# Séance 4: Mesures de tendances centrales

## Et Séance 5: Mesures de dispersion (à venir)

Visseho Adjiwanou, PhD.

26 January 2023

# Plan de présentation

1. Description statistique des variables quantitatives
   - Paramètres de position
   - Paramètres de dispersion (prochain cours)

# Description statistique des variables qualitatives (rappel)

# Description statistique des variables qualitatives

Soit une série de valeurs qualitative: H, F, F, F, H, F, H, F, F, F, F, H, H, F, H, H, . . . , F

- donner les effectifs de chaque modalité
- donner les proportions (= fréquences) de chaque modalité par rapport au total
- combiner si besoin les proportions, notamment des proportions cumulées pour des variables ordinales)

# Description statistique des variables qualitatives

La variable X prend les valeurs $x_1, x_2, ..., x_p$, n valeurs avec p occurrences différentes:

| Occurence de $X$ | $x_1$ | $x_2$ | $\dots$ | $x_i$ | $\dots$ | $x_p$ | total |
|---|---|---|---|---|---|---|---|
| Effectifs | $n_1$ | $n_2$ | $\dots$ | $n_i$ | $\dots$ | $n_p$ | $n$ |
| Fréquence | $f_1$ | $f_2$ | $\dots$ | $f_i$ | $\dots$ | $f_p$ | 1 |

# Description statistique des variables qualitatives

- Nombre total d'observation
- $$n = \sum_{i=1}^{p} n_i$$
- Fréquence relative

$$f_i = \frac{n_i}{n}$$

- Somme des fréquences

$$\sum_{i=1}^{p} f_i = 1$$

# Description statistique des variables quantitatives

# Description statistique des variables quantitatives

Les variables sont décrites numériquement par :

1. des **paramètres de position** encore appelés **mesures de tendance centrale**

- moyenne
- percentiles, dont :
- médiane
- premier (Q1) et troisième quartile (Q3)
- percentiles p
- autres : tiertiles, déciles, etc
- mode
- médiale
- minimum et maximum

# Description statistique des variables quantitatives

Mais aussi :

2. des **paramètres de dispersion**

- variance
- écart-type
- écart inter-quartile
- étendue ou amplitude
- coefficient de variation

Plus skewness et kurtosis, paramètres d'étalement et d'asymétrie.

# Paramètres de position

# Paramètres de position

- Une mesure de tendance centrale est une valeur **typique** ou **représentative** d'un ensemble de scores
- Il existe différentes façons de caractériser le centre d'une distribution. Nous en présenterons les trois façons les plus utilisés:
- La moyenne
- la Médiane
- le mode

# Moyenne arithmétique

**La Moyenne (arithmétique)** = Somme des valeurs divisée par l'effectif de la série

- Son calcul dépend du type de données:
  - données individuelles
  - données agrégées

# Moyenne sur données individuelles

- Soit un échantillon de taille n $X_1, X_2, ..., X_n$
- Moyenne

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Exemple : calcul de la moyenne arithmétique pour les données suivantes :

- données individuelles : 6, 7, 7, 7, 8, 8, 8, 9, 9, 10
- La moyenne vaut $(6 + 7 + \ldots + 10)/10 = 7{,}9$

# Moyenne sur données groupées

- $x_1, x_2, ..., x_p$ étant les p occurrences observées
- $n_1, n_2, ..., n_p$, les effectifs correspondants de ces occurrences.
- $n = n_1 + n_2 + ... + n_p$
- $f_1, f_2, ..., f_p$ sont les fréquences relatives.
- $f_1 = \frac{n_1}{n}$
- La moyenne $\bar{X}$ vaut:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{p} n_i * x_i = \frac{1}{n}(n_1 * x_1 + n_2 * x_2 + ... + n_p * x_p)$$

# Moyenne sur données groupées

$$\bar{X} = \sum_{i=1}^{p} f_i * x_i = f_1 * x_1 + f_2 * x_2 + ... + f_p * x_p$$

# Moyenne sur données groupées : exemple

Données individuelles : 6, 7, 7, 7, 8, 8, 8, 9, 9, 10

Ces données peuvent être regroupées en (les parenthèses font références aux formules en haut):

- 6 ($x_1$) –> 1 ($n_1$) ==> fréquence relative 1/10
- 7 ($x_2$) –> 3 ($n_2$) ==> fréquence relative 3/10
- 8 ($x_3$) –> 3 ($n_3$) ==> fréquence relative 3/10
- 9 ($x_4$) –> 2 ($n_4$) ==> fréquence relative 2/10
- 10 ($x_5$) –> 1 ($n_5$) ==> fréquence relative 1/10
- Nombre d'éléments de la série n = n1 + n2 + ... + n5 = 10

–> La moyenne vaut alors 1/10*6 + 3/10*7 + ... + 1/10*10 = 7,9

# Propriété de moyenne

1. Est affecté par les cas déviants ou extrêmes (scores particulièrement bas ou élevés)
   - Exemple: Remplacer dans les scores précédents 10 par 700
2. Lorsque la moyenne est soustraite de chaque valeur individuelle et que ces différences sont additionnées, la valeur donne 0. Autrement: $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$
   - La moyenne "équilibre" pour ainsi dire une distribution
3. La moyenne minimise la somme des déviations au carré de chaque score par rapport à la moyenne
   - Le mot déviance renvoie à la différence entre un score et la moyenne

# La Médiane

**La Médiane** = valeur telle que la moitié des observations lui sont inférieures et donc la moitié lui sont supérieures.

- Deux cas se présentent:

1. le nombre de valeurs est impair (n impair)

- si n = 15, (n + 1)/ 2 = 8 –> la médiane est la huitième valeur de la série.
- Exemple: 1, 1, 2, 2, 3, 4, 5, 6, 6, 7, 8, 9, 9, 9, 10
  - Médiane = 6

# La Médiane

**La Médiane** = valeur telle que la moitié des observations lui sont inférieures et donc la moitié lui sont supérieures.

- Deux cas se présentent:

2. le nombre de valeurs est pair (n pair)

- Tout nombre compris entre (n/2) et (n/2)+1 répond à la définition.
- On définit alors généralement la médiane par : médiane = $(x_{n/2} + x_{n/2+1})/2$
- Exemple: si la série des valeurs est {1, 1, 2, 2, 3, 4, 5, 6, 6, 7, 8, 9}
- n = 12, la médiane se trouve alors entre la 6e (=12/2) et la 7e (6 + 1) valeur

# Le Mode

**Le Mode** = encore appelée **valeur dominante**: valeur observée de fréquence maximum.

- le mode est la valeur la plus fréquente mais de manière relative et pas absolue (donc pas forcément la majorité des valeurs)
- il peut y avoir deux ou plusieurs modes :
  - 1, 2, 3, 3, 3, 3, 4, 5, 6, 6, 6, 6, 7, 15 : modes = 3 et 6
- lorsqu'une distribution est bimodale, on peut penser que l'échantillon est en réalité issu de deux populations différentes
- si toutes les valeurs sont différentes, autant de modes que de valeurs :
  - 1, 2, 3, 5, 6, 9, 14, 16 –> chaque valeur = mode

# Quel paramètre de tendance centrale utilisée (moyenne, médiane, mode)

Dépend de :

1. du niveau de mesure de la variable;

- La moyenne se prête bien aux variables de ratio et d'intervalles
- Parce qu'elle utilise l'ensemble des scores, elle contient plus d'informations que le mode et la médiane;
- La médiane se prête mieux aux variables ordinales
- Le mode est à privilégier pour les variables nominales

# Quel paramètre de tendance centrale utilisée (moyenne, médiane, mode)

2. de sa distribution
   - dans une distribution **symétrique** et **unimodale**, le mode, la médiane et la moyenne affichent la même valeur.
   - Par contre, dans une distribution **asymétrique**, les trois valeurs sont différentes
   - la moyenne est plus petite que la médiane lorsque l'asymétrie de la distribution se trouve à gauche;
   - la moyenne est plus grande que la médiane lorsque l'asymétrie de la distribution se trouve à droite;
   - Le médiane se situe toujours entre la moyenne et le mode
   - Donc employer la médiane si la distribution est très asymétrique
3. de ce que l'on tente de découvrir avec cette variable

# Quartiles

**Quartiles** Les trois quartiles divisent l'ensemble de la distribution en 4 ensembles de même taille (au moins approximativement):

- Q1 –> 25% des valeurs sont inférieures à Q1
- Q2 –> Médiane –> 50% des valeurs sont inférieures à Q2
- Q3 –> 75% des valeurs sont inférieures à Q3

# Formes générales: quantiles

**Quantiles / Fractiles** Le quantile d'ordre $k$ est la valeur qui sépare la distribution en k classes de même effectifs (au moins approximativement).

- déciles: divise la distribution en . . . ,
- quartiles: divise la distribution en . . . ,
- quintiles: divise la distribution en . . . ,
- tertiles: divise la distribution en . . . ,
- centiles: divise la distribution en . . . , etc.

# Formes générales: quantiles

**Percentile** le percentile p divise la distribution en deux groupes tel que p% des valeurs soient situées sous p et (100 - p%) des valeurs soient situés au-dessus.

- Les quantiles sont pertinents surtout quand le nombre de valeurs est suffisant pour les calculer de manière précise ($n > 100$)

# Paramètres de position - code

```
age <- c(1, 2, 3, 3, 3, 3, 4, 5, 6, 6, 6, 6, 7, 15)

age_moyen <-  mean(age)
age_moyen
```

```
## [1] 5
```

```
age_median <- median(age)
age_median
```

```
## [1] 4.5
```

# Paramètres de position - code

- Il n'y a pas de fonction pour le mode
- Je vous montrerai en classe comment cela se calcule.

# Paramètres de dispersion

# Paramètres de dispersion

- Bien que la moyenne soit la caractéristique la plus importante résumant une distribution à l'aide d'un seul nombre, il est nécessaire aussi d'étudier comment les observations sont dispersées, ou variées.
- On donne l'exemple d'homme qui s'est noyé dans un ruisseau qui avait en moyenne 10 centimètres de profondeur
- De même qu'il existe différentes mesures de valeur centrale, on trouve de nombreuses mesures de la dispersion.
- deux d'entre elles sont généralement utilisées:
  - l'**intervalle interquartile** et
  - l'**écart type**
- Nous en citerons d'autres tout au long de la présentation

# Étendue

- L'**étendue** (ou *range* ou *amplitude*) est simplement la différence entre la plus grande et la plus petite valeur de la variable.
- Étendue = plus grande observation - plus petite observation

# Étendue Interquartile (EIQ)

- Au lieu d'utiliser les deux observations extrêmes, prenons les deux quartiles.
- les deux quartiles sont beaucoup plus stables (i.e. stables à l'influence indue d'une seule observation).
- La distance séparant les quartiles mesure la dispersion de la moitié centrale des observations: c'est pourquoi on l'appelle **étendue interquartile (EID)**, ou **dispersion centrale**.
- EIQ = 3ème quartile - 1er quartile
- Limite: Elle n'utilise pas l'ensemble des observations de la distribution.

# Variance

- La **variance** est la moyenne arithmétique des carrés des écarts à la moyenne
- Elle mesure la dispersion, l'étalement, et la variabilité des valeurs
- Pour une distribution, la variance est:

$$Variance = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$s^2 = \frac{1}{n-1} * [(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + ... + (X_n - \bar{X})^2]$$

- $X_1, X_2, ..., X_n$ sont les n valeurs observées et $\bar{X}$ = moyenne de la distribution

## Variance

- Pour les données classées, il faut modifier cette formule, en pondérant chaque écart par sa fréquence.

$$Variance, s^2 = \frac{1}{n-1}\sum_{i=1}^{p} f_i(x_i - \bar{x})^2$$

- $x_1, x_2, ..., x_p$ étant les p occurrences observées avec $n_1, n_2, ..., n_p$, les effectifs correspondants de ces occurrences.
- $f_1, f_2, ..., f_p$ sont les fréquences relatives et $\bar{x}$ = moyenne de la distribution groupée (classée)

# Variance

- la variance est elle aussi très sensible aux valeurs extrêmes
- soit la série de 9 valeurs suivantes : 1, 2, 3, 4, 6, 5, 9, 7, 2.
- on trouve :
- moyenne = 4,3 et variance = 7
- si la valeur 9 est plutôt 90, alors la moyenne = 14,1 et la variance = 816,1

## Autres exemples

| Groupe A: | Groupe B | Groupe C |
|---|---|---|
| Relativement homogène | Entre les deux | Relativement hétérogènes |
| 64 | 44 | 34 |
| 68 | 63 | 58 |
| 70 | 80 | 90 |
| 71 | 91 | 101 |
| 69 | 74 | 79 |
| 66 | 56 | 46 |
| **68** | **68** | **68** |

- En gras, moyenne de chaque groupe

# Écart type

- Pour éliminer le fait d'avoir utilisé le carré des écarts, on calcule finalement la racine carrée de la variance: ceci donne la façon la plus générale de mesurer l'écart par rapport à la moyenne, appelée pour cette raison son écart type **s**
- **écart-type** = racine carrée de la **variance**

# Paramètres de position - code

```
age <- c(1, 2, 3, 3, 3, 3, 4, 5, 6, 6, 6, 6, 7, 15)

age_variance <-  var(age)
age_variance
```

```
## [1] 11.53846
```

```
age_ecart_type <- sd(age)
age_ecart_type
```

```
## [1] 3.396831
```

```
# Vérification
sqrt(age_variance)
```

```
## [1] 3.396831
```

# En résumé : Quel type de résumé pour quel type de variable?

| Type de variable | Fréquence | Pourcentage | Commentaire |
|---|---|---|---|
| Nominale | Oui | Oui | Toujours |
| Ordinale | Oui | Oui | Toujours |
| Ratio/Intervalle | Pas souhaité | Pas souhaité | Oui si peu de modalit |
| Ratio/Intervalle (données groupées) | Oui | Oui | Toujours |

# En résumé : Quel type de résumé pour quel type de variable?

| Type de variable | Moyenne | Mode | Médiane | Variance | Écart-type |
|---|---|---|---|---|---|
| Nominale | Non | Oui | Non | Non | Non |
| Ordinale | Possible | Oui | Oui | Possible | Possible |
| Ratio/Intervalle | Oui | Oui | Oui | Oui | Oui |
| Ratio/Intervalle (données groupées) | Oui | Oui | Oui | Oui | Oui |

# En résumé : Mesure de tendance centrale (paramètres de position)

| Symbole | Définition | Formules |
|---|---|---|
| Moyenne | Somme des valeurs divisée par l'effectif de la série | $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ |
| Médiane | Valeur qui divise la distribution en deux parties égales | |
| Mode | Valeur observée de fréquence maximum | |
| Percentile | Valeurs qui divisent la distribution en 100 parties égales | |

# En résumé : Mesure de dispersion

| Symbole | Définition | Formules |
|---|---|---|
| Étendue | Différence entre la plus grande et la plus petite valeur de la variable | G - P |
| EIQ | 3ème quartile - 1er quartile | Q3 - Q1 |
| Déviation | La distance d'une valeur à la moyenne | $X - \bar{X}$ |
| Sommes des carrés | Somme des carrés des déviations | $SC = \sum_{i=1}^{n}(X_i - \bar{X})^2$ |
| Variance | Moyenne des carrés des déviances | $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ |
| Écart-type | Racine carrée de la variance | $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ |

# Application: Labo 4 et 5

## Pour la semaine prochaine

1. Lecture
   - Paramètres de variation (ou de dispersion) - Fox : chapitre 4, pp.91-103
   - Distribution d'échantillonnage - Fox : Chapitre 4, pp.103-120
2. Application
   - https://juba.github.io/tidyverse/01-presentation.html
   - https://juba.github.io/tidyverse/02-prise_en_main.html
   - https://juba.github.io/tidyverse/03-premier_travail.html