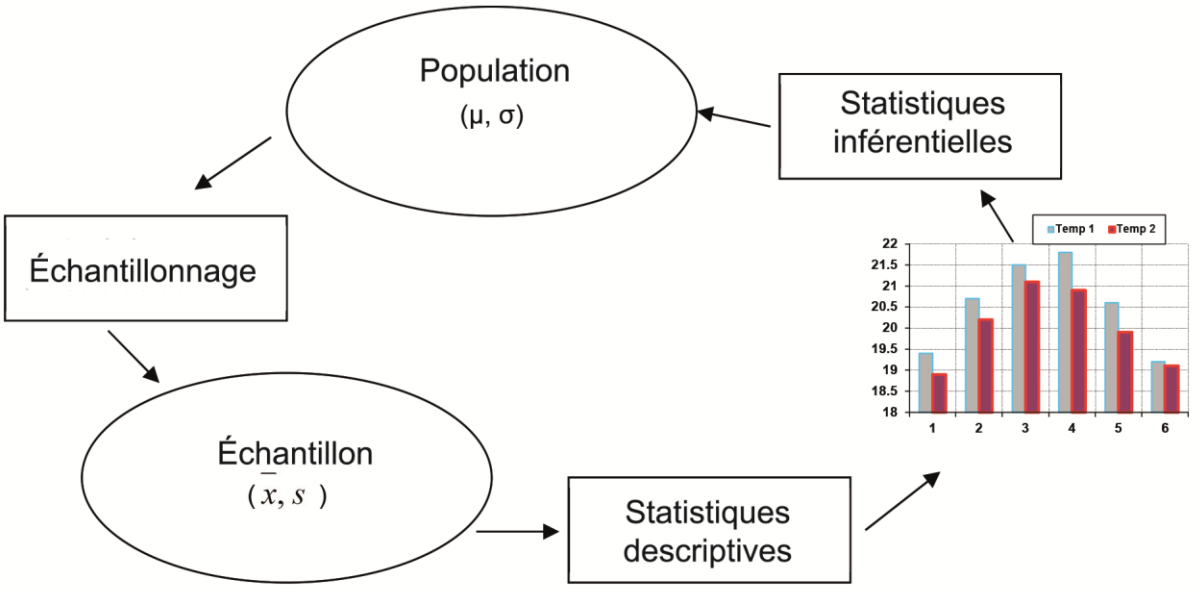
Les statistiques interviennent presque partout, non seulement pour évaluer des données de mesure en sciences ou en économie, mais aussi quotidiennement dans les médias, dans le cadre de statistiques quelconques relatives à la société.
Malheureusement, ces affirmations dans les médias manquent souvent de précision et sont peu parlantes. Pour chaque statistique, il faudrait par exemple indiquer au minimum combien de personnes ont été interrogées / combien de valeurs mesurées ont été saisies, ainsi que la manière dont l'enquête / la mesure a été effectuée, etc.

Les laboratoires, mais aussi des entreprises ou des administrations entières ont besoin de données qui les informent de la situation et des prévisions à l'intérieur et à l'extérieur de l'entreprise. Sans ces données, la planification, la commande et le contrôle d'une entreprise ou d'une administration seraient difficilement possibles.

Les études produisent très souvent d'énormes quantités de données peu claires qui doivent ensuite être traitées au moyen de méthodes statistiques appropriées, afin d'obtenir un nombre restreint mais significatif de paramètres et de graphiques.

Les statistiques sont souvent divisées en domaines partiels:

- échantillonnage: sélection d'un échantillon adapté, représentatif (significatif) pour pouvoir se prononcer sur une „population"
- statistiques descriptives: traitement d'un ensemble relativement important de données, analyse de données, notamment le calcul de chiffres clés et la création de diagrammes correspondants
- statistiques inférentielles (inductives, d'évaluation): formuler des conclusions appropriées pour l'ensemble de la population sur la base des chiffres clés et des graphiques de l'échantillon analysé



Chaque analyse statistique traite d'un «objet», p.ex. d'un nouveau médicament, de l'eau du lac de Zurich, de la population suisse, … .
Cet «objet» est appelé **porteur de caractéristiques**. Il s'agit de l'objet de l'analyse statistique. Il contient les informations qui nous intéressent et qui doivent être examinées de plus près.

La **population** dans laquelle l'examen statistique doit être effectué correspond à la quantité de tous les porteurs de caractéristiques qui possèdent des valeurs concordantes en ce qui concerne les caractéristiques de délimitation, par exemple toute l'eau du lac de Zurich à un moment précis, l'ensemble de la population suisse qui a un certain âge à une date de référence, ...
Or, il n'est généralement pas possible d'étudier l'ensemble de la population (p. ex. l'ensemble des eaux du lac de Zurich), pour des raisons pratiques, mais aussi économiques.

**L'échantillon** est une quantité partielle de la population qui est sélectionnée en fonction de certains aspects. Pour des raisons de coûts, mais aussi de temps, il faut choisir dans la population un échantillon représentatif (aléatoire ou systématique) qui doit autant que possible être une image non biaisée de la population. Dans l'échantillon, toutes les valeurs de caractéristiques d'intérêt devraient donc apparaître avec la même occurrence relative que dans la population. Il est souvent très difficile de trouver un échantillon vraiment représentatif. Il est donc d'autant plus important d'en tenir compte lorsque l'on tire des conclusions sur la population.

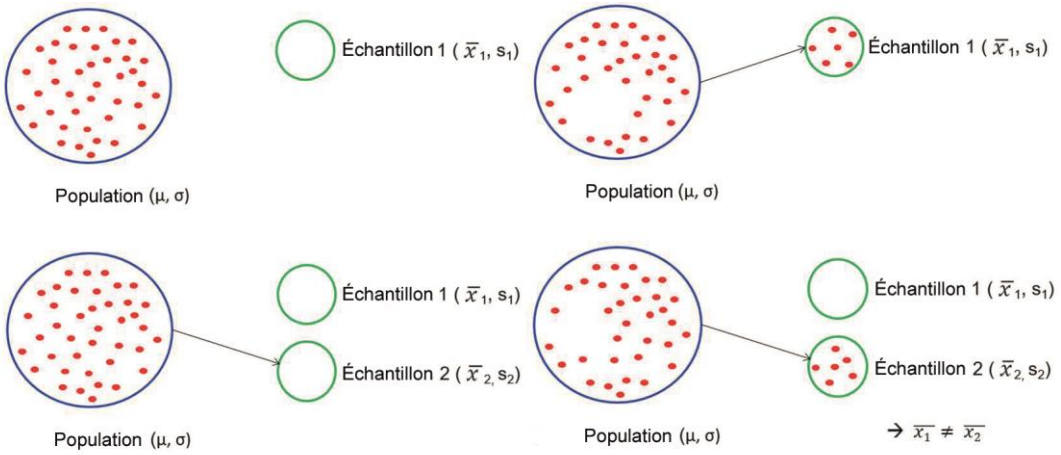
Les attributs qui sont à présent examinés, mesurés, demandés pour ce porteur de caractéristiques sont appelés **caractères** (p. ex. quantité d'un ingrédient spécifique, stabilité dimensionnelle, temps jusqu'à ce que le principe actif est absorbé, teneur en oxygène de l'eau à différentes profondeurs, âge des Suisses, appartenance à différentes nationalités, connaissances linguistiques, …..).

Toutes les valeurs qu'un caractère peut adopter sont appelées les **valeurs caractéristiques**.

Selon les valeurs des caractéristiques, les données doivent être préparées différemment, certains types de caractères ne peuvent être représentés qu'au moyen de graphiques, tandis que pour d'autres types, il est possible de calculer pratiquement toutes les valeurs statistiques. Les "**types de caractères quantitatifs**" (également appelés types de caractères **métriques** ou **numériques**) sont bien adaptés aux analyses statistiques. Il s'agit de caractères pour lesquels des valeurs numériques peuvent être indiquées (p. ex. quantité en µg, temps en s, âge en années, taille en m, ..).
Les caractères quantitatifs peuvent être classés par ordre de grandeur, on peut déterminer différents ratios à leur sujet, on peut estimer dans quelle mesure un échantillon s'écarte significativement de la population ou d'un autre échantillon.
En fonction du nombre de valeurs différentes (valeurs d'expression) pouvant apparaître pour un tel caractère quantitatif, on distingue les caractères quantitatifs continus et les caractères quantitatifs discrèts.

**Les caractères constants** (ou **continus**) peuvent adopter pratiquement toutes les valeurs au sein de certaines limites (p. ex. pour le salaire mensuel – chaque nombre réel au sein de certaines limites).

## Échantillons




Quel est le problème des échantillons?

Problème: les résultats tirés d'échantillons ne peuvent pas être appliqués directement à l'ensemble, car les données de l'échantillon sont empreintes d'une certaine incertitude.

Les moyennes d'échantillonnage sont aléatoires et varient autour de la moyenne réelle de la population μ.

### Exemples de types de caractéristiques quantitatives, continues

| Caractéristique | Valeurs caractéristiques |
|---|---|

**Les caractéristiques discrètes** ne peuvent prendre qu'un nombre limité de valeurs différentes au sein de certaines limites (p. ex. pour le nombre de frères et soeurs, seulement des nombres naturels dans une certaine plage)

**Exemples de types de caractéristiques quantitatives, discrètes**

| Caractéristique | Valeurs caractéristiques |
|---|---|

Outre les types de caractéristiques quantitatives, il existe aussi les **caractéristiques qualitatives**, souvent également désignées en tant que **types de caractéristiques catégorielles**. Leur donnée ne peut pas être saisie sous forme de valeur numérique (p. ex. toxicité, couleur des yeux, profession, nationalité, groupe sanguin, ..).
On ne peut pas calculer grand-chose avec les données des caractéristiques qualitatives, on ne peut les classer que partiellement, on ne peut pas calculer de moyenne, etc., on peut tout au plus compter combien de fois une certaine caractéristique/valeur apparaît.

Ces caractéristiques qualitatives peuvent elles aussi être encore subdivisées, en caractéristiques **ordinales** et **nominales**, en fonction de ce qu'elles expriment.

Les caractéristiques ordinales peuvent également être placées dans un certain ordre ou classées judicieusement.

**Exemples de types de caractéristiques qualitatives, ordinales**

| Caractéristique | Valeurs caractéristiques |
|---|---|

Les caractéristiques nominales ne peuvent pas être placées dans un classement/ordre judicieux.

**Exemples de types de caractéristiques qualitatives, nominales**

| Caractéristique | Valeurs caractéristiques |
|---|---|

## 3. Données d'un échantillon

Il existe un grand nombre de possibilités différentes pour représenter les données recueillies ou pour les résumer avec des chiffres clés individuels.

Les données peuvent par exemple être recueillies par recensements, par expérimentations, par contrôles, par mesures ou par enquêtes.

Ces données recueillies et non traitées sont appelées **données brutes et doivent toujours être stockées / conservées en tant que tels "sans traitement"** ! Souvent, elles sont rassemblées sous la forme d'une liste, dite:

**Liste de référence**

Valeurs de caractéristiques relevées/mesurées, dans l'ordre dans lequel elles ont été relevées/mesurées

Cette liste de référence est généralement d'abord transformée en **liste classée** ou en liste à tirets. Si les valeurs de caractéristiques quantitatives et en partie aussi qualitatives sont classées par ordre de grandeur, il faut veiller à ce que toutes les données supplémentaires soient également converties / classées (en tenir compte dans Excel).

**Liste à tirets**

est possible pour les types de caractéristiques discrètes, mais aussi pour les types de caractéristiques ordinales ou nominales; c'est la représentation graphique la plus simple

*Exemple: Dé*

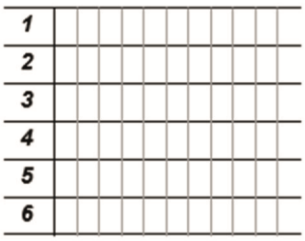
*Liste de référence*

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | |

*Liste classée*

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | |

*Liste à tirets*



## 3.1 Fréquence, histogramme

**Fréquence** indique combien de fois la même valeur est présente dans l'échantillon. La fréquence peut donc être indiquée pour les caractéristiques quantitatives et pour les caractéristiques qualitatives.

On indique la fréquence **absolue** (sous forme de nombre) ou **relative** (sous forme de pourcentage).
On indique également une **fréquence cumulée** absolue ou relative, qui est cependant uniquement judicieuse dans le cadre des caractéristiques quantitatives. La fréquence cumulée est utilisée pour des questions telles que: quelle part de données est plus petite ou égale à une certaine valeur; combien de valeurs mesurées ont une valeur plus élevée qu'une grandeur donnée; quelle part des données analysées fournit des valeurs dans une certaine plage; ...

Ces quatre différentes données de fréquence sont rassemblées dans un tableau de fréquence.

**Étendue** le nombre d'objets/valeurs analysés est appelé „étendue" de l'échantillon (population) et est désigné par **n**.

Ainsi, si l'on additionne les fréquences absolues, la somme sera égale à n ou, dans le cas des fréquences relatives, à 100% (écarts d'arrondi possibles). Dans les colonnes de la fréquence cumulée, l'étendue n ou les 100% pour la fréquence cumulée relative apparaissent également sur la ligne inférieure.

*Exemple:* *Exemple du dé de la page 7 (propre dé / classe entière)*

| | **Fréquence** | | | | **Fréquence cumulée** | | | |
|---|---|---|---|---|---|---|---|---|
| Nombre de points | absolue | | relative | | absolue | | relative | |
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | | | | | | | | |
| | | | | | | | | |
| n = | | | | | | | | |

## Répartition des classes

Si les données de mesure sont des données continues, il n'est pas utile / possible d'indiquer une fréquence pour chaque valeur, car chaque valeur n'apparaît généralement qu'une seule fois et de nombreuses valeurs intermédiaires n'apparaissent pas du tout.
Dans ce cas, on regroupe certaines valeurs dans ce que l'on appelle des classes.

Dans le cadre de la répartition des classes, il faut tenir compte des points suivants:

**quand**: toujours lorsqu'un caractère est présent dans un très grand nombre d'expressions différentes (donc continues); la répartition en **k** classes améliore la vue d'ensemble de l'échantillon entier, c'est-à-dire de toutes les valeurs mesurées

**nombre k**: d'après une règle empirique, il devrait y avoir au moins environ 5 classes mais pas plus de 20 classes, donc
une approche pour des échantillons plus petits dit : (n étendue de l'échantillon)
une approche pour des grands échantillons dit:

**largeur**: toutes les classes doivent avoir la même largeur, cette largeur peut être estimée à l'aide de l'étendue (maximum - minimum) et doit être adaptée de manière à ce que la largeur prenne des valeurs simples.

**milieux**: les milieux des classes sont les valeurs qui sont généralement indiquées dans le diagramme pour la classe, ces milieux doivent donc être des nombres simples et raisonnables ; (par ex. choisir la classe de 2.5 à 7.5, cela donne le milieu à 5 ; c'est mieux qu'une classe de 0 à 5, qui donnerait le milieu à 2.5)

*Exemple:* Un loueur de vélos propose deux modèles différents de vélos électriques. Il les teste tous les deux et roule pour cela 10 fois avec chacun des deux modèles avec une batterie pleine, aussi loin qu'il peut aller. Les autonomies ainsi obtenues sont rassemblées dans le tableau.
Réfléchis à une répartition judicieuse des classes (pour les 20 trajets)

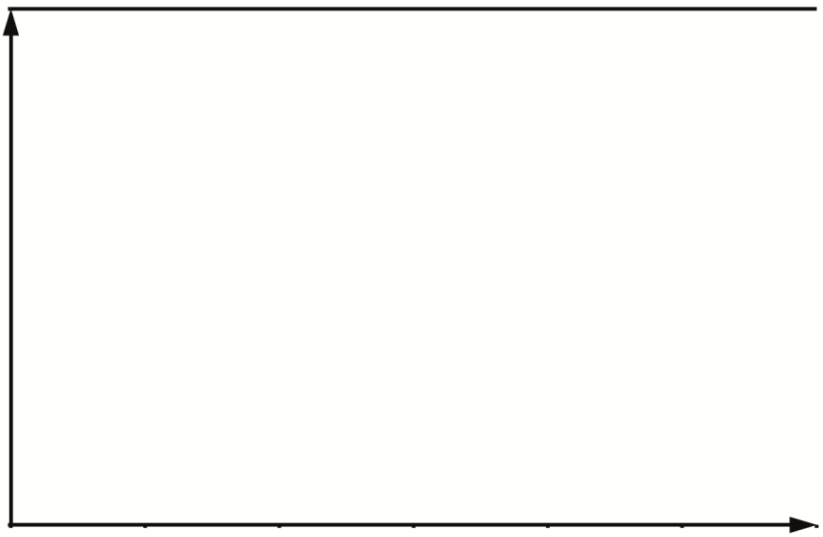
| Modèle A | Modèle B |
|---|---|
| Autonomie en km | Autonomie en km |
| 54.8 | 45.2 |
| 57.4 | 70.1 |
| 55.3 | 52.3 |
| 53.4 | 58.1 |
| 52.6 | 69.5 |
| 55.5 | 37.5 |
| 52.4 | 55.4 |
| 55.7 | 49.7 |
| 57 | 46.6 |
| 54.6 | 90.2 |

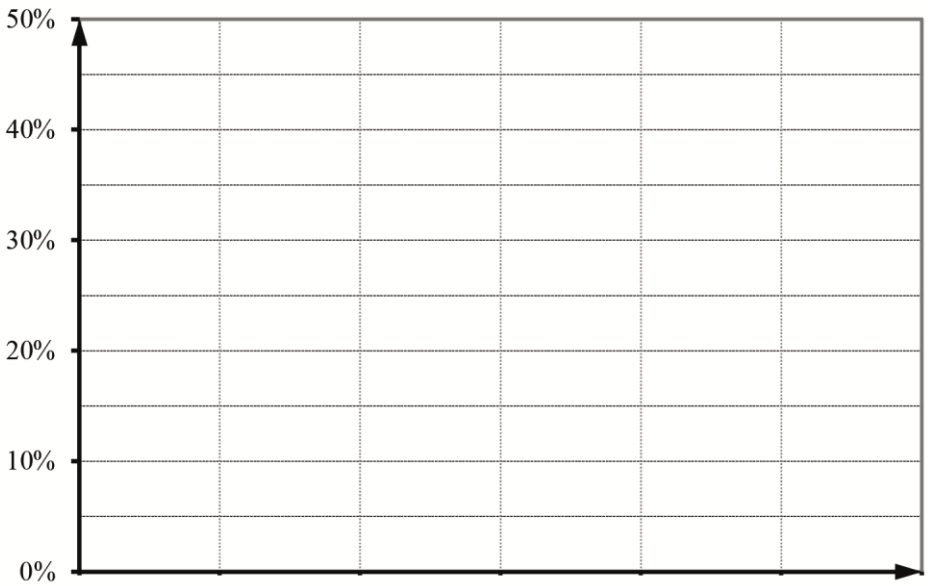
| **Classe** (à partir de jusqu'à et avec) | **Moyenne de classe** | **A + B** | **A** | **B** |
|---|---|---|---|---|

## Histogramme

Les fréquences (relatives et absolues) ne sont pas seulement reportées dans un tableau de fréquences. Elles sont souvent représentées graphiquement avec un diagramme de colonnes, l'histogramme.

*Exemple:* Trace l'histogramme de l'exemple des vélos électriques (A+B), resp. ton exemple avec les dés.





## 3.2 Indicateurs de position

L'ensemble des données observées ou mesurées doit être décrit par un nombre aussi réduit que possible d'indicateurs pertinents. Il est ainsi possible de comparer relativement facilement différentes séries de mesures / échantillons entre eux, sans devoir toujours consulter les listes de référence éventuellement très longues.

Les indicateurs de position permettent de décrire où se situent, à peu de choses près, les valeurs, autrement dit, quel est l'ordre de grandeurs des valeurs mesurées.

Pour pouvoir estimer la plage de données dans son ensemble, les deux valeurs extrêmes donnent rapidement un bon aperçu, c'est-à-dire le maximum et le minimum.

**Minimum** Le minimum $x_{min}$ est la plus petite valeur apparaissant pour la caractéristique analysée.

*Exemple:*

**Maximum** Le maximum $x_{max}$ est la plus grande valeur apparaissant pour la caractéristique analysée.

*Exemple:*

Ces deux valeurs extrêmes peuvent être déterminées très rapidement, mais ne sont pas non plus très parlantes pour indiquer l'emplacement des données. D'une part, elles ne dépendent chacune que d'une seule valeur mesurée, qui peut ne pas être du tout représentative des autres valeurs, et d'autre part, elles sont très fortement dépendantes des "valeurs aberrantes ou extrêmes", c'est-à-dire des valeurs qui s'écartent très fortement de toutes les autres valeurs.

**Médiane** **La médiane** (**ou valeur centrale**) est la valeur au centre d'un ensemble de valeurs ordonnées.
Dans la liste ordonnée ci-dessus, on laisse par conséquent autant de valeurs de côté de part et d'autre, pour qu'il n'y ait plus qu'une seule valeur, c'est-à-dire la valeur qui se trouve au milieu de la liste (valeur centrale). Si on dispose d'un nombre pair de données de mesure, il reste deux valeurs au milieu. On prendra alors pour médiane, la moyenne de ces deux valeurs.
Exprimé de façon générale, on peut dire:
pour un nombre impair de valeurs de mesure, la médiane correspond à la (n+1)/2ème valeur de mesure dans une liste ordonnée, soit

$$\tilde{x} = x_{\frac{n+1}{2}}$$

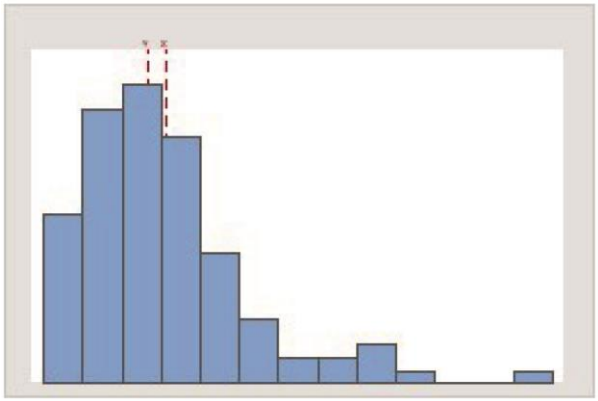
pour un nombre pair de valeurs de mesure, la médiane est définie comme étant la moyenne des deux valeurs centrales de la liste, soit

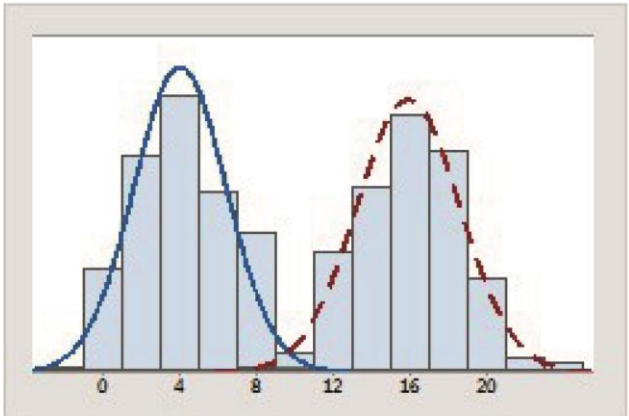
*Exemple:*

$$\tilde{x} = \frac{1}{2} \cdot \left( x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)$$

La médiane est robuste aux valeurs aberrantes. Elle ne se modifie donc pas aussi facilement à cause de ces valeurs extrêmes qui se situent effectivement aux extrémités. Même si ces valeurs extrêmes s'écartent très fortement de toutes les autres valeurs, la médiane reste constante. Elle est donc souvent une bonne mesure de la situation des valeurs dans la série de mesures /l'échantillon.

**Valeur modale** Le mode ou valeur modale **m** (ou également $x_{mod}$) est la valeur qui affiche la fréquence la plus élevée. Il s'agit donc de la valeur de mesure qui apparaît nettement plus souvent que toutes les autres valeurs au sein des valeurs de mesure enregistrées. S'il n'y a pas de valeur dominante, alors la valeur modale n'est pas définie. Il n'existe donc pas de valeur modale dans tous les cas.
La valeur modale est surtout clairement définie lorsqu'une seule valeur est dominante.
Si deux valeurs apparaissent à peu près aussi souvent l'une que l'autre ou nettement plus souvent que toutes les autres, on parle d'une distribution bimodale (distribution avec 2 pics clairs), si même trois valeurs différentes sont représentées un peu à la même fréquence, il s'agit d'une distribution trimodale ou exprimée plus généralement „multimodale".






Quand les valeurs de mesure sont continues, ou quand les valeurs de mesure sont discrètes avec un grand nombre de valeurs différentes, les valeurs sont alors regroupées dans des classes et la valeur modale est alors déterminée de manière simplifiée pour les milieux de classe.

La valeur modale est un bon indicateur, très significatif, surtout pour les distributions multimodales. En effet, avec les autres indicateurs de position, il est plus difficile de dégager une signification s'il y a plusieurs valeurs très fréquentes (plusieurs pics).

*Exemple:*

**Moyenne**

La moyenne également connue sous le nom de moyenne arithmétique, correspond à la somme de toutes les valeurs ($x_i$) divisée par le nombre de valeurs (étendue n)

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

La valeur moyenne est de loin l'indicateur le plus connu et le plus utilisé pour enregistrer la position des valeurs d'un échantillon. Pour calculer la moyenne, toutes les valeurs de mesure sont prises en compte, et pas seulement certaines, comme c'est le cas pour d'autres indicateurs.
La moyenne ne révèle pas toujours des valeurs „judicieuses". P. ex. si l'on considère le nombre moyen de frères et soeurs, la valeur peut s'élever à 1,7 ce qui n'est pas un chiffre représentatif dans ce cas.
La valeur moyenne est sensible aux valeurs aberrantes, autrement dit aux valeurs qui sont très distantes des autres valeurs, raison pour laquelle ce sont les autres indicateurs de position qui sont souvent donnés pour décrire un échantillon (valeur modale et médiane).

*Exemple:*

**Moyenne tronquée**

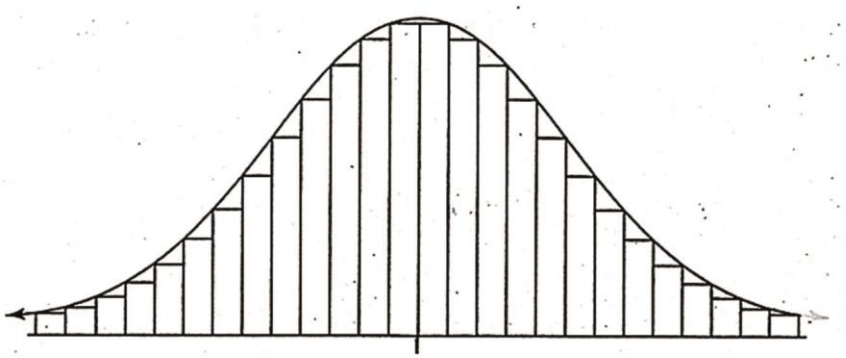
Pour calculer une moyenne tronquée, on prend toutes les données en excluant les 5 % de valeurs les plus élevées (la plupart du temps) et les 5 % de valeurs les plus basses. Ainsi, la moyenne tronquée est moins sensible aux valeurs aberrantes, qu'il s'agisse des valeurs les plus faibles ou les plus élevées.
Si les données contiennent des valeurs aberrantes, la moyenne tronquée peut être une mesure plus appropriée que la moyenne. Il ne faut cependant pas "vendre" la moyenne tronquée comme une "moyenne normale", car elle a été établie à partir d'un groupe de données restreint.

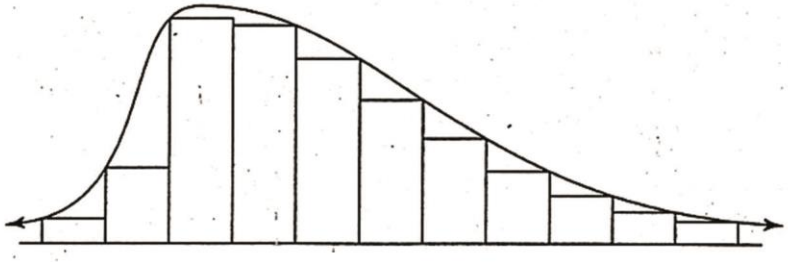
*Exemple:*

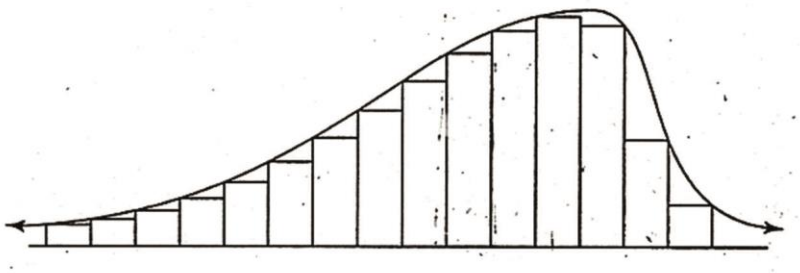
## Comparaison des trois indicateurs de position

Si ces trois valeurs, soient la moyenne, la valeur modale et la médiane, sont proches l'une de l'autre, alors il y a une distribution symétrique, si elles sont très différentes, alors on est en présence d'une distribution oblique.

Estime, pour les 3 distributions indiquées, où se situent approximativement la moyenne, la médiane et la valeur modale (les fréquences absolues des valeurs ou des classes sont représentées dans un histogramme).







**Comparaison des différents indicateurs de position**

| | Minimum | Maximum | Médiane | Valeur modale | Moyenne |
|---|---|---|---|---|---|
| déterminable pour les caractéristiques **quantitatives** | | | | | |
| déterminable pour les caractéristiques **qualitatives** | | | | | |
| Effort pour déterminer le chiffre clé | | | | | |
| Nombre de valeurs prises en compte | | | | | |
| Sensibilité par rapport aux valeurs qui s'écartent très fortement | | | | | |

| | |
|---|---|
| **La médiane** est adaptée / judicieuse pour ... | |
| **La valeur modale** est adaptée / judicieuse pour ... | |
| **La moyenne** est adaptée / judicieuse pour ... | |

Jusqu'à présent, une seule valeur était toujours spécifiée pour décrire la position des données. Il est toutefois plus parlant d'utiliser plusieurs indicateurs pour décrire l'échantillon / la série de mesures. Il peut notamment s'agir de la valeur moyenne/médiane/modale ou la position des données peut aussi être décrite avec les **quartiles** et les **quantiles**.

La médiane répartit l'ensemble des valeurs en deux moitiés, 50% de toutes les valeurs sont inférieures (ou égales) à la médiane et les autres 50% sont supérieures (ou égales).

La première moitié (les 50% de valeurs les plus petites) peut aussi être divisée en deux, on forme alors pratiquement la "médiane de la première moitié", et on peut en faire de même avec l'autre moitié. Au total, l'ensemble des valeurs est ainsi divisé en quatre quarts, on parle alors des différents quartiles.

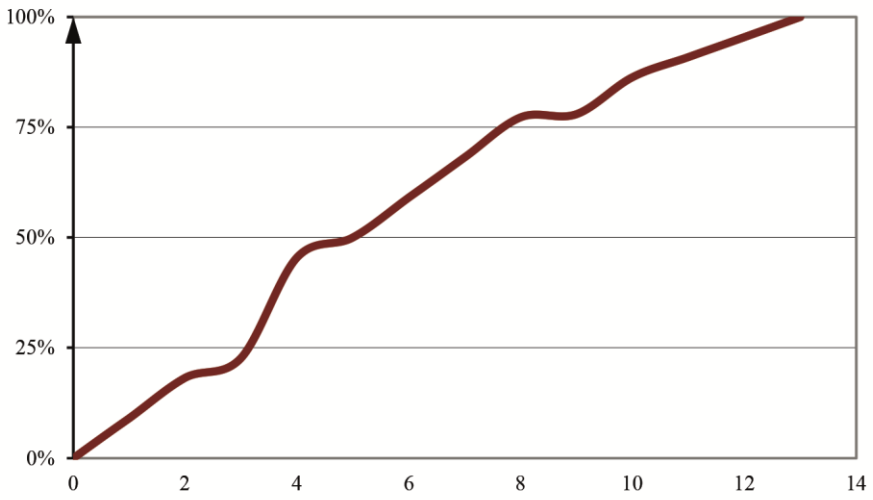
**Quartiles**

**1er quartile $Q_1$** indique la valeur qui sépare les 25% des valeurs les plus petites du reste des valeurs, c'est-à-dire que 25% de toutes les valeurs sont inférieures ou égales à la valeur Q1 (ou 75% sont supérieures)

**3ème quartile $Q_3$** indique la valeur qui sépare les 25% des valeurs les plus grandes du reste des valeurs, c'est-à-dire que 25% de toutes les valeurs sont supérieures à la valeur $Q_3$ , ou 75% de toutes les valeurs sont inférieures ou égales à la valeur $Q_3$

**2ème quartile $Q_2$** indique la valeur qui se trouve au milieu des valeurs (donc à 50%) et correspond donc aussi à la **médiane**

Si l'on trace la "fréquence relative cumulée" pour les valeurs, on obtient assez rapidement les valeurs pour ces différents quartiles.



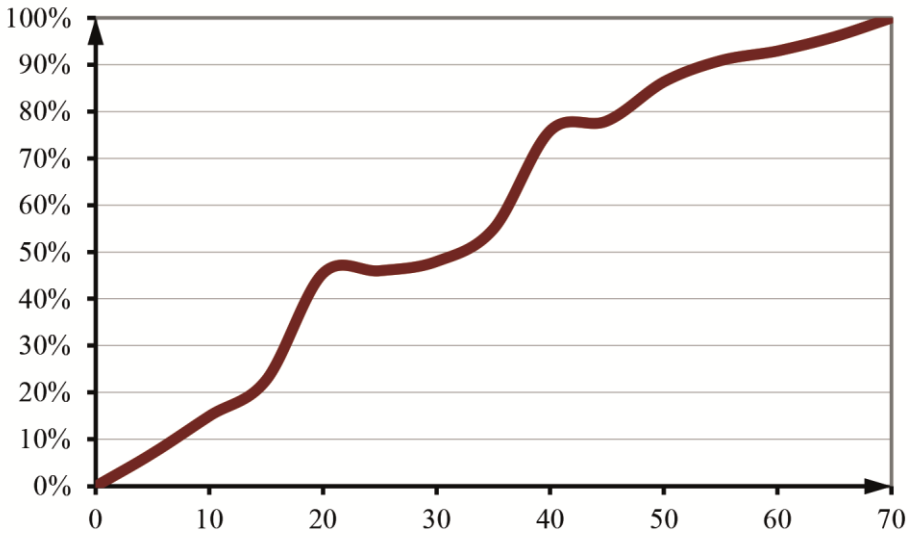

*Exemple:*

Si l'on affine encore davantage la subdivision, on parle alors par exemple de **déciles**, pour une subdivision en dix dixièmes ou de **percentiles**, pour une subdivision en cent centièmes, etc. De manière générale, on définit ces subdivisions plus fines par le terme **quantiles** (terme générique pour les catégories ci-dessus).
Dans ces subdivisions plus fines, les parts les plus grandes et les plus petites sont très souvent intéressantes, par ex. les 5% de valeurs les plus petites parmi toutes les valeurs (quantile 5%), ou les 5% de valeurs les plus grandes parmi toutes les valeurs (quantile 95%).

**Quantile**

Le quantile d'ordre p désigné par $\mathbf{q_p}$ indique la valeur qui sépare les p% de valeurs les plus petites parmi le reste des valeurs, autrement dit les p% parmi toutes les valeurs sont inférieures ou égales à la valeur $q_p$.
Soit par exemple le quantile 5% $\mathbf{q_{0.05}}$. Il indique que 5% de toutes les valeurs sont inférieures ou égales à la valeur $q_{0.05}$, ou exprimé différemment, 95% de toutes les valeurs de l'ensemble sont supérieures (ou égales) à ce quantile 5%.

Le quantile 25% $q_{0.25}$ sera alors égal au 1er quartile $Q_1$, et le quantile 50% $q_{0.5}$ correspond au quartile $Q_2$, ou médiane et ainsi de suite.



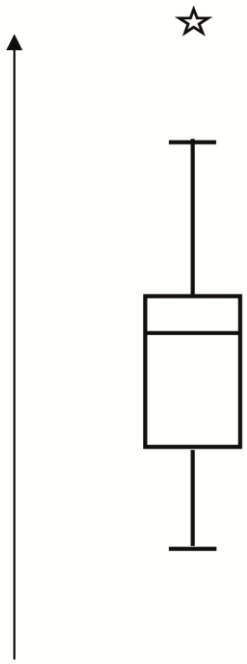

*Exemple:*

## 3.3 Boxplot (diagramme en boîte) ou boîte à moustaches (Whisker - Plot)

**Boxplot**

Il s'agit d'une représentation graphique simple et parlante de l'ensemble des valeurs.
Les données sont rendues compréhensibles par plusieurs caractéristiques et sont parfois encore plus claires que la répartition des fréquences.

Le boxplot, autrement dit le rectangle représenté, va du premier au troisième quartile, donc de $Q_1$ à $Q_3$. La hauteur de la médiane est symbolisée dans le rectangle par un trait horizontal. Les prolongements vers le haut et le bas sont seulement représentés par une ligne qui aboutit généralement au maximum resp. au minimum.
S'il y a des valeurs extrêmes (valeurs aberrantes) qui sont très éloignées de la boîte, elles sont seulement symbolisées par une croix/étoile/cercle …. La ligne de prolongation atteint donc seulement la deuxième plus grande ou deuxième plus petite valeur.
S'il y a beaucoup de valeurs aberrantes, il y aura par conséquent plusieurs croix, les lignes de prologation allant généralement seulement jusqu'aux valeurs comprises au sein d'une bande qui, à partir de la fin de la boîte, a une distance de 1,5 fois le quartile Q3 - Q1.



*Exemple*:

## 3.4 Rang (chiffre indiquant l'ordre)

**Rang**

Le rang, ou chiffre indiquant l'ordre d'une valeur de mesure/d'une observation indique la place de la valeur de mesure dans la liste ordonnée d'un échantillon. En d'autres termes, la valeur la plus petite (minimum) correspond au rang 1, la valeur la plus grande (maximum) correspond au rang n (n = étendue de l'échantillon).

Le rang de la valeur de mesure indique donc également combien de valeurs sont inférieures ou égales à cette valeur de mesure (le rang 8 signifie par conséquent que 8 valeurs de l'échantillon sont inférieures ou égales à cette valeur).

Si certaines valeurs mesurées apparaissent plusieurs fois dans l'échantillon, la valeur moyenne des rangs correspondants est indiquée comme rang pour ces valeurs (p. ex. 4ème, 5ème et 6ème rang ont tous la même valeur de mesure, ces trois valeurs de mesure identiques obtiennent donc toutes les trois le rang 5, il n'y aura donc pas de rang 4 et 6 dans cet échantillon).

*Exemple 1*: L'échantillon avec les valeurs suivantes (liste de référence):

5 / 9 / 4 / 13 / 2 / 4 / 9 / 6 / 10 / 3 / 9 / 8

donne la liste ordonnée ci-dessous:

2 / 3 / 4 / 4 / 5 / 6 / 8 / 9 / 9 / 9 / 10 / 13

et cela donne alors les rangs:

$x = 2 \rightarrow Rang\ 1$
$x = 3 \rightarrow Rang\ 2$
$x = 4 \rightarrow Rang\ 3.5$

$x = 5 \rightarrow Rang\ 5$
$x = 6 \rightarrow Rang\ 6$
$x = 8 \rightarrow Rang\ 7$

$x = 9 \rightarrow Rang\ 9$
$x = 10 \rightarrow Rang\ 11$
$x = 13 \rightarrow Rang\ 12$

*Exemple 2*: Échantillon avec les valeurs:

3 / 6 / 5 / 7 / 3 / 3 / 10 / 5 / 8 / 7 / 2 / 10 / 3 / 11 / 4 / 7

Liste ordonnée:

Rangs:

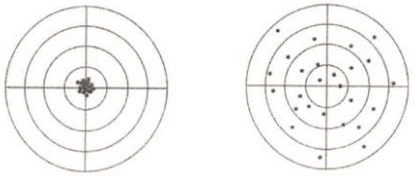
$x = \_\_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_\_ \rightarrow Rang \_\_\_\_\_$

$x = \_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_ \rightarrow Rang \_\_\_\_\_$
$x = \_\_\_\_ \rightarrow Rang \_\_\_\_\_$

## 3.5 Indicateurs de dispersion

La dispersion est une autre caractéristique très importante d'une série de mesures/ de données.
La plupart du temps, il est très intéressant de savoir si les données sont dispersées dans une zone très étroite ou si elles sont réparties sur une très grande zone.
Les indicateurs de dispersion permettent d'évaluer si la répartition est symétrique et où se situe la valeur type/valeur moyenne, mais pas à quel point les valeurs sont dispersées autour de cette moyenne.
Deux ensembles de données peuvent présenter la même valeur moyenne, mais être très différents.

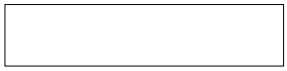
Même moyenne



Petite dispersion Grande dispersion

Petite dispersion Grande dispersion

**<u>Étendue</u>** (largeur de variation, rangs)
L'étendue montre la distance entre les points extrêmes



L'étendue est une mesure de dispersion très simple, mais aussi très grossière, qui ne dépend pas de la taille de l'échantillon et qui est calculée à partir de deux valeurs seulement de l'ensemble des données. C'est pourquoi l'étendue ne donne qu'une impression très vague de la manière dont les données sont réparties, mais elle ne décrit pas la manière dont les valeurs des caractéristiques se dispersent à l'intérieur de cette fourchette. L'étendue réagit fortement aux valeurs aberrantes et n'est donc généralement pas une valeur adaptée pour décrire la dispersion des données.
L'étendue est principalement utilisée pour pouvoir former une répartition judicieuse des classes en cas de caractères répartis de manière continue.

*<u>Exemple:</u>*

**Écart interquartile**

L'écart interquartile correspond à la différence entre le troisième et le premier quartile:

$$QA = IQR = Q_3 - Q_1$$

Dans la plage donnée par cet écart interquartile se trouvent les 50% des valeurs intermédiaires. L'écart interquartile est aussi une valeur relativement simple qui peut être déterminée relativement facilement (s'il existe une liste ordonnée) et qui donne une meilleure image de la dispersion des valeurs.
L'écart interquartile est assez robuste par rapport aux valeurs aberrantes, car seules les valeurs moyennes sont prises en compte, les 25% de valeurs les plus grandes et les 25% de valeurs les plus petites ne sont pas considérées. L'écart interquartile est nécessaire pour le boxplot, mais il est également très approprié comme mesure de la dispersion lorsque c'est surtout la zone moyenne des valeurs qui est déterminante.

*Exemple:*

Si toutes les données doivent être utilisées pour la dispersion et pas seulement des valeurs individuelles comme pour l'étendue ou l'écart interquartile, on peut par exemple calculer les écarts des différentes valeurs mesurées par rapport à la moyenne et additionner toutes ces différences, alors $\sum_{i=1}^{n} \left(x_i - \overline{x}\right)$ , et on obtient ...

Une possibilité d'obtenir une valeur plus judicieuse et plus parlante consiste à additionner le montant des différences, à en faire la moyenne, ce qui donne la valeur d'„écart moyen absolu".
Ainsi, seule la distance par rapport à la valeur moyenne est déterminante, et non pas le fait que la valeur soit supérieure ou inférieure à .
Dans la deuxième possibilité, qui est le plus souvent utilisée, on élève les différences au carré afin que la présentation de la distance ne joue aucun rôle. Cette mise au carré est utilisée pour la variance ou l'écart-type.

**Variance**

**$s^2$** décrit donc l'**écart moyen au carré** des valeurs mesurées par rapport à la moyenne, soit:

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

On divise par (n-1), si l'on se base sur des données provenant d'un échantillon. Par contre, si l'on dispose des données d'une population, il suffit de diviser par n (dans le cas d'une population, l'étendue, c'est-à-dire n, est généralement très grande et la différence entre diviser par exemple par 1'000'000 ou par 1'000'001 est infinitésimale).

*Exemple* : Dés (calcul „manuel“ de la variance/de l’écart-type)

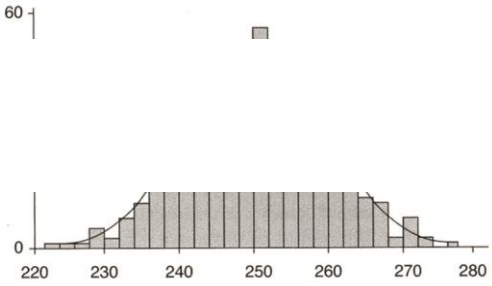
| Nombre de points | Fréquence absolue (ex. 1) | $x_i - \overline{x}$ | $\sum (x_i - \overline{x})^2$ | Fréquence absolue (ex. 2) | $x_i - \overline{x}$ | $\sum (x_i - \overline{x})^2$ |
|---|---|---|---|---|---|---|
| **1** | **5** | | | **21** | | |
| **2** | **2** | | | **8** | | |
| **3** | **6** | | | **4** | | |
| **4** | **4** | | | **9** | | |
| **5** | **1** | | | **9** | | |
| **6** | **2** | | | **9** | | |
| $\overline{x}$ | | ----- | ----- | | ----- | ----- |
| n resp. Σ | | | | | | |
| resp. s | ----- | s = | $s^2$ = | ----- | s = | $s^2$ = |

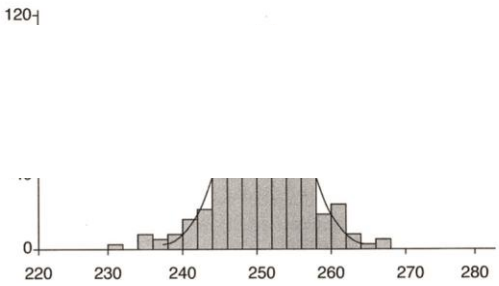
**Écart-type**

Comme les unités de la variance apparaissent également au carré, il est difficile de se représenter cette valeur. C’est pourquoi on tire la racine (toujours positive) de la variance et on obtient ainsi **l’écart-type.**

$$s = +\sqrt{s^2} = +\sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n-1}}$$

L’écart-type a donc les mêmes unités que les valeurs mesurées elles-mêmes et donne une indication sur la dispersion des données. Très souvent, la série de données est alors décrite par les deux valeurs importantes que sont la moyenne et l’écart-type sous la forme:
(ici aussi, pour léchantillon, s est calculé avec n-1, et pour la population, σ est calculé de manière analogue mais avec n)




*Exemple:*

Selon la taille de la moyenne d'une série de données, il n'est pas possible de comparer directement les différents écarts-types des différentes séries de données pour savoir où les données sont proches les unes des autres et où elles sont très dispersées. Comme l'écart-type a la même unité que les valeurs elles-mêmes, on ne peut par exemple pas comparer directement les écarts-types de grandeurs en mètres avec ceux en millimètres. Il n'est pas non plus possible de comparer directement deux séries de mesures avec pour indications resp. . L'écart-type de la deuxième indication est dans ce cas deux fois moins grande, donc les données sont aussi deux fois moins dispersées? Ce n'est évidemment pas le cas ici !

**Coefficient de variation** (= écart-type relatif)

Le coefficient de variation est donné en % et correspond à l'écart-type par rapport à la moyenne, (ce pourquoi il est appelé écart-type relatif).

en %

Il permet de comparer directement différents échantillons en fonction de leur dispersion. (ex. ci-dessus: premières indications donne un v de ..... %, la deuxième indication donne un v de ..... % donc environ trois fois plus grand que pour la première indication, les données de la deuxième série de mesures sont donc nettement plus dispersées que celles de la première)

*Exemple:*

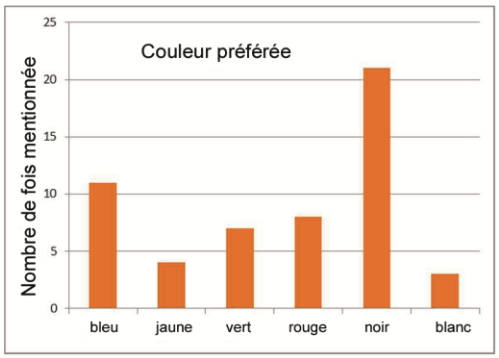
# 4. Représentation graphique

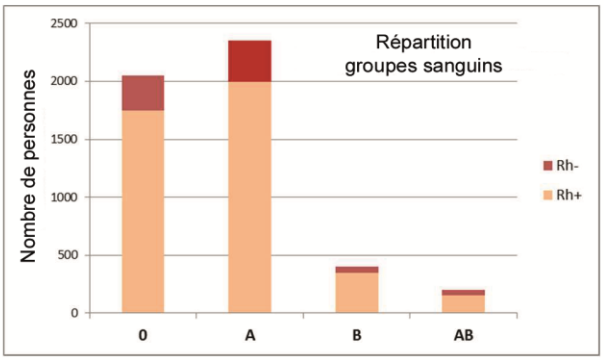
## 4.1 Diagrammes

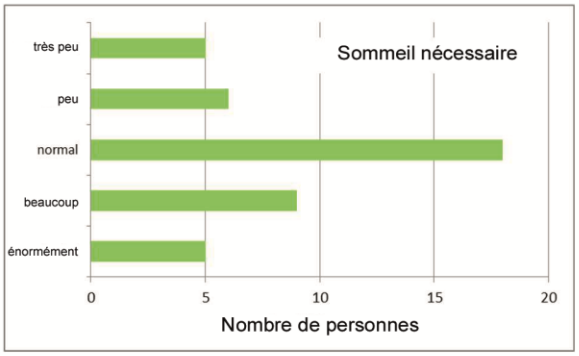
Les représentations graphiques donnent une première impression rapide de la répartition des données. Certaines représentations graphiques ont déjà été expliquées dans les documents précédents.

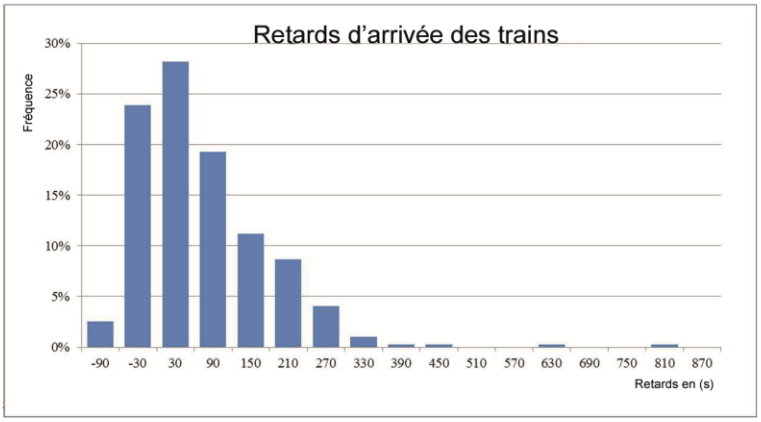
**Liste à tirets** Il s'agit de la représentation graphique la plus simple, qui peut être créée assez rapidement pour des types de caractéristiques discrètes, ordinales et nominales, si les données ne sont pas trop nombreuses.

**Graphique à barres** ou **diagrammes en bâtons** sont utilisés pour visualiser des fréquences absolues ou relatives. Il est possible de représenter des données discrètes, ordinales ou même nominales.
Si les données prennent des valeurs continues (souvent des échantillons à grande étendue), elles sont d'abord regroupées en classes (p. 123), puis représentées sous forme de diagramme en bâtons, appelé **histogramme**.



**Boxplot**

donne une idée de la zone dans laquelle se trouvent les données et de la manière dont elles se répartissent sur cette zone. Le Boxplot réduit les données d'un échantillon à cinq indicateurs ($Q_1$, $Q_2$ = , $Q_3$, $x_{min}$, $x_{max}$) plus les éventuelles valeurs aberrantes. Les Boxplots sont utilisés pour les caractères quantitatifs, avant tout pour les grands échantillons. La représentation au moyen d'un Boxplot convient particulièrement pour comparer plusieurs échantillons entre eux. Cette représentation convient moins aux distributions multimodales, car il est impossible de reconnaître les différents modes dans les boxplots.

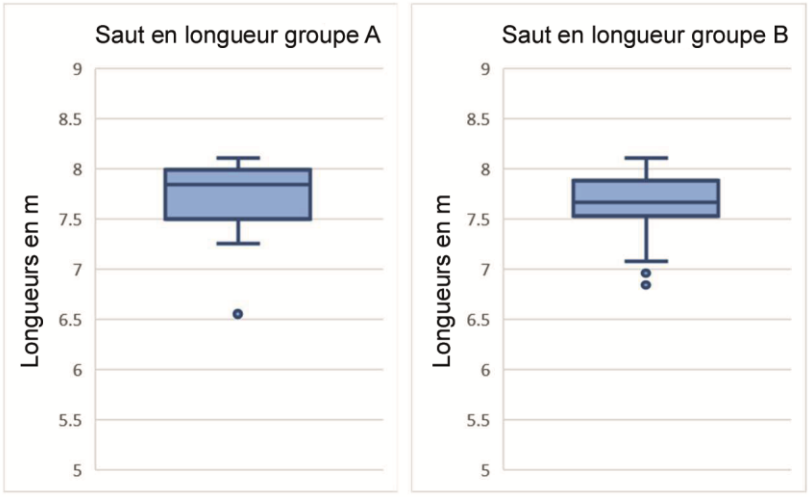
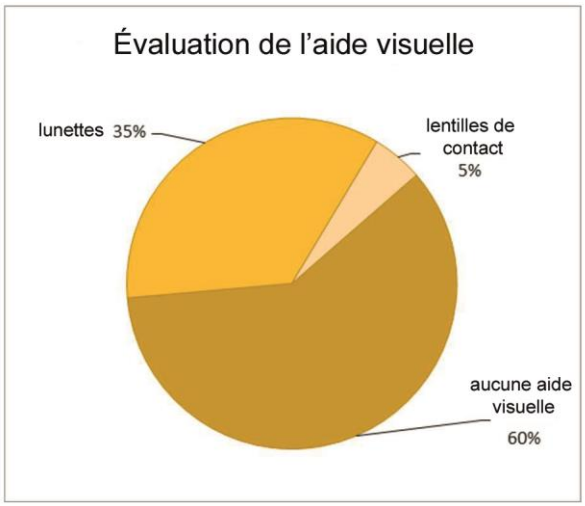
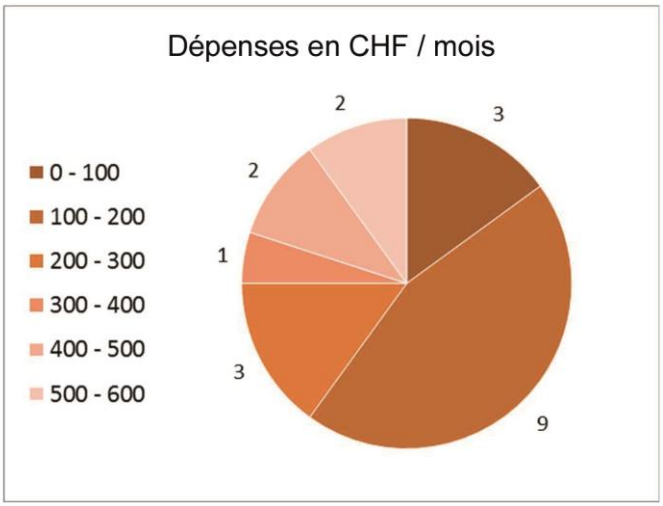


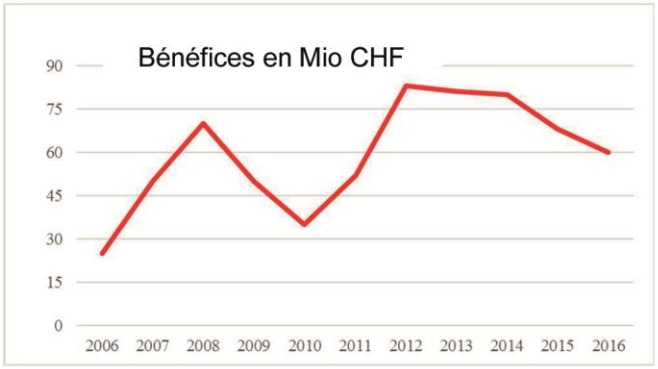

**Diagramme circulaire**

également appelé **diagramme en camembert**. Les diagrammes circulaires peuvent être tracés pour tous les types de caractères (qualitatifs et quantitatifs). Dans un diagramme circulaire, les fréquences relatives des expressions des caractéristiques sont représentées, l'angle du secteur circulaire étant proportionnel à la fréquence relative. La surface du secteur est donc également proportionnelle à la fréquence. Le diagramme circulaire étant une représentation facile à lire, il est très souvent utilisé dans les médias.



**Graphique en ligne** sont particulièrement bien adaptés pour vérifier les tendances dans le temps. Très souvent, les unités de temps telles que les années, les mois, ... sont inscrites dans le diagramme linéaire dans le sens horizontal (axe x) et dans le sens vertical, par exemple, le taux de chômage moyen, le taux de natalité, le prix moyen des loyers, les bénéfices d'une entreprise, etc. Les graphiques linéaires permettent de représenter les variations temporelles de caractères quantitatifs.



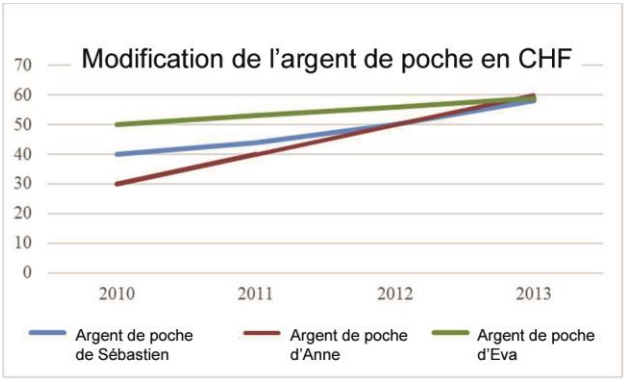
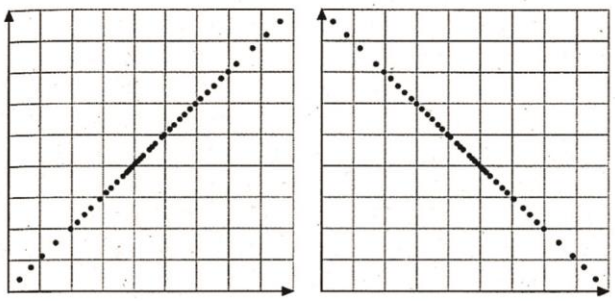
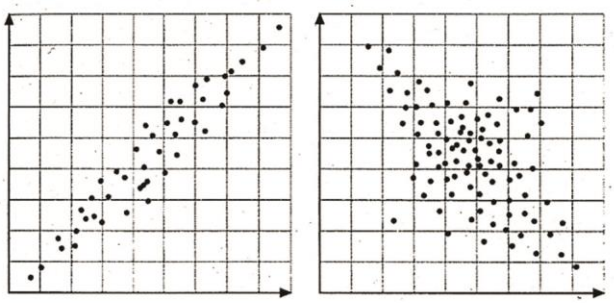
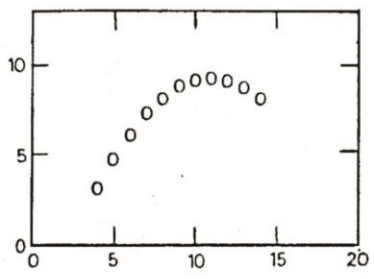



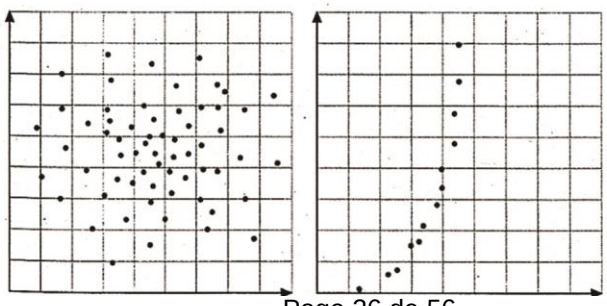

**Diagramme de dispersion**

également appelé **diagramme x-y** ou **nuage de points**. Il est utilisé quand **deux caractères** ( = **données bivariées**) doivent être enregistrés simultanément, c'est-à-dire l'un avec l'autre. Les deux caractères ont des attributs quantitatifs, ce sont donc des valeurs numériques. Normalement, la grandeur "indépendante" est placée dans le sens horizontal, la grandeur qui en dépend dans l'axe vertical. On parle également de caractère de cause et de caractère d'effet qui en découle (p. ex. les deux caractères âge et salaire: plus on est âgé, plus le salaire a tendance à être élevé, donc la cause "âge" sur l'axe x et l'effet "salaire" sur l'axe y). Un point du diagramme x-y est alors attribué à chaque paire de valeurs. Un diagramme de dispersion permet souvent aussi de déterminer s'il existe ou non une corrélation entre les deux caractères. Cette corrélation peut être linéaire (on dit alors que les deux caractéristiques sont **corrélées**) mais aussi carrée ou exponentielle.





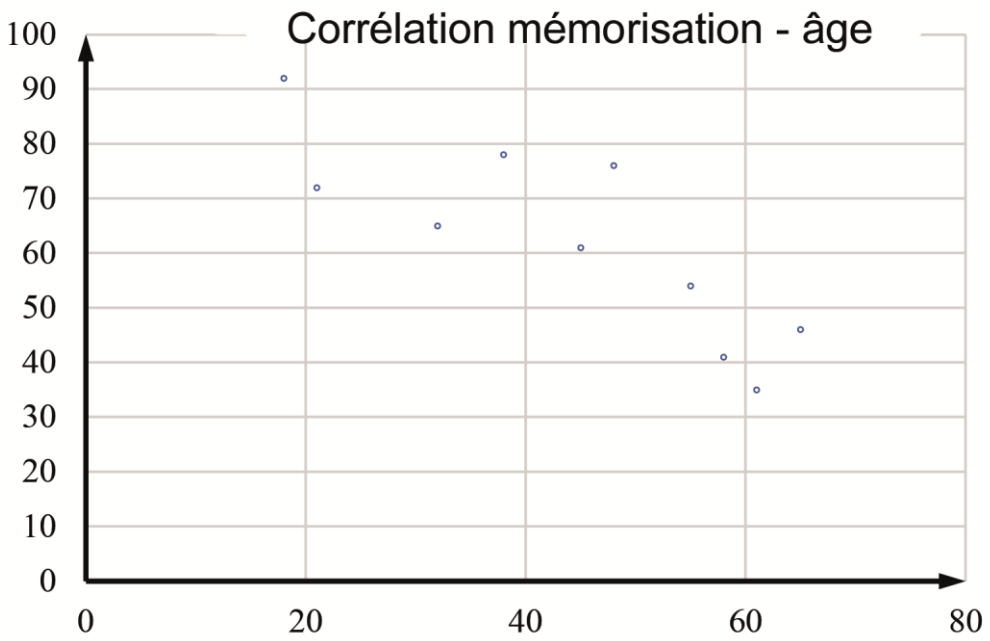


*Exemple:* 10 personnes différentes participent à un test de mémoire.
On consigne simultanément les deux caractères "résultat du test" et "âge" des personnes. On obtient les résultats suivants:

| **Personne** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Résultat / Pts** | 65 | 41 | 54 | 72 | 46 | 92 | 78 | 76 | 35 | 61 |
| **Âge / ans** | 32 | 58 | 55 | 21 | 65 | 18 | 38 | 48 | 61 | 45 |

Représentez la corrélation dans un diagramme de dispersion



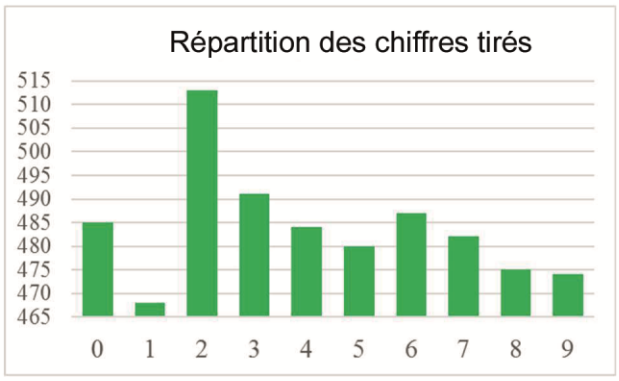

Y-a-t-il un lien entre les deux caractères?
Les deux caractères sont-ils corrélés?

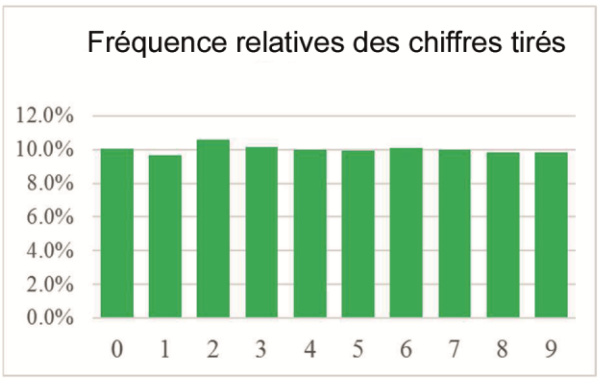
## 4.2 Ce à quoi il faut faire attention dans les diagrammes (influence volontaire!)

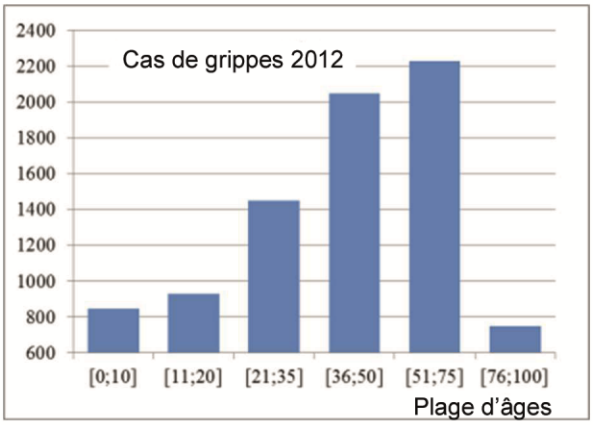
Pour vérifier l'exactitude statistique des graphiques, la première chose à faire est de vérifier que les légendes des axes et du titre sont complètes et claires (unités). En outre, selon le diagramme, certains points doivent être examinés en particulier.

**Graphiques à barres ou diagrammes en bâton**

- si les données ont été regroupées en classes, les largeurs des classes doivent toutes être identiques
- la répartition des classes doit être adaptée aux données (valeurs et taille de l'échantillon)
- l'unité de l'axe vertical (axe horizontal) doit être adaptée aux données ; une échelle inappropriée peut fortement influencer le message du graphique
- il faut indiquer clairement sur l'axe s'il s'agit de fréquences relatives ou absolues

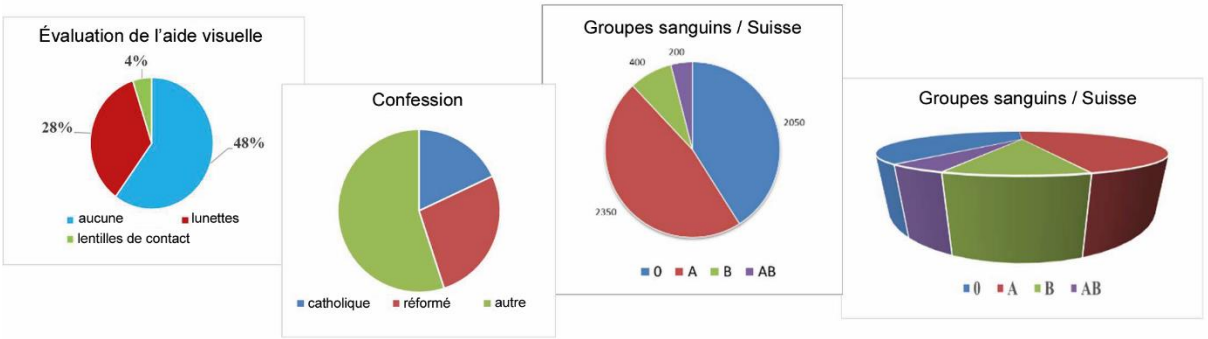


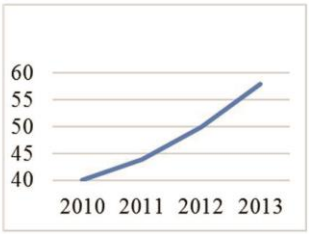
## Diagramme circulaire

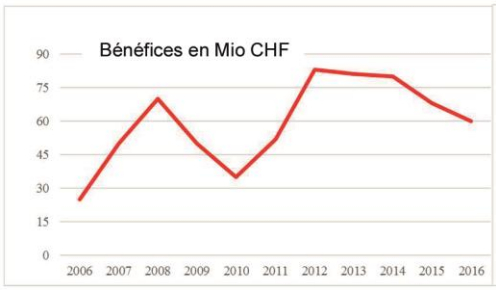
- la somme totale des pourcentages doit être égale à 100%, resp. la somme des fréquences doit être égale à la somme totale de l'échantillon
- aucun segment portant la mention "autre" ou "divers" ne devrait être plus grand que la plupart des autres segments (aucune valeur informative)
- la taille de l'échantillon du caractère étudié doit être connue ou peut être clairement déterminée à partir du graphique
- si le diagramme circulaire est représenté en trois dimensions, l'angle doit être choisi de manière à ce qu'il n'y ait pas de fortes distorsions dans les tailles des segments

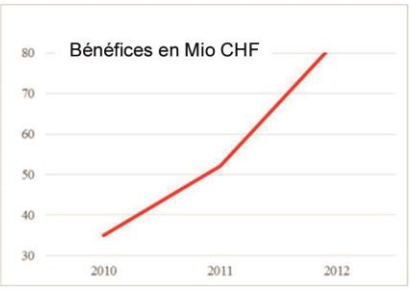


## Graphique en ligne

- l'unité de l'axe vertical (axe horizontal) doit être adaptée aux données; une échelle inappropriée peut fortement influencer le message du graphique
- l'axe temporel doit également être choisi de manière judicieuse et régulière; s'il manque des données pour une année, cette année ne peut pas être simplement omise
- l'ensemble du temps saisi est-il représenté? La représentation d'une partie seulement de l'ensemble de la période peut modifier considérablement le message du graphique.

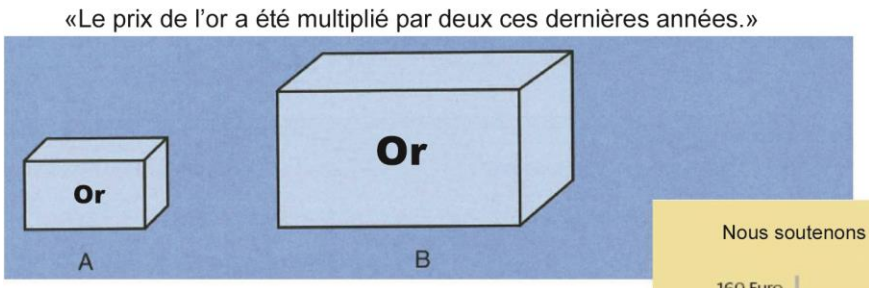



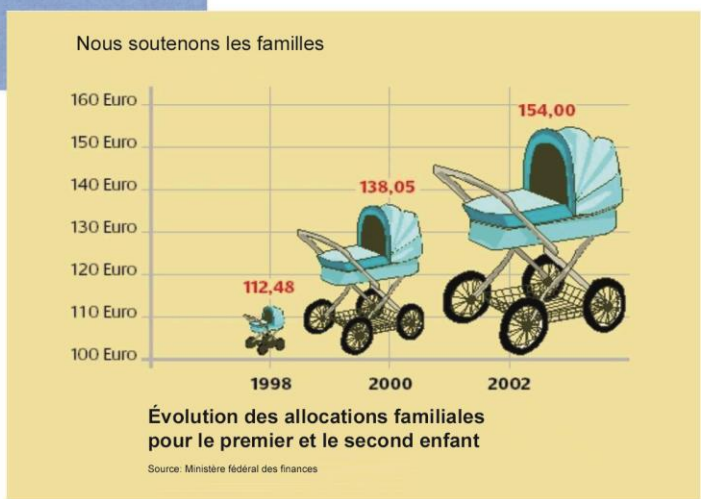





Les tailles des différents **pictogrammes** doivent également être proportionnelles aux indications, sinon le "message" est également fortement modifié.



# 5. Corrélation entre 2 caractères / données bivariées

Lorsque plusieurs caractères sont considérés simultanément lors de mesures ou d'enquêtes, on parle de caractère multidimensionnel. Les caractères multidimensionnels sont surtout pris en compte lorsqu'on suppose l'existence de liens entre les différents caractères. Le plus simple est de considérer la corrélation entre deux caractères (par exemple x et y), de telles données sont appelées données bivariées, dans chaque mesure les deux caractères sont mesurés et enregistrés simultanément.

Les différents caractères d'un porteur de caractéristiques peuvent être liés entre eux d'une manière ou d'une autre, les mesures peuvent par exemple dépendre de la température ou de l'humidité de l'air.

Déterminer le degré de dépendance et le lien fonctionnel conduit aux **calculs de corrélation et de régression.**

Il est donc fréquent que dans une expérience ou une enquête, on obtienne non seulement des observations / caractères individuels $x_i$ ou $y_i$, mais aussi des **paires de valeurs liées** ($x_i$ ; $y_i$).

*Exemples*:

Une première étape dans l'étude des deux variables consiste à tracer le diagramme de points afin d'avoir une vue d'ensemble de la relation entre les deux caractères x et y.

Pour choisir judicieusement les axes, il faut en premier lieu se demander si une dépendance dirigée pourrait exister ou si elle est non dirigée.

Dépendance non dirigée: les deux caractères sont équivalents, il n'est donc pas clair quel caractère est la cause et quel caractère est l'effet; les axes peuvent donc être choisis arbitrairement.

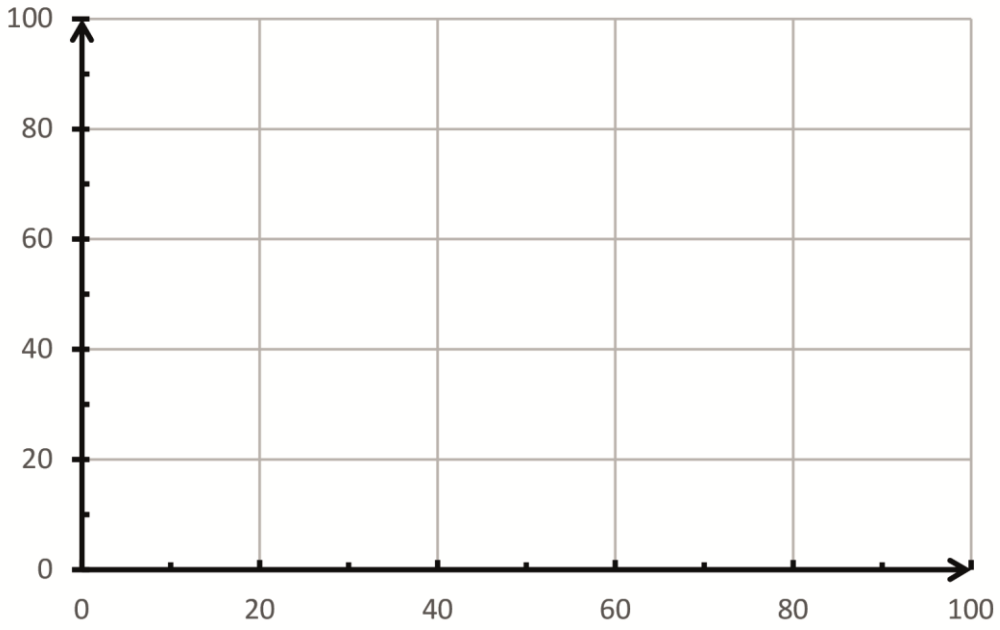
*Exemples*:

Dépendance dirigée: il est clair quel caractère est la cause et quel caractère est l'effet; on place donc toujours le caractère qui est la cause sur l'axe horizontal (axe des x) (variable indépendante) et sur l'axe vertical (axe des y), on place la variable dépendante, autrement dit le „caractère de l'effet".

*Exemples*:

On réalise le même test sur 10 personnes. Le résultat du test et l'âge de la personne sont consignés. (Le test consiste à présenter 100 objets aux personnes pendant un certain temps. On recouvre ensuite les objets puis on vérifie le nombre d'objets dont ils se souviennent.)

| Personne | Âge (ans) | Mémorisation (points) |
|---|---|---|
| A | 32 | 65 |
| B | 58 | 41 |
| C | 55 | 54 |
| D | 21 | 72 |
| E | 65 | 46 |
| F | 18 | 92 |
| G | 38 | 78 |
| H | 48 | 76 |
| I | 61 | 35 |
| J | 45 | 61 |




Exsite-t-il une corrélation entre les 2 grandeurs „âge“ et „mémorisation“?

Il existe de nombreuses corrélations différentes que l'on pourrait vérifier, mais très souvent, seule la dépendance linéaire est examinée de plus près. Pour pouvoir indiquer la tendance de la relation, le diagramme de points x-y donne toujours une bonne première impression.

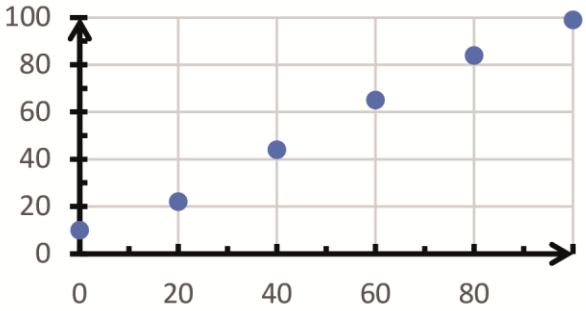
Corrélations possibles ainsi que les types de fonctions correspondants qui permettent d'exprimer mathématiquement cette dépendance:

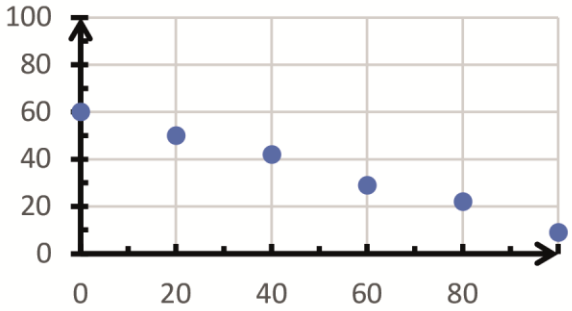
- corrélation linéaire par le point zéro
- corrélation linéaire générale
- corrélation carrée
- corrélation exponentielle
- et bien plus encore

## 5.4 Coefficient de corrélation

Si le diagramme semble révéler une corrélation linéaire, il faudrait en premier lieu vérifier à l'aide du **coefficient de corrélation** r, si cela a vraiment un sens de calculer une droite de régression, il faut donc aussi vérifier "mathématiquement" s'il existe effectivement une relation linéaire.

Selon la manière dont les caractères se comportent, on parle d'une corrélation positive ou d'une corrélation négative.
Corrélation positive: alors que la taille de l'un des caractères augmente, l'autre augmente également




Le coefficient de corrélation peut prendre n'importe quelle **valeur entre -1 et +1**.
Il dépend des écarts-types des deux caractères et d'un écart-type "lié" des deux grandeurs, la covariance.

**Coefficient de corrélation r**

avec: $s_x$ , $s_y$ les 2 écarts-types des caractères x et y
$s_{xy}$ la covariance des caractères x et y

La covariance est en fait une variance "mixte", car les écarts des deux caractères se produisent par rapport à leur valeur moyenne:

covariance $s_{xy}$

$$s_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left( (x_i - \bar{x}) \cdot (y_i - \bar{y}) \right)$$

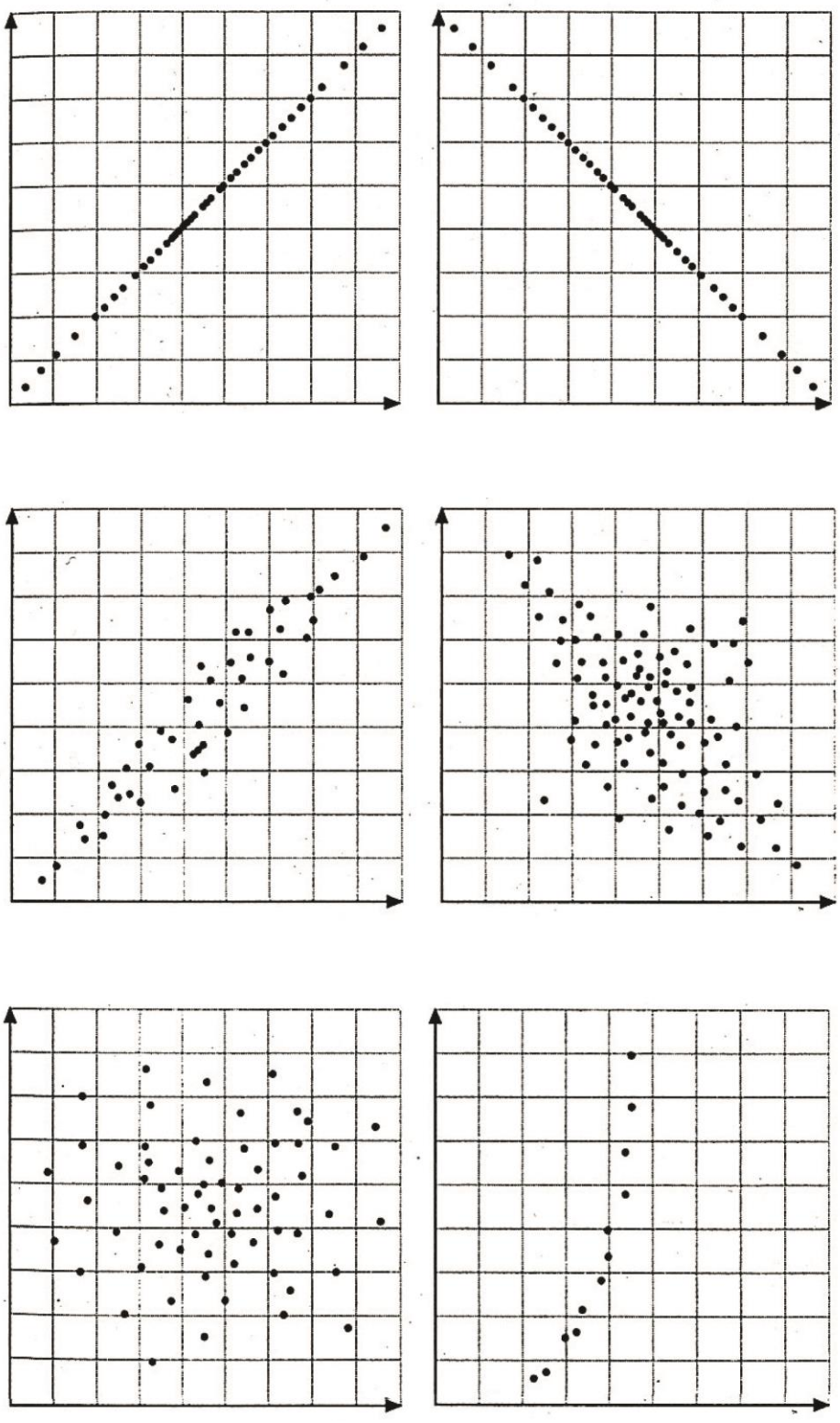
par des transformations, on peut aussi calculer la covariance avec la formule :

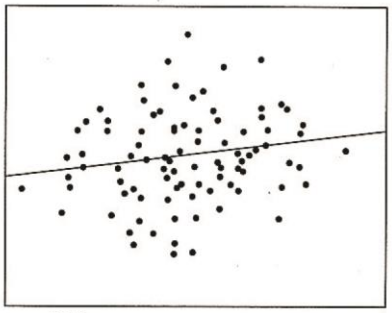
$$s_{xy} = \frac{1}{n-1} \cdot \left( \sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y} \right)$$

Sur la base de la valeur que prend le coefficient de corrélation r, on peut généralement observer une relation linéaire:

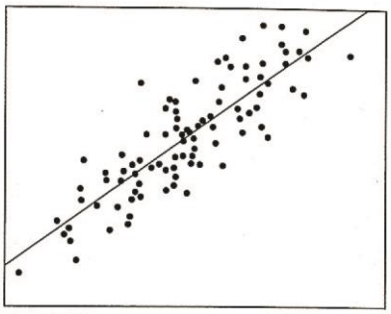
- si le coefficient de corrélation est proche de **+1**, il y aura probablement une forte **corrélation linéaire, positive**
- si le coefficient de corrélation est proche de **-1,** il y aura probablement une forte **corrélation linéaire, négative**
- si le coefficient de corrélation est proche de **0**, il n'y a probablement **pas de corrélation linéaire,** mais éventuellement une autre

Même si le coefficient de corrélation r est proche de +1 ou -1, il n'existe **pas nécessairement** une corrélation linéaire. À l'inverse, il peut y avoir une corrélation linéaire même si r n'est pas proche de +1 ou -1.
C'est pourquoi la représentation graphique à l'aide du diagramme de points x-y est généralement toujours nécessaire. Ce diagramme de points permet également de déterminer s'il existe une autre relation non linéaire, par exemple une relation carrée ou exponentielle.

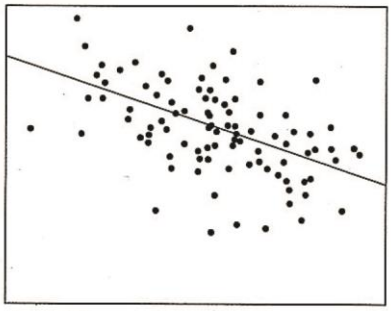




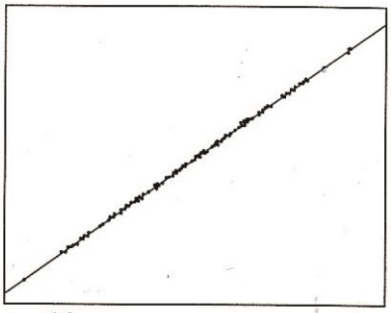
$r = +0{,}15$



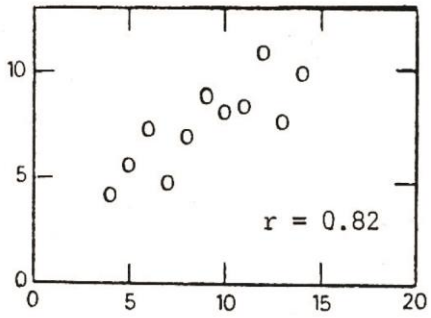
$r = +0{,}85$

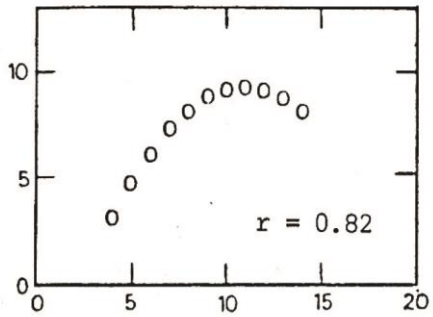


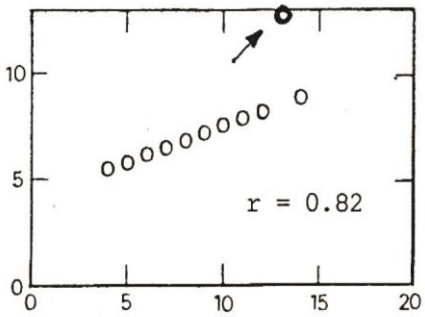
$r = -0{,}50$

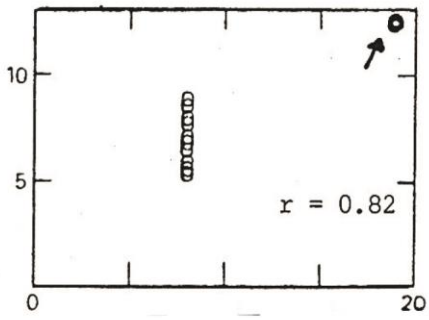


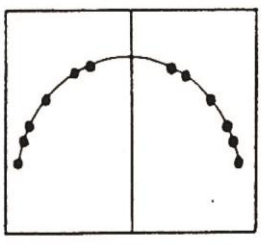
$r = +1{,}0$



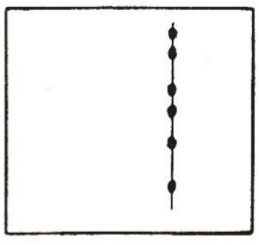



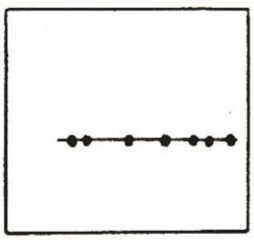



r = 0



r non défini



*Dans l'exemple „âge-mémorisation“,* voici les grandeurs suivantes:

| Personne | Âge (ans) | Écart par rapport à la moyenne $^2$ | Mémorisation (points) | Écart par rapport à la moyenne $^2$ | Produit des deux grandeurs |
|---|---|---|---|---|---|
| A | 32 | *146.41* | 65 | *9* | *2080* |
| B | 58 | *193.21* | 41 | *441* | *2378* |
| C | 55 | *118.81* | 54 | *64* | *2970* |
| D | 21 | *533.61* | 72 | *100* | *1512* |
| E | 65 | *436.81* | 46 | *256* | *2990* |
| F | 18 | *681.21* | 92 | *900* | *1656* |
| G | 38 | *37.21* | 78 | *256* | *2964* |
| H | 48 | *15.21* | 76 | *196* | *3648* |
| I | 61 | *285.61* | 35 | *729* | *2135* |
| J | 45 | *0.81* | 61 | *1* | *2745* |
| Somme | ***441*** | ***2448.9*** | ***620*** | ***2952*** | ***25078*** |

Moyenne âge $\bar{x} =$

Écart-type âge $s_x =$

Moyenne mémorisation $\bar{y} =$

Écart-type mémorisation $s_y =$

Covariance $s_{xy} =$

Coefficient de corrélation $r =$

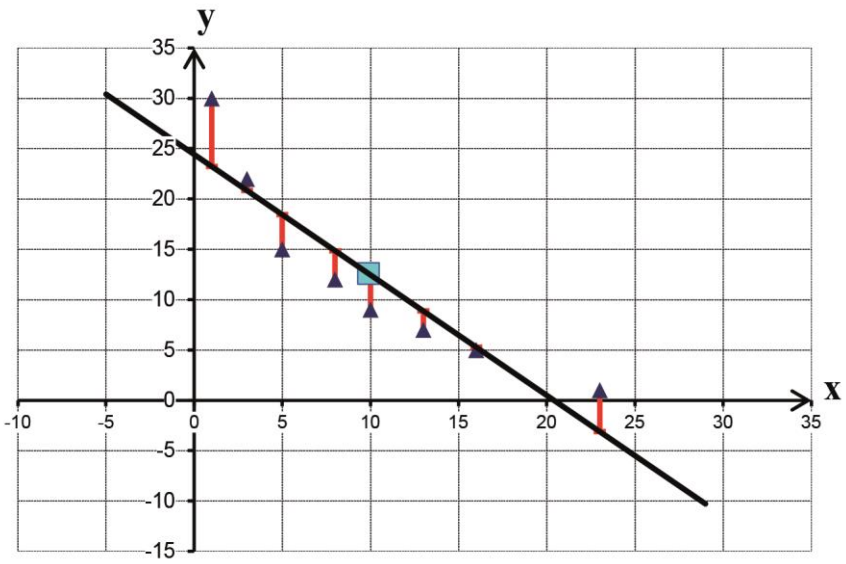
On peut donc admettre une corrélation .....

Si à la fois le graphique et le coefficient de corrélation indiquent une relation linéaire, une droite peut être tracée de manière à ce que les points de mesure soient aussi proches que possible de cette droite. Cette droite ainsi "ajustée au mieux" est appelée **droite de régression.**

## 5.5 Analyse de régression

https://studyflix.de/statistik/regressionsanalyse-2081

Il existe de nombreuses possibilités pour définir la droite comme étant la plus adaptée, mais on choisit à nouveau l'approche selon laquelle la **somme des écarts quadratiques** des écarts verticaux entre les points de mesure et la droite doit être la plus petite possible.




La droite de régression peut, comme c'était le cas avant, avec les fonctions linéaires, être définie avec 2 grandeurs, avec le **segment d'axe y** q et la **pente ascendante** m.

**Droite de régression**: $y = m \cdot x + q$

La pente de la droite de régression peut être calculée à partir des écarts-types des deux caractères et du coefficient de corrélation, ou encore à l'aide de la covariance:

**Pente ascendante**: $m = \frac{s_{xy}}{s_{x^2}} = r \cdot \frac{s_x}{s_y}$

Pour obtenir encore le décalage dans la direction y par rapport au point zéro, (comme jusqu'à présent pour les fonctions linéaires), il faut insérer un point, dans ce cas le point .

**Segment d'axe y**: $q = \bar{y} - m \cdot \bar{x}$

L'équation de la droite permet désormais de calculer des valeurs qui n'ont pas été mesurées, donc des **valeurs escomptées** ou des **prédictions** pour la grandeur dépendante, à condition que la grandeur indépendante soit donnée, ou l'inverse.

**Interpolation**: Il s'agit d'une prédiction où l'on se limite aux valeurs qui se situent dans la plage des valeurs mesurées, c'est-à-dire dans la plage observée. Ce calcul ne pose généralement pas de problème.

**Extrapolation**: Il s'agit d'une prédiction dans laquelle on choisit une valeur qui se situe en dehors de la plage des valeurs mesurées, donc en dehors de la plage observée. Ce calcul est généralement problématique. Il faudrait éviter les extrapolations, à moins d'avoir de très bonnes raisons de penser que la relation linéaire peut se poursuivre au-delà de la zone observée.

*Exemple âge/mémorisation:*

Droite de régression:

pente ascendante $m$ =

segment d'axe y $q$ =

équation de la droite $y$ =

Interpolation:

on estime la capacité de mémorisation d'une personne de 25 ans à ..... points

on estime la capacité de mémorisation d'une personne de 40 ans à ..... points

on estime la capacité de mémorisation d'une personne de 50 ans à ..... points

Extrapolation:

on estime la capacité de mémorisation d'une personne de 15 ans à ..... points

on estime la capacité de mémorisation d'une personne de 70 ans à ..... points

on estime la capacité de mémorisation d'une personne de 10 ans à ..... points

on estime la capacité de mémorisation d'une personne de 100 ans à ..... points

on estime la capacité de mémorisation d'une personne de 1 an à ..... points

Calculer les valeurs x (arguments):

une mémorisation de 25 points correspond à une personne de ..... ans

une mémorisation de 85 points correspond à une personne de ..... ans

une mémorisation de 5 points correspond à une personne de ..... ans

une mémorisation de 99 points correspond à une personne de ..... ans

# 6. Distributions

Dans le deuxième chapitre, des possibilités ont été données sur la manière de représenter les données graphiquement et quels indicateurs doivent être calculés à cet effet. Ces indicateurs et ces graphiques permettent de représenter uniquement la série de mesures, donc on obtient seulement une impression de l'échantillon, de la série de mesures elle-même. Mais rien n'est encore dit sur la manière dont on peut tirer des conclusions sur la **population** dont ces données sont issues, en se basant sur ces données observées/mesurées.

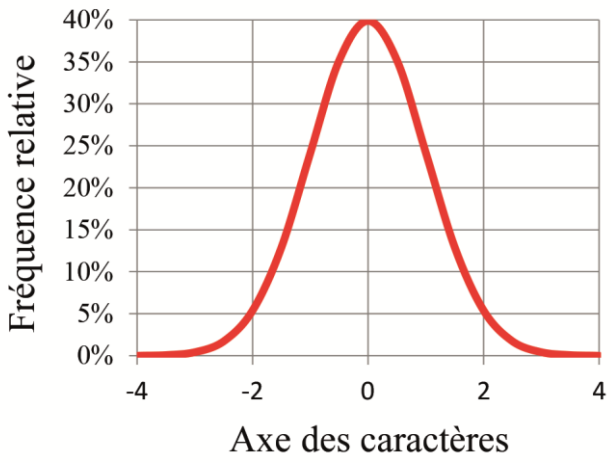
Par population de référence (ensemble, population), on désigne l'ensemble de toutes les observations imaginables du même type, ou le groupe complet des personnes sur lesquelles on veut faire des déclarations statistiques. La population n'est pas toujours connue, très souvent elle n'est malheureusement pas non plus clairement délimitée (pour les enquêtes, il convient d'indiquer clairement quel groupe de personnes constitue la population de référence; pour les mesures, il convient de définir clairement ce qui constitue la population de référence).

Les seules informations dont on dispose souvent sur une population, sont les données d'un échantillon de cette population. Une enquête exhaustive (c'est-à-dire portant sur l'ensemble de la population) n'est souvent pas possible ou bien trop chère. Si l'on veut tirer des conclusions sur l'ensemble de la population à partir de l'étude d'un **échantillon**, il faut non seulement que l'échantillon soit suffisamment grand, mais aussi qu'il soit **représentatif**, c'est-à-dire qu'il ait été **prélevé de manière aléatoire** dans l'ensemble de la population (nous y reviendrons).

La population est un tout qui est stable et dont il est possible de prélever une infinie variété d'échantillons différents. Ces échantillons varient de manière aléatoire, de sorte qu'une conclusion sur la population n'est jamais absolue, mais toujours entachée d'une certaine incertitude. Les indicateurs des échantillons tels que la moyenne et l'écart-type sont donc également des variables aléatoires qui suivent une certaine loi de distribution. Cette loi indique, par exemple, la probabilité selon laquelle les indicateurs des échantillons se situent dans une plage donnée.
Une des lois les plus importantes de la distribution en statistiques est la **distribution normale**.

## 6.1 Distribution normale (DN)

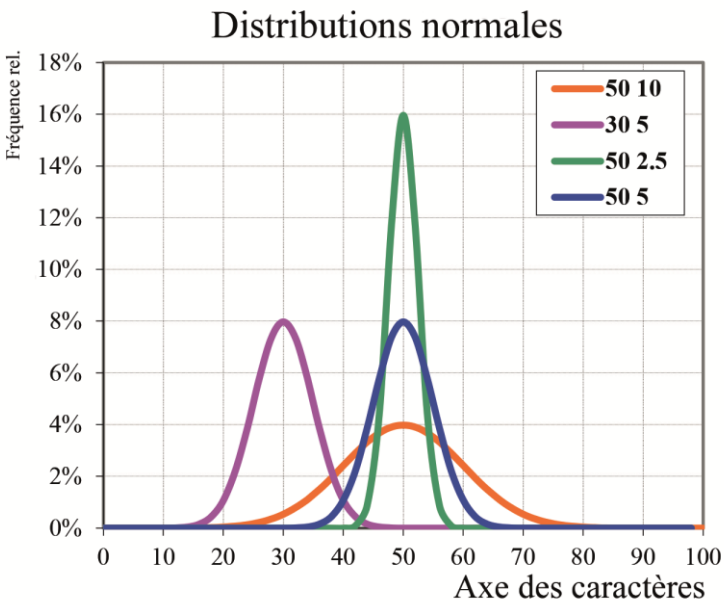
Une grande partie de ce qui est mesuré montre une distribution caractéristique, où une très grande partie de toutes les valeurs mesurées est concentrée autour de la valeur moyenne de la série de données. Plus on est éloigné de cette valeur moyenne, plus la fréquence avec laquelle cette valeur de mesure apparaît est réduite. Cette distribution (diagramme de fréquence, histogramme) a donc une forme symétrique, atteint sa valeur maximale à la moyenne, puis diminue des deux côtés jusqu'à zéro. Cette forme de cloche est très typique et est aussi très souvent obtenue dans la pratique (au moins approximativement), si la taille de l'échantillon n'est pas trop petite.




Cette distribution de probabilité continue la plus importante est la **distribution normale**. Elle a été analysée par Abraham de Moivre et plus tard Carl Friedrich Gauss. La contribution de Gauss a été si importante que la distribution normale est souvent connue sous le nom de **distribution de Gauss**.

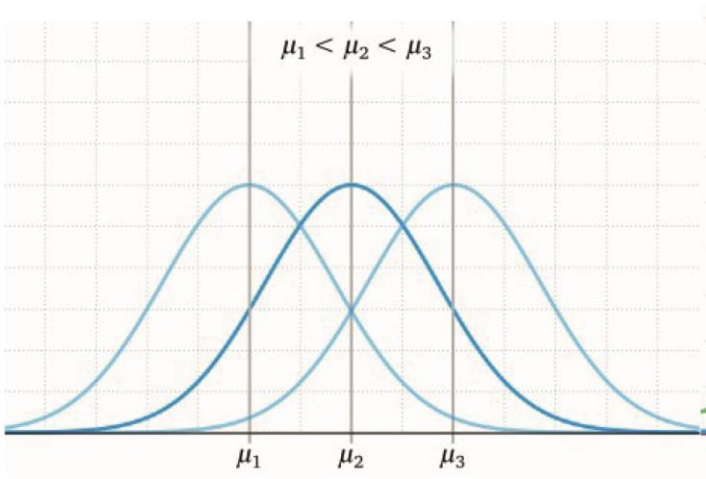
En raison de sa forme caractéristique, elle est parfois appelée tout simplement **courbe en cloche**, même s'il existe d'autres fonctions de distribution qui ont un graphique en forme de cloche (p. ex. la loi de Poisson ou loi binomiale).

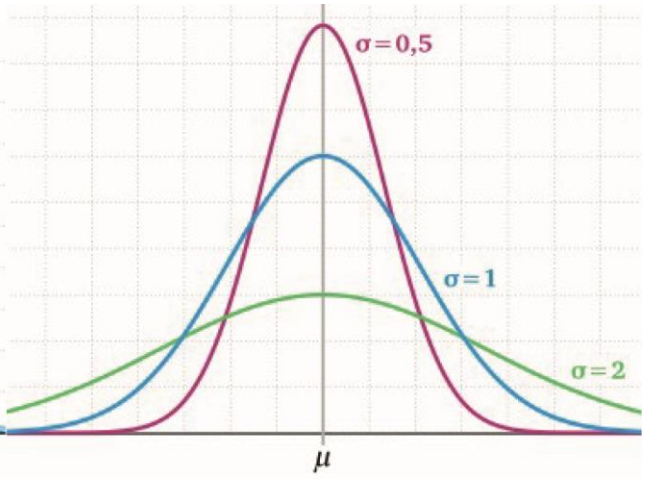
La distribution normale est la distribution la plus importante en statistiques. Elle est aussi bien utilisée en sciences naturelles qu'en sciences humaines et économiques, surtout lorsque la fonction de distribution réelle sur laquelle reposent les données est inconnue.
Les valeurs des caractéristiques observées dans la pratique peuvent donc souvent être considérées comme normalement distribuées. En outre, les valeurs moyennes issues d'échantillons suivent une distribution à peu près normale, à condition que la taille de l'échantillon soit suffisamment grande. Cela vaut également lorsque la population étudiée ne suit pas une distribution normale.
Les propriétés spécifiques à la distribution normale (DN) permet une utilisation simple.
La DN est déterminée de manière **univoque** par **2 grandeurs** seulement, ces deux grandeurs étant la valeur moyenne et l'écart-type.
Dans les échantillons, la valeur moyenne est désignée par , dans la population, elle est donnée par la lettre grecque μ. L'écart-type d'un échantillon est symbolisée par le signe s, celui de la population par σ.




L'illustration montre différentes DN, avec les indicateurs respectifs pour la valeur moyenne et pour l'écart-type.
La valeur moyenne indique donc à nouveau la position de la courbe, l'écart-type indique la largeur et donc la hauteur de la valeur maximale.
Si, comme illustré, la fréquence relative est tracée comme axe vertical, on obtient toujours la même surface sous la courbe (intégrale) pour chaque DN, quelles que soient les valeurs, à savoir la **surface 1**.

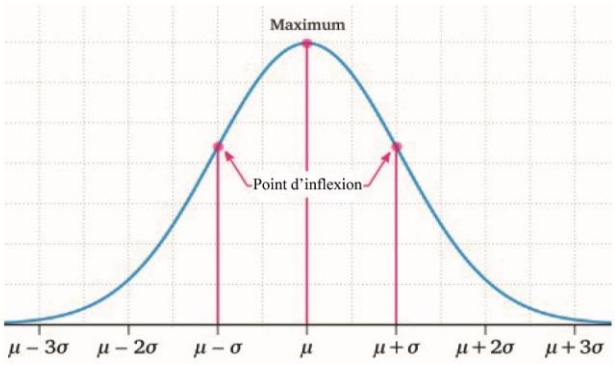
Si on compare différentes DN qui ont le même écart-type, c'est-à-dire qui ne diffèrent que par leur moyenne, les courbes ont l'air identiques, il s'agit simplement de courbes décalées.
Mais si on compare des DN qui ont la même moyenne et des écarts-types différents, les courbes sont au même endroit, elles sont toutefois „étirées" ou „comprimées".




Un autre avantage important de la DN est la possibilité de lire directement sur la forme de la courbe, non seulement la valeur moyenne mais aussi l'écart-type.
Si l'on observe le tracé de la courbe avec plus de précision (différenciation, déduction), elle augmente d'abord lentement à partir de petites valeurs, elle devient de plus en plus raide, avant que la pente ne diminue à nouveau à mesure que l'on se rapproche du maximum de la courbe. Au maximum proprement dit, la pente de la courbe est exactement nulle et devient ensuite négative si l'on va vers des valeurs
encore plus grandes.

Le passage d'une pente de plus en plus forte vers une pente plus plate est appelé **le point d'inflexion**. La distance (horizontale) de ce point d'inflexion jusqu'à la moyenne correspond précisément à l'écart-type s de cette distribution.




Comme il n'est pas possible d'indiquer avec précision les fréquences de toutes les valeurs d'un échantillon, on utilise la DN. Avec la DN, il est possible de calculer la fréquence pour des valeurs quelconques en fonction de leurs deux paramètres  et s.

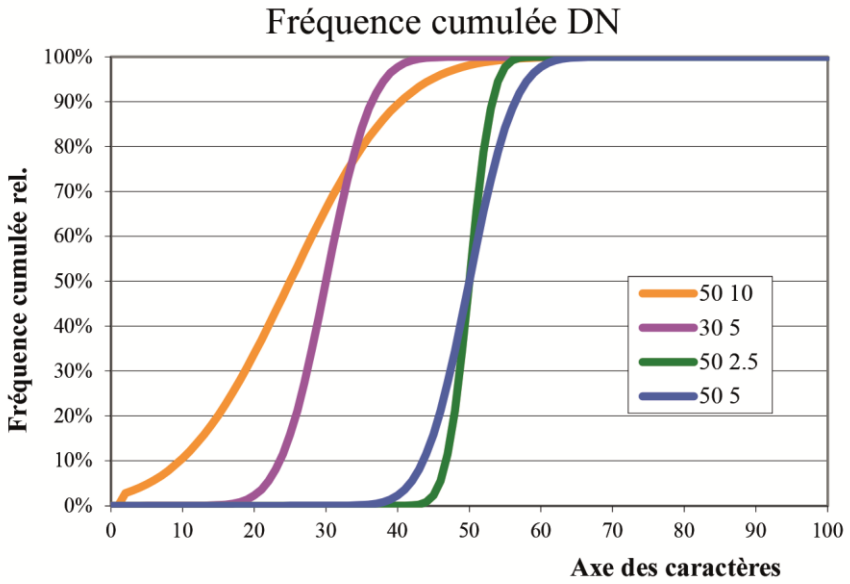
Ce que l'on appelle la **fonction de densité**, avec laquelle il est possible de calculer de telles valeurs, est la suivante:

$$f(x) = \frac{1}{s \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot \frac{(x-\overline{x})^2}{s^2}}$$

ou alors, par analogie, avec les données d'une population:

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

Comme précédemment pour les indicateurs, la fréquence cumulée / le quantile est une grandeur pratique pour décrire la distribution des valeurs. Pour la DN aussi, c'est très souvent la fréquence cumulée qui est indiquée (ce qui correspond à la surface sous la courbe). Comme la surface totale sous la courbe correspond à la fréquence cumulée totale de l'échantillon, elle est toujours égale à 1, c'est-à-dire à 100%. Plus l'écart-type est petit, plus cette courbe de fréquence cumulée monte abruptement; plus l'écart-type est grand, plus la courbe s'aplatit.




Cette fréquence cumulée, autrement dit la surface sous la courbe DN, n'est pas si simple à calculer, si elle doit être déterminée de manière précise pour toutes les valeurs x possibles et il faut refaire tout le calcul pour chaque DN en fonction de ses paramètres spécifiques.

*Exemple* L'exemple du groupe sanguin mentionné ci-dessus donne la fréquence cumulée pour les deux valeurs:

| | |
|---|---|
| pour 120 mmHg | 7.69% |
| pour 160 mmHg | 92.31% |
| pour 140 mmHg | |
| si pour 100 mmHg | 0.22% |
| alors pour | etc. |

Pour décider si une série de données peut être considérée comme normalement distribuée, quelques considérations suffisent:

- .....
- .....
- .....

## 6.2 Distribution normale standard (DNS)

En règle générale, on spécifie toutes les valeurs de la distribution et de la fréquence cumulée pour une seule DN particulière et on peut ensuite convertir toutes les autres DN en cette seule **distribution normale standard** (DNS). Elle est pour ainsi dire l'unité de base de la DN.

Pour cette DNS, on prend le cas le plus simple pour les deux paramètres d'une DN.
La **DNS** a la **moyenne** $\bar{x}$ = 0 et l'**écart-type** $s$ = 1.

Si on applique ces valeurs dans l'équation fonctionnelle de la DN, on obtient, pour la DNS, la fonction suivante:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot z^2}$$

Chaque DN peut être transformée en cette DNS et inversement, la DNS peut être transformée en n'importe quelle DN. Il suffit de décaler les données (jusqu'à ce que la moyenne soit ) et d'étirer/de compresser les données de manière à ce que l'écart-type atteigne .

Si dans une série de mesure, chaque valeur de mesure est **décalée** d'une certaine valeur a (p. ex. conversion des températures de °C en K), cela donne pour la série décalée

une moyenne de:

un écart-type de:

Si dans une série de mesure, chaque valeur de mesure est **multipliée** par une certaine valeur b (étiréé) (p. ex. conversion de masses de g en kg), cela donne pour la série étirée

une moyenne de:

un écart-type de:

Ces considérations permettent maintenant de convertir les données, c'est-à-dire de les déplacer et de les étirer de manière à ce qu'elles correspondent ensuite à la DNS ou inversement.

**Transformation d'un DN en DNS**:

donc calculer à partir du x (de la DN) un z correspondant (de la DNS):

**Transformation de la DNS en une DN**:

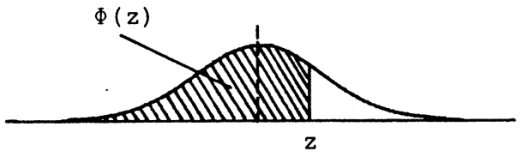
donc calculer à partir du z (de la DNS) un x correspondant (de la DN):

Les surfaces sous les différentes courbes des DN se comportent de la même façon, puisque pour chaque DN, la surface totale est identique et donne toujours 100%. À partir des données de la DNS, il est donc posssible de lire les pourcentages (quantiles) au niveau de la valeur z transformée et de témoigner directement que le même pourcentage est également atteint pour la valeur x correspondante de la DN initiale.

Ainsi, il est possible de répondre à des questions sur la base des données d'échantillonnage, telles que:

- Combien de pourcentages des données de mesure seront plus grands / plus petits qu'une valeur donnée?
- Combien de pourcentages des données de mesure se trouveront entre / en dehors de deux limites prédéfinies?
- Quelle valeur dépassera / ne dépassera pas un pourcentage donné de toutes les données de mesure?
- etc.

Tableaux relatifs aux DNS (issus de „Statistik im Laboratorium" de G. Rey et U. Kreuter)




| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| .0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| .1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| .2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| .3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| .4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| .5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| .6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| .7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| .8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| .9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 | .9292 | .9306 | .9319 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| 1.8 | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| 1.9 | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |
| 2.0 | .9772 | .9778 | .9783 | .9788 | .9793 | .9798 | .9803 | .9808 | .9812 | .9817 |
| 2.1 | .9821 | .9826 | .9830 | .9834 | .9838 | .9842 | .9846 | .9850 | .9854 | .9857 |
| 2.2 | .9861 | .9864 | .9868 | .9871 | .9875 | .9878 | .9881 | .9884 | .9887 | .9890 |
| 2.3 | .9893 | .9896 | .9898 | .9901 | .9904 | .9906 | .9909 | .9911 | .9913 | .9916 |
| 2.4 | .9918 | .9920 | .9922 | .9925 | .9927 | .9929 | .9931 | .9932 | .9934 | .9936 |
| 2.5 | .9938 | .9940 | .9941 | .9943 | .9945 | .9946 | .9948 | .9949 | .9951 | .9952 |
| 2.6 | .9953 | .9955 | .9956 | .9957 | .9959 | .9960 | .9961 | .9962 | .9963 | .9964 |
| 2.7 | .9965 | .9966 | .9967 | .9968 | .9969 | .9970 | .9971 | .9972 | .9973 | .9974 |
| 2.8 | .9974 | .9975 | .9976 | .9977 | .9977 | .9978 | .9979 | .9979 | .9980 | .9981 |
| 2.9 | .9981 | .9982 | .9982 | .9983 | .9984 | .9984 | .9985 | .9985 | .9986 | .9986 |
| 3.0 | .9987 | .9987 | .9987 | .9988 | .9988 | .9989 | .9989 | .9989 | .9990 | .9990 |
| 3.1 | .9990 | .9991 | .9991 | .9991 | .9992 | .9992 | .9992 | .9992 | .9993 | .9993 |
| 3.2 | .9993 | .9993 | .9994 | .9994 | .9994 | .9994 | .9994 | .9995 | .9995 | .9995 |
| 3.3 | .9995 | .9995 | .9995 | .9996 | .9996 | .9996 | .9996 | .9996 | .9996 | .9997 |
| 3.4 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9998 |

Comme la moyenne est à 0 et qu'il s'agit d'une distribution symétrique, seules les valeurs positives sont inscrites dans le tableau, les valeurs négatives sont à déterminer de façon identique avec $\Phi(z) = 1 - \Phi(-z)$

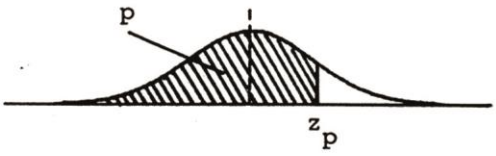
*Exemple de lecture*:

La valeur z = 1,84 a le quantile ……, donc …… % de toutes les données sont inférieures à z = 1.84, ou …… % de toutes les données sont supérieures à z = 1.84.

La valeur z = -2,68 le quantile …… , donc …… % de toutes les données sont inférieures à z = -2,68, ou …… % de toutes les données sont supérieures à z = -2,68.

Si le quantile est prédéfini et qu'on cherche la valeur z correspondante, on a le tableau suivant pour quelques valeurs (pas aussi détaillé que tableau 1)

*Tableau A-2:* quantiles de la distribution normale standard: $z_p$




| p | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| .00 | – | –2.33 | –2.05 | –1.88 | –1.75 | –1.64 | –1.55 | –1.48 | –1.41 | –1.34 |
| .10 | –1.28 | –1.23 | –1.18 | –1.13 | –1.08 | –1.04 | –0.99 | –0.95 | –0.92 | –0.88 |
| .20 | –0.84 | –0.81 | –0.77 | –0.74 | –0.71 | –0.67 | –0.64 | –0.61 | –0.58 | –0.55 |
| .30 | –0.52 | –0.50 | –0.47 | –0.44 | –0.41 | –0.39 | –0.36 | –0.33 | –0.31 | –0.28 |
| .40 | –0.25 | –0.23 | –0.20 | –0.18 | –0.15 | –0.13 | –0.10 | –0.08 | –0.05 | –0.03 |
| .50 | 0.00 | 0.03 | 0.05 | 0.08 | 0.10 | 0.13 | 0.15 | 0.18 | 0.20 | 0.23 |
| .60 | 0.25 | 0.28 | 0.31 | 0.33 | 0.36 | 0.39 | 0.41 | 0.44 | 0.47 | 0.50 |
| .70 | 0.52 | 0.55 | 0.58 | 0.61 | 0.64 | 0.67 | 0.71 | 0.74 | 0.77 | 0.81 |
| .80 | 0.84 | 0.88 | 0.92 | 0.95 | 0.99 | 1.04 | 1.08 | 1.13 | 1.18 | 1.23 |
| .90 | 1.28 | 1.34 | 1.41 | 1.48 | 1.55 | 1.64 | 1.75 | 1.88 | 2.05 | 2.33 |

| p | .995 | .990 | .975 | .950 | .900 |
|---|---|---|---|---|---|
| $z_p$ | 2.576 | 2.326 | 1.960 | 1.645 | 1.282 |

*Exemple*:

**Loi des grands nombres (pour toutes les DN)**:

Combien de données (en %) se trouvent:

- dans la plage $\mu \pm \sigma$ ……
- dans la plage $\mu \pm 2\sigma$ ……
- dans la plage $\mu \pm 3\sigma$ ……

*Exemple*:

On a analysé la teneur en nitrates d'échantillons d'épinards. Dans l'un des échantillons, on a relevé les grandeurs $\bar{x}$ = 3%  nitrate et $s$ = 0.8%  nitrate. On souhaite savoir plusieurs choses à ce sujet:

**a)** Sur l'ensemble des échantillons d'épinards, quel est le pourcentage d'échantillons ayant une teneur en nitrate supérieure à 4%?

**b)** Quelle est la teneur en nitrates qui n'est dépassée que par 2% de tous les échantillons ?

**c)** Quel est le pourcentage de tous les échantillons ayant une teneur en nitrates comprise entre 2,5% et 3,5%

**d)** Quel est le pourcentage d'échantillons dont la valeur est inférieure ou égale à 1% de nitrates ?

**e)** Dans quelle fourchette de teneur en nitrates se situent les 56% des échantillons moyens?

## 6.3 Intervalle de confiance

La moyenne  d'un échantillon **ne donne pas une estimation exacte** de la **vraie** moyenne μ de la population, mais lorsqu'il s'agit d'un échantillon représentatif et suffisamment grand et que les données présentent une distribution plutôt faible (donc un petit écart-type), la moyenne de cet échantillon constitue sans doute une très bonne évaluation de la **vraie** moyenne μ. Pour la grandeur de la moyenne μ, on indique donc normalement une **fourchette** dans laquelle la valeur moyenne réelle se situera avec une très grande certitude. Cette fourchette est appelée **intervalle de confiance** (de la moyenne).

La moyenne obtenue à partir d'un échantillon est la meilleure estimation possible de la vraie moyenne. On ajoute à cette moyenne un montant adéquat pour définir la limite supérieure de l'intervalle de confiance et de la même manière, on soustrait le même montant de la moyenne pour déterminer la limite inférieure. La véritable moyenne μ se situera dans cet intervalle avec une grande certitude.

Quelle sera la largeur d'un tel intervalle de confiance, c'est-à-dire quelle sera la valeur ajoutée/soustraite de la valeur moyenne de l'échantillon, dépend de différents facteurs:

- quel doit être le niveau de sécurité (ou à quel point l'erreur doit être petite), avec lequel on peut supposer que μ se situe dans cet intervalle
- quel est le degré de fiabilité, c'est-à-dire de précision, des données de mesure utilisées dans l'échantillon
- quelle est la taille de l'échantillon dont on utilise la moyenne comme estimation de μ
- à quel point l'échantillon utilisé est représentatif par rapport à la population totale

**Probabilité p**

$$\frac{\text{Nombre de résultats favorables}}{\text{Nombre de résultats possibles}}$$

Définition de la probabilité: rapport entre les résultats favorables et l'ensemble des résultats possibles.

P =

Exemple: quelle est la probabilité de tirer un nombre impair avec un dé (6 surfaces avec 1, 2, 3, 4, 5, 6 points → 6 résultats possibles)?

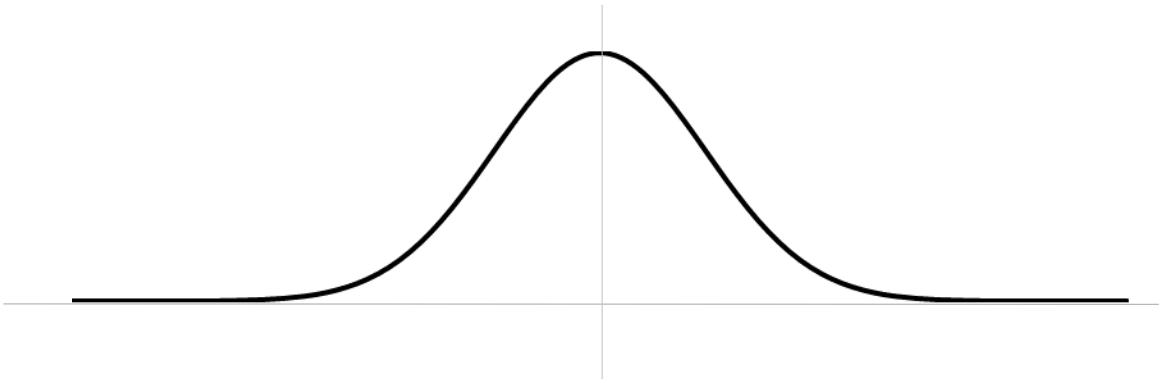
La probabilité que l'intervalle de confiance ne contienne pas la vraie valeur de la population est appelée probabilité d'erreur α (« niveau de signification »).

α = 0,05 signifie que l'intervalle de confiance estimé contient la vraie valeur de la population avec une **probabilité de 95%**.

**Sécurité: probabilité d'erreur α, ou coefficient de confiance 1- α**

On se demande quel doit être le degré de sécurité avec lequel la valeur moyenne estimée doit correspondre à la valeur moyenne effective, c'est-à-dire vraie, de la population. En premier lieu, on aimerait bien sûr être tout à fait sûr que la valeur concorde, mais avec la sécurité élevée souhaitée, la largeur de l'intervalle augmente fortement.

Avec le **coefficient de confiance 1-α,** on indique donc quel doit être le niveau de sécurité (ou quelle est l'ampleur de l'erreur, c'est-à-dire la probabilité d'erreur α). Si, par exemple, α = 10% cela signifie que dans 10 cas sur 100, on admet une moyenne qui ne correspond pas à la moyenne effective (erreur de 10%), où dans 5 cas, la VM effective sera plus grande que l'intervalle estimé et dans 5 cas, elle sera plus petite (distribution normale). Dans 90 cas sur 100, la VM effective se situe toutefois dans la fourchette estimée (sécurité 1-α = 90%).



Ainsi, plus on choisit un intervalle de confiance large, plus on peut être sûr que la vraie moyenne se situe dans cet intervalle. Un intervalle très large signifie donc une erreur très faible, mais en contrepartie, un intervalle large ne dit plus grand-chose sur la valeur moyenne effective.

*Exemple* : *âge des participant(e)s à l'examen professionnel supérieur (EPS)*:
Si l'on prend pour échantillon de tous les participant(e)s, une année EPS p.ex. au cours des 10 dernières années et des 10 prochaines années, on a p.ex. pour l'échantillon (HFPL15), les données suivantes $\bar{x}$ = 28,6 années, avec s = 5,3 années , pour une taille d'échantillon de $n$ = 20..
En supposant que les valeurs sont distribuées approximativement selon une loi normale, cela donne pour l'estimation de la moyenne effective μ pour l'âge de tous les participant(e)s à l'EPS:

- à faible sécurité (α = 20%) 28,6 ans ± 1,5 ans
- à moyenne sécurité (α = 5%) 28,6 ans ± 2,3 ans
- à très grande sécurité (α = 0.01%) 28,6 ans ± 4,6 ans

Pour les études réelles, il faut un intervalle suffisamment large pour contenir la vraie moyenne avec une fiabilité/certitude raisonnable. En même temps, il doit être suffisamment étroit pour que l'on puisse encore y lire des informations. Le „compromis" souvent utilisé est **l'intervalle de confiance de 95%** (donc un intervalle avec un coefficient de confiance de 1-α = 95%), et donc une probabilité d'erreur de **α = 5%.**
Pour des applications spécifiques, on utilise parfois la valeur α = 10%, α = 1% voire α = 0,1%.

## Influence de l'échantillon sur l'intervalle de confiance

On part du principe que l'échantillon est représentatif de la population.
Mais même un échantillon représentatif peut encore fortement influencer la largeur de l'intervalle de confiance:

- Plus la dispersion des valeurs mesurées est grande, plus les données sont réparties largement et, en conséquence, l'intervalle de confiance qui en résulte est également beaucoup plus large (pour une même sécurité, c'est-à-dire pour un même α)

- Plus la taille de l'échantillon est grande, plus l'estimation que l'on fait pour la population est sûre, car beaucoup plus d'éléments de la population ont été pris en compte. Cela signifie qu'avec un échantillon plus grand, l'intervalle de confiance se rétrécit, tout en conservant la même sécurité.

**Calcul de l'intervalle de confiance**

Si les données sont distribuées de manière approximativement symétrique, l'intervalle de confiance pour la moyenne attendue μ peut maintenant être calculé à l'aide de la formule suivante:

$$\bar{x} \pm \frac{z_{1-\alpha/2} \cdot s}{\sqrt{n}}$$

avec:

| | |
|---|---|
| $\bar{x}$ | moyenne calculée de l'échantillon |
| s | écart-type calculé de l'échantillon |
| n | taille de l'échantillon |
| $z_{1-\alpha/2}$ | quantile de la DNS, des tableaux 1 ou 2 en page 33/34 |

*Exemple* : *âge des participant(e)s à l'examen professionnel supérieur (EPS)*:
$\bar{x}$ = 28,6 années avec s = 5,3 années, pour une taille d'échantillon de $n$ = 20
Quelle sera la taille de l'intervalle de confiance pour la moyenne avec une probabilité d'erreur de α = 10%? Et quelle est la taille de cet intervalle de confiance si α = 1%?

Quelle doit être la taille minimale de l'échantillon, pour que (à écart-type identique) l'intervalle de confiance $\bar{x} \pm 1$ année, avec une probabilité d'erreur de α = 1%?

# 7. Le concept du test statistique

Ce chapitre est un chapitre complémentaire qui permet de mieux comprendre les applications.

## 7.1 Échantillonnage

- Dans un échantillon, on observe une indication qu'il existe un lien entre deux caractéristiques (p. ex. taille et poids corporel) ou une différence entre deux groupes (p. ex. femmes / hommes).

- *Problème: Il n'est pas clair si cette observation n'est apparue que par hasard dans notre échantillon ou si elle est réellement valable pour l'ensemble de la population.*

## 7.2 Hypothèse

Supposition: en statistique, on part du principe qu'il n'existe pas de différences ou de corrélations dans la population.
→ concept des hypothèses pour la population

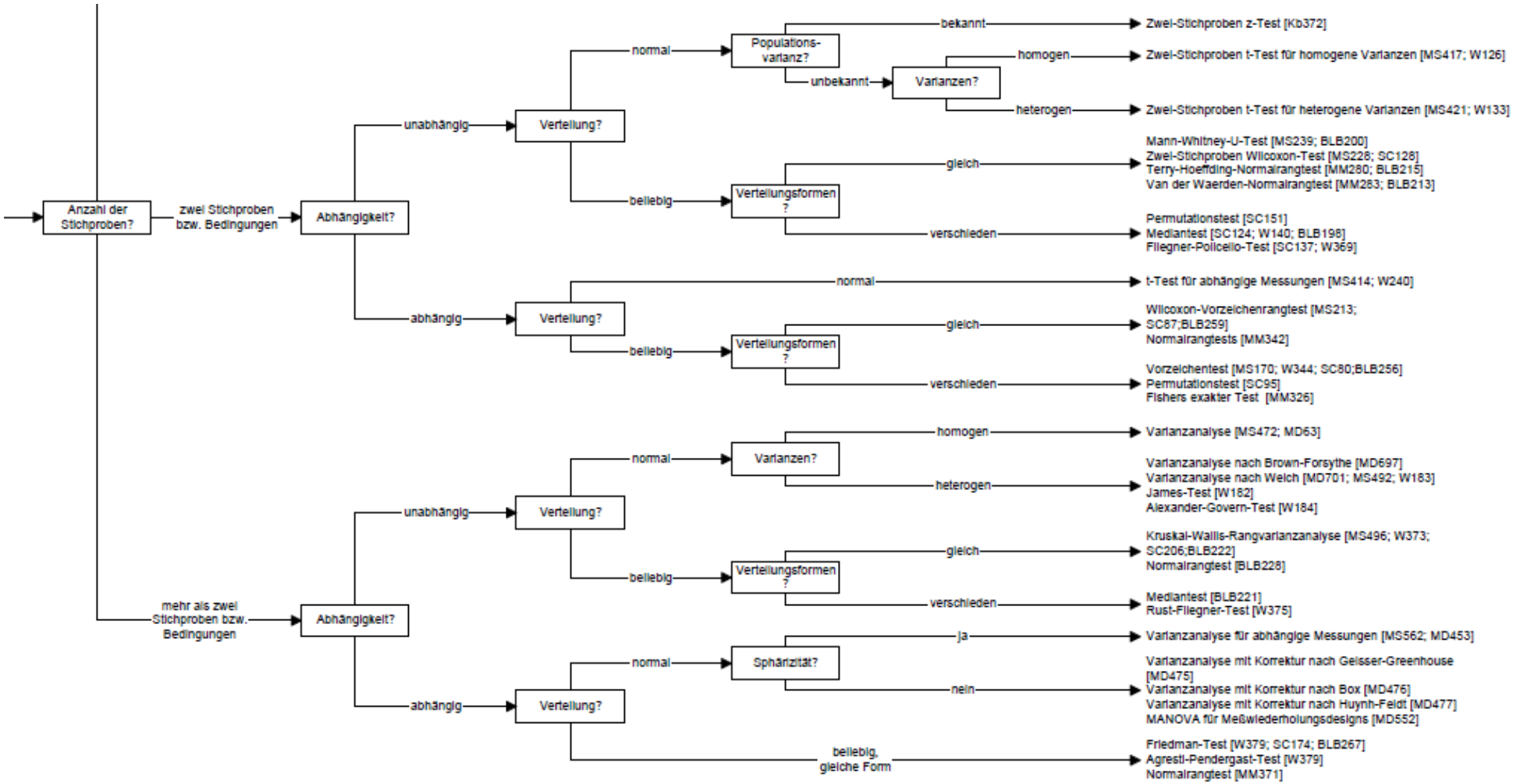
***Hypothèse nulle $H_0$ =*** *il n'y a pas de corrélation ou pas de différence dans la population.*
***$H_0$*** : $\mu_x = \mu_y$

***Hypothèse alternative $H_A$*** *= il y a une corrélation ou une différence dans la population.*
***$H_A$*** : $\mu_x \neq \mu_y$

Liens d'information**:**
https://studyflix.de/statistik/hypothesentest-1719
https://studyflix.de/statistik/nullhypothese-1586

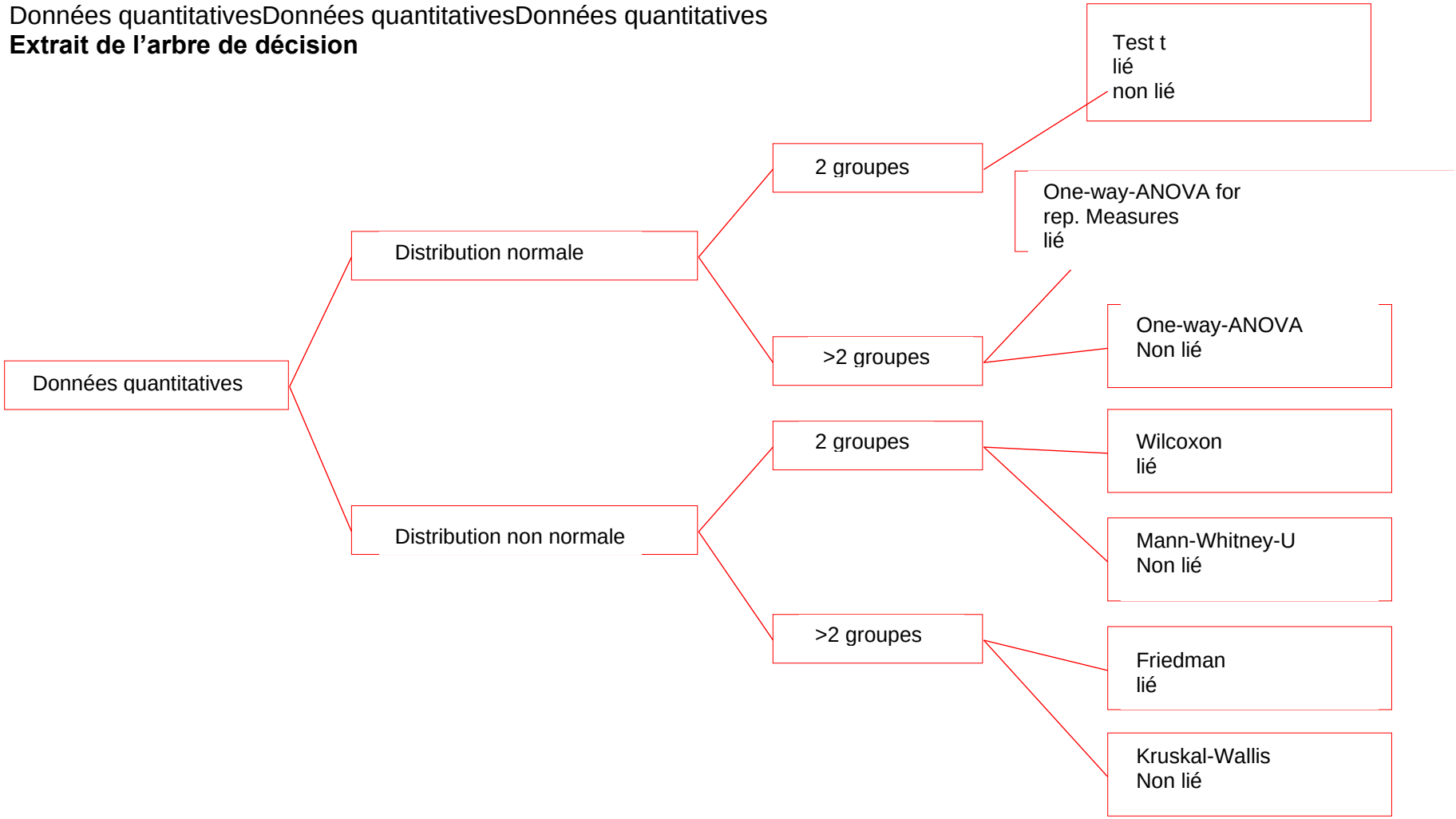
## 7.3 Statistiques de test

Il existe différentes statistiques de test qui sont appliquées par exemple en fonction de la question, des formes de distribution, du nombre d'échantillons



# STATISTIQUES

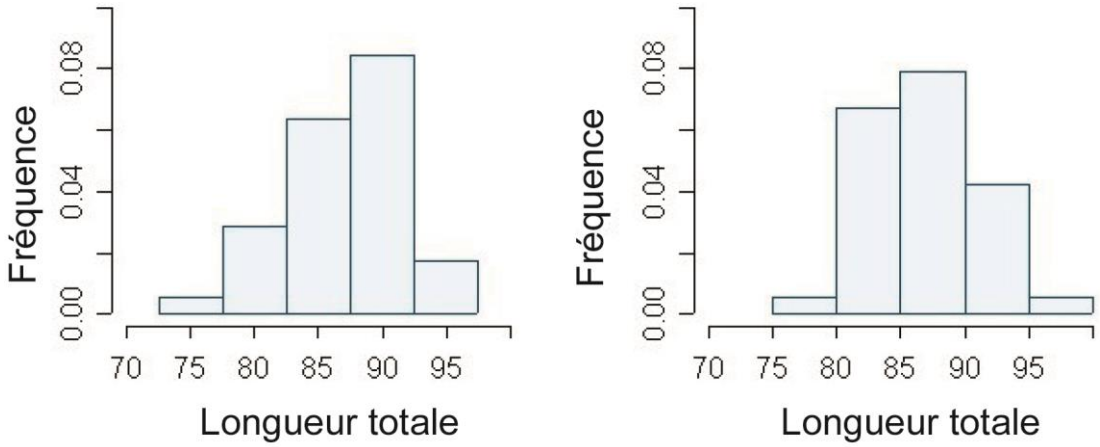
Données quantitativesDonnées quantitativesDonnées quantitatives
**Extrait de l’arbre de décision**




**Liens d’information:**
https://studyflix.de/statistik/anova-2213
https://studyflix.de/statistik/t-test-1584

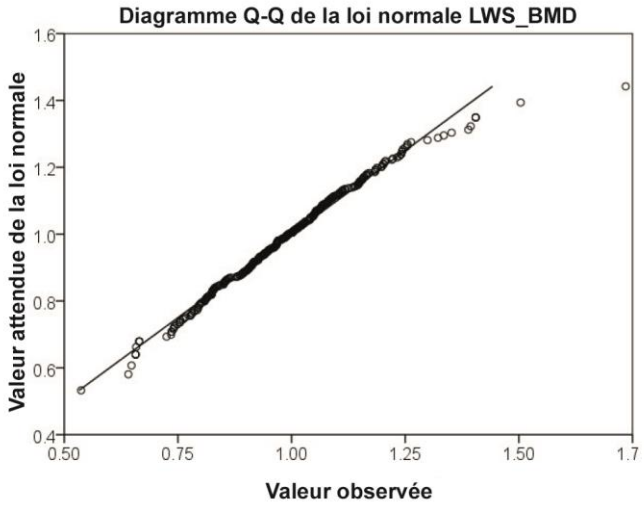
## 7.4 Vérification de la distribution normale

1. *À l'aide d'un histogramme*: souvent insuffisant, car l'aspect de l'histogramme est influencé par la largeur et les limites des barres



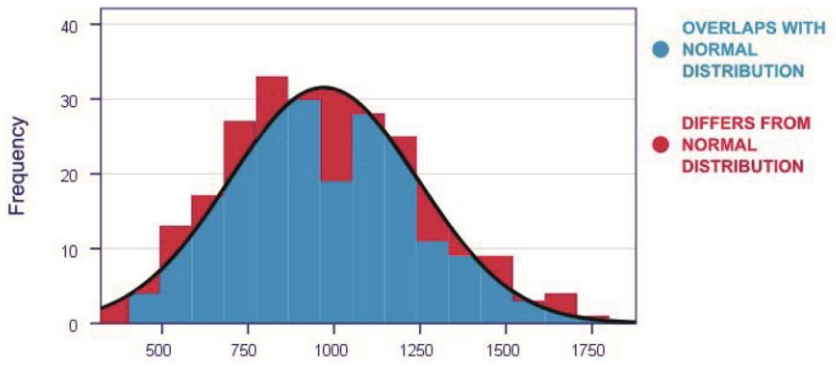

➔ données analogiques représentées avec des limites de barres différentes

2. *À l'aide d'un diagramme quantile-quantile («Q-Q-Plot»)*: les quantiles théoriques des données sont représentés par rapport aux quantiles empiriques.




➔ La linéarité est considérée comme un indice d'une distribution normale

3. *À l'aide de tests statistiques:* p.ex. «Shapiro-Wilks», «Anderson-Darling», «Kolmogorow-Smirnow»



## 7.5 Niveau de signification ou niveau α  Power β

Le *niveau de signification* désigne la **probabilité** avec laquelle, dans le cadre d'une procédure de test statistique, l'**hypothèse nulle** ($H_0$) peut être rejetée à tort, même si H0 est en fait correcte.

→ avec un niveau de signification α = 0.05 on accepte que la décision de test (p.ex. «rejeter $H_0$») est correcte dans 95% des cas et n'est pas correcte dans 5% des cas.

| Situation dans l'échantillon | Situation dans la population : $H_0$ est vraie | Situation dans la population : $H_0$ est fausse |
|---|---|---|
| Accepter $H_0$ → résultat test non significatif | vrai | faux<br>erreur de type 2 |
| Rejeter $H_0$ → résultat test significatif | faux<br>erreur de type 1 | vrai |

- **Erreur de type 1 (erreur α):** probabilité de rejeter une véritable hypothèse nulle

  → *erreur de rejet (souvent 5%)*

- **Erreur de type 2 (erreur β):** probabilité d'accepter une véritable hypothèse nulle
  → *erreur de non-rejet*

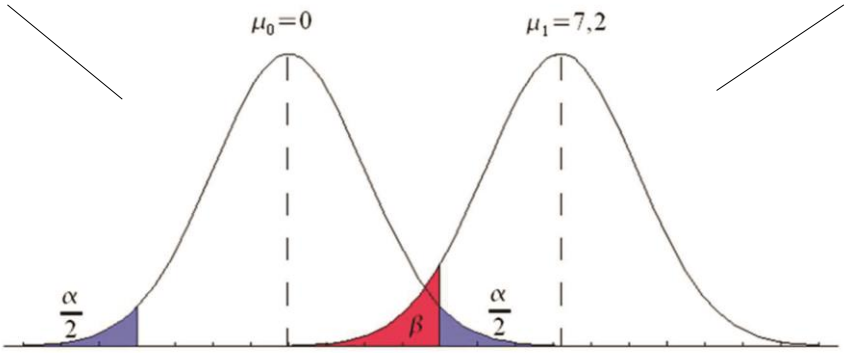
La «contrepartie» de l'erreur β est la **puissance**:

| *Puissance = 1 – β* | *β = «erreur de non-rejet»* |
|---|---|

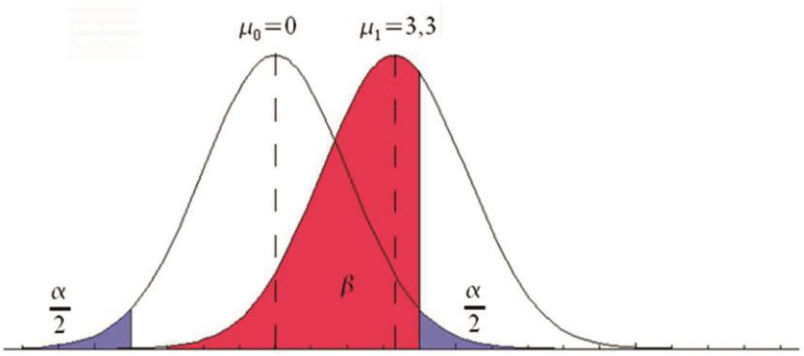
En statistiques, la «**puissance**» décrit la **force** d'un test statistique ou sensibilité de test.

La **puissance** indique la **probabilité** de détecter une différence présente dans la population lors d'une étude statistique.

Distribution sous $H_0$ Distribution sous $H_A$




→ La puissance(1-β) est grande. La sensibilité du test dans cette situation est élevée.




→ La puissance (1-β) est petite. La sensibilité du test dans cette situation est réduite.