Chapitre 1 – Variables observées et distributions expérimentales - Statistique descriptive -

1. Notion de variable observée

Individus ou unités statistiques: objet concernés par la question que l'on se pose.

Variables: caractéristiques des individus étudiés.
Forte variabilité
=> *variable aléatoire*

Critères de choix des variables:
- **complétude des variables,**
- **pertinence,**
- **indépendance.**

Types de variables	Exemples
Qualitatives	
Binaires (2 descriptions)	Absence/présence d'une espèce
Multiples (Plusieurs descriptions)	≠ phénotypes possibles
Non ordonnées (échelle nominale)	e.g. sexe, couleur les yeux, etc.
Ordonnées (échelle ordinale)	e.g. rare, présent, abondant…
Semi-quantitatives (intervalles variables entre classes)	Etat d'un malade (en danger, sérieux, sans danger, etc.)
Quantitatives (intervalles de classe connu)	
Discontinues	Nb d'œufs par nid
Continues	Taille

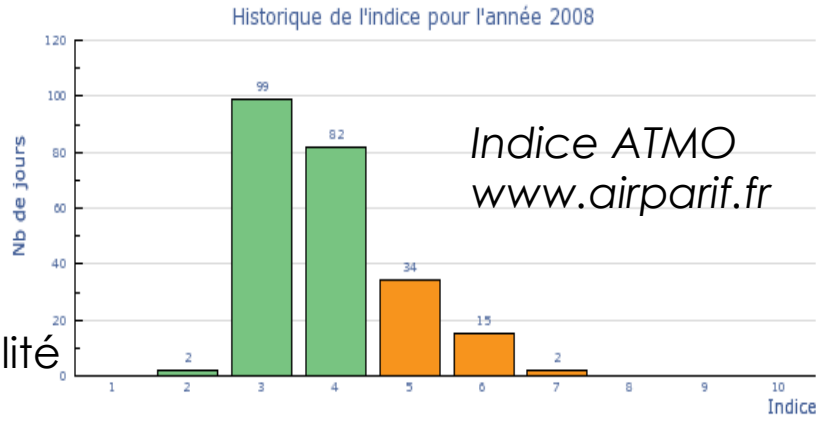
2. Distributions observées

Echantillon représentatif = *n* unités statistiques ➔ forment une distribution

Afin de caractériser les échantillons, la 1ere étape est de résumer les informations collectées et les caractériser.

Représentation graphique

Diagramme en bâtons:
Variables qualitatives et quantitatives discrètes



Représente :
- le nombre d'occurrences n_i pour chaque modalité
- ou la fréquence $f_i = n_i/n$

Histogramme:
Variables quantitatives continues

Les valeurs observées sont **discrétisées**.
On trace le nombre de valeurs observées n_i pour chaque classe *i*
ou la fréquence $f_i = n_i/n$ en fonction du centre de la classe.

Nombre de classes: choisi pour une représentation claire
Règle de Sturge $k=1+3.3\ ln(n)$; *ou règle de Yule:* $k=2.5\ n^{1/4}$

3. Caractérisation des distributions observées

3.1 – Indicateurs de la localisation des valeurs

- **valeurs maximale et minimale** ;
- **le mode** : valeur qui revient le plus souvent ;
- **la médiane** : paramètre de tendance centrale qui sert à résumer une série de valeurs d'une variable quantitative.
= valeur pour laquelle il y a 50% de chances d'être plus grand ;
- **les quantiles ou percentiles** :

Pour toute série numérique de données dans un intervalle I, on définit le quantile par :

$$e_{n,U}=inf\{x \text{ t.q. } F(x) >= u\},$$

où F(x) = la fréquence des éléments de la série inférieurs ou égaux à x.

P-Quantile: Fraction de données se situant sous une valeur limite p.

On rencontre aussi le terme **quartile**, Q25 est le premier quartile (25% des données sont inférieures à Q25), Q50 le 2ème quartile et Q75 le 3ème quartile ;

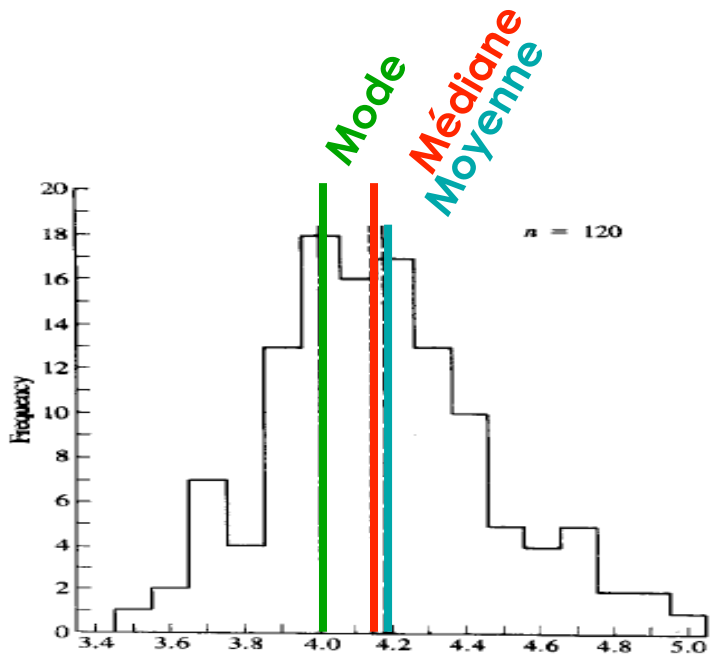
• **la moyenne** : paramètre de tendance centrale qui sert à résumer une série de données d' une variable quantitative.

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

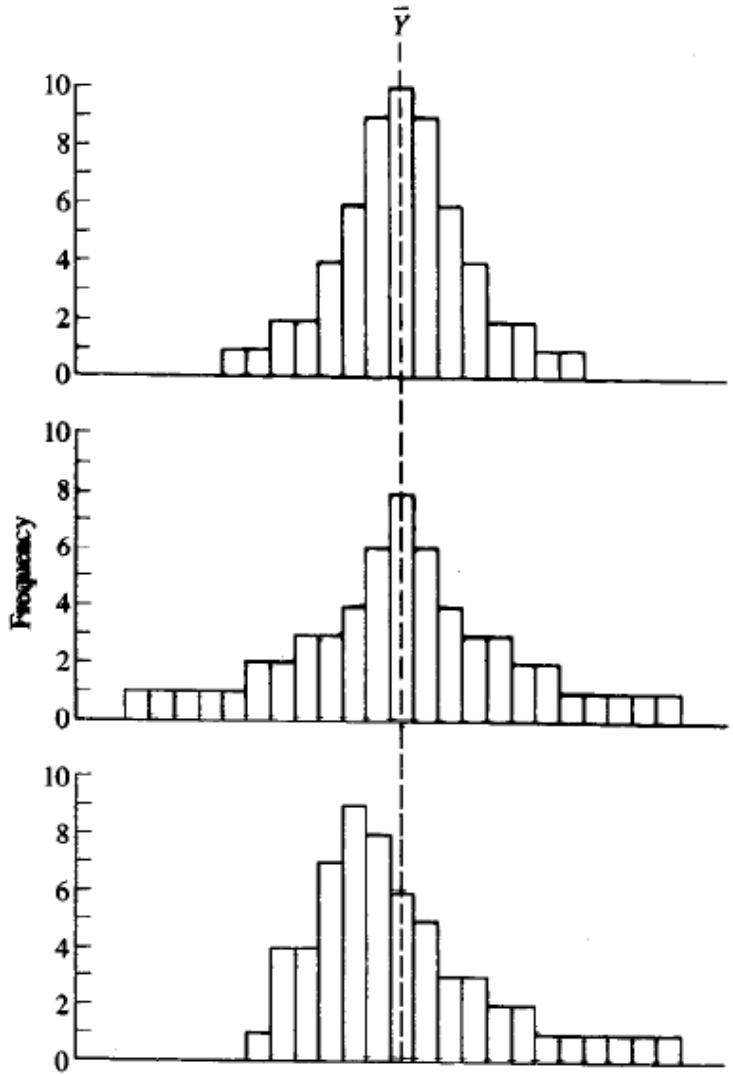
Ou lorsque les valeurs sont réparties en k classes (histogr.)

$$\mu = \frac{1}{n}\sum_{j=1}^{k} n_j e_j \quad \text{avec} \quad k = \sum_{j=1}^{k} n_j$$

Exemple de distribution et paramètres de position correspondants:



Exemple de distributions ayant la même moyenne:



➔ La moyenne ne permet pas la description complète de la distribution

3.2 – Indicateurs de la dispersion des valeurs

Variance:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\mu^2\right)$$

Ou, pour des variables réparties en *k* classes:

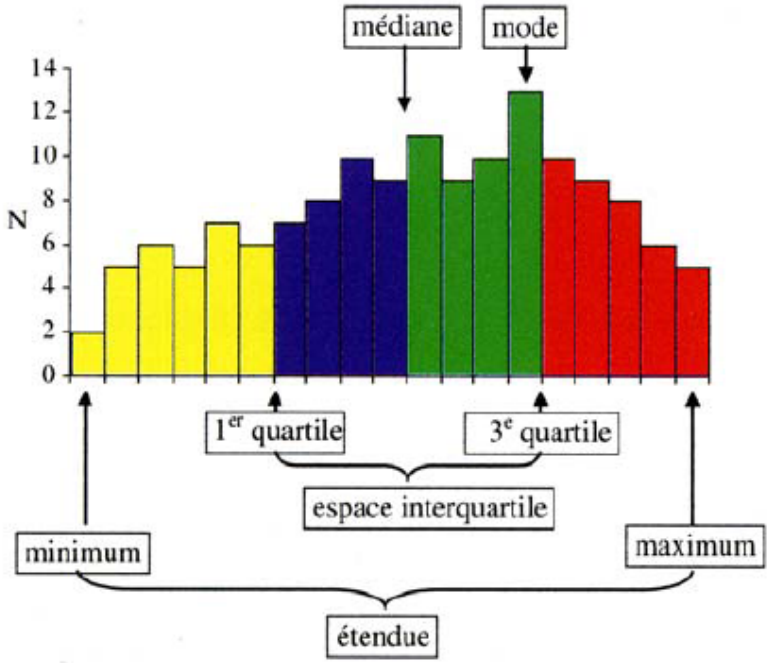
$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{k}\left(n_j e_j - \mu\right)^2 = \frac{1}{n}\left(\sum_{j=1}^{k} n_j e_j^2 - \frac{\left(\sum_{i=1}^{k} n_j e_j\right)^2}{n}\right) = \frac{1}{n}\left(\sum_{j=1}^{k} n_j e_j^2 - n\mu^2\right)$$

Ecart type: $\sigma = \sqrt{\sigma^2}$

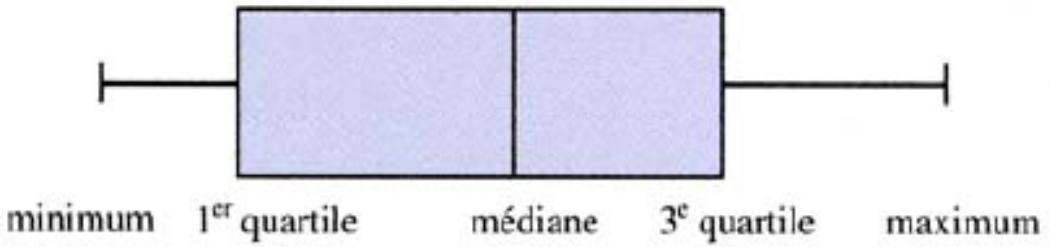
Coefficient de variation: $CV = 100\ \frac{\sigma}{\mu}$

3.3 – Les boîtes à moustache

Résumé graphique des paramètres de la distribution observée



Paramètres résumés dans la boîte à moustaches



3.4 – Distributions observées et Loi de Probabilité

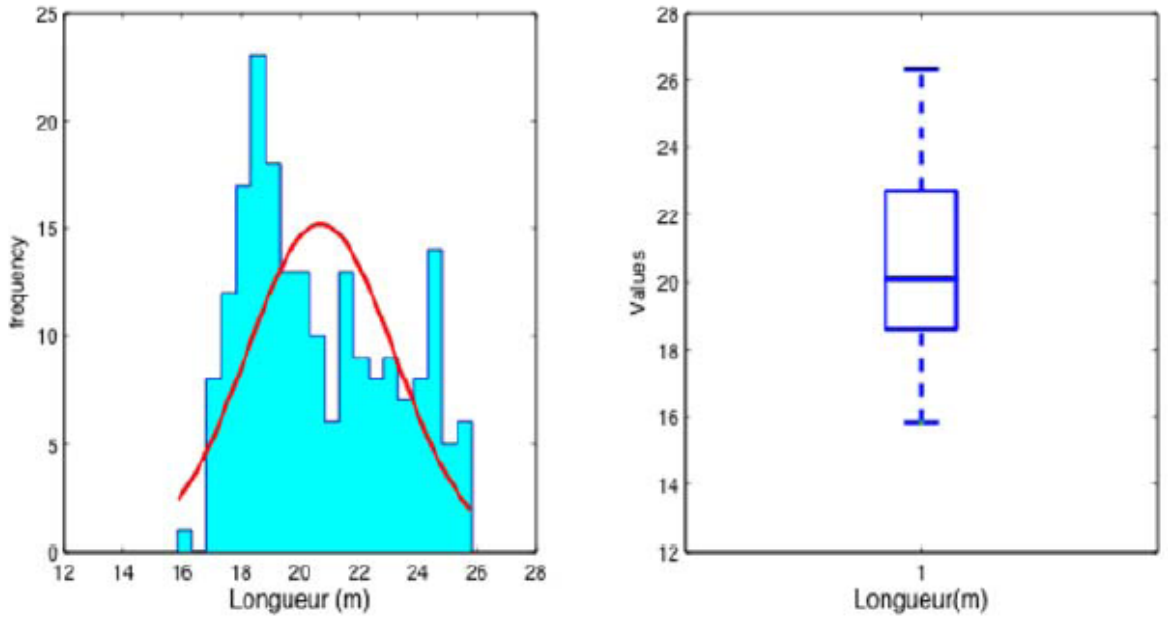


Figure 6 *: Distribution observée des longueurs de baleines pêchées au début du XXème siècle : Histogramme et boîte à moustaches. Sur l'histogramme, une loi de probabilité est ajustée.*

Si l' échantillon est représentatif et composé de variables aléatoires, la distribution pourra être résumée par un seul objet mathématique == **loi de probabilité**.