# Séance 3: Bref Survol de l'analyse descriptive

## Partie 1: description univariée

Visseho Adjiwanou, PhD.

Département de Sociologie - UQAM

# Base de données et analyse descriptive

# Statistiques

1. Statistiques en tant que nombre
   - Le revenu moyen des habitants de Wakanda est de 1200w
   - la température à Montréal aujourd'hui est de -3 dégrés Celcuis
2. Statistiques en tant que méthodes de calcul
   - Les statistiques sont des méthodes résumant quantitativement et généralisant des informations

# Données

1. Information non résumée, brute que les statistiques rendent plus aisément manipulables.

- Différentes types:
  - Images
  - Vidéo
  - Textes
  - Chiffres ...

2. Banque de données : organisation systématiques des données
3. Fichiers de données : Quand les banques de données peuvent être lues par les ordinateurs
4. Unités d'analyse : est la personne, l'objet ou l'événement que le chercheur étudie.

# Sources de données en sciences sociales

1. Données que vous collectez vous-mêmes
2. Données qui existent déjà

# Collecter vos propres données

- Avantages
  - Vous collectez ce qui vous intéresse si vous devez faire une collecte formelle
  - Peut aussi recourir à collecter les données des médias et réseaux sociaux
- Inconvénients
  - Peut demander beaucoup de temps de préparation
  - Peut demander de la programmation
  - Coûteux
  - Disponibilités de multiples données qui existent déjà, pourquoi ne pas utiliser une de ses données?

# Collecter vos propres données

- Exemple : Collecter les données twitter sur le premier ministre Trudeau

# Exemples de données qui existent déjà

1. Sur les pays en développement
   - Enquêtes démographique et de santé
   - https://dhsprogram.com/data/
2. Sur le Canada
   - Recensements
   - Enquêtes sociales générales
   - Pleins d'autres
   - Sondage d'opinions
     - https://www.queensu.ca/cora/our-data/data-holdings
3. Sur les USA
   - http://www.pewresearch.org/

# Survol des statistiques

Deux branches des statistiques:

1. Statistiques descriptives : méthodes résumant l'information afin de la rendre plus intelligible, plus utile ou plus aisément communicable.
    - Exemple: Age moyen des étudiants de ma classe.
- Cependant, il y a perte d'information.
- Choix judicieux du type de statistiques à utiliser

# Survol des statistiques

2. Statistiques inférentielles renvoient aux procédures par lesquelles nous généralisons l'information concernant un échantillon à la population de laquelle fut tiré l'échantillon en question.
   - marge d'erreur; p value ou p <0,05; test de chi-carré, chi 2 . . .
3. Le cours concerne à la fois les statistiques descriptives et les statistiques inférentielles

# Les échantillons et les populations

1. Données de population : proviennent de tous les cas auxquels un chercheur veut appliquer ses conclusions: on dit dans ce cas qu'on fait un **recensement** de la population.

   - Données parfois introuvables
   - Onéreuses
   - Longues à collecter

2. Données d'échantillon : provient d'une partie de la population.

3. Paramètre : résumé basé sur une population

4. Statistique (3e définition du mot) : caractéristique d'un échantillon

   - Exemple : âge moyen (paramètre si calculé sur toute la population) et statistique (si calculé à partir d'un échantillon)

# Les variables

- Une variable est une caractéristique ou une propriété quelconque dont la valeur diffère d'un cas à l'autre.
- Elles se retrouvent souvent en colonnes dans les fichiers de données
- Echelle : série des valeurs possibles d'une variable.
- Scores : valeur possible d'une variable

# Les niveaux de mesures

Quatre catégories de variables selon la façon dont elles sont mesurées:

1. Variables nominales : se mesure de telle façon que ses valeurs ou ses attributs diffèrent les uns des autres. Les valeurs ne peuvent pas être disposées selon un ordre logique ou naturel.

- variable nominale dichotomique: prend deux valeurs. ex, sexe
- variable nominale non dichotomique: ex. religion

# Les niveaux de mesures

2. Variable ordinale : variables dont les valeurs peuvent êtres ordonnées.

   - classe sociale
   - quintile de revenu

Remarques:

- Variable nominale + variable ordinale = variable qualitative.
- Variables ordinales plus informatives que les variables nominales

# Les niveaux de mesures

Quatre catégories de variables selon la façon dont elles sont mesurées:

3. Variable d'intervalle a non seulement des valeurs qui peuvent être ordonnées, mais elle se mesure également à l'aune d'une unité de mesure fixe ou standard.

- Température
- Date de naissance

# Les niveaux de mesures

4. Variable de ratio : est semblable à une variable d'intervalle. Mais, en plus, elle a un zéro non arbitraire.

- Nombre d'habitants
- Âge

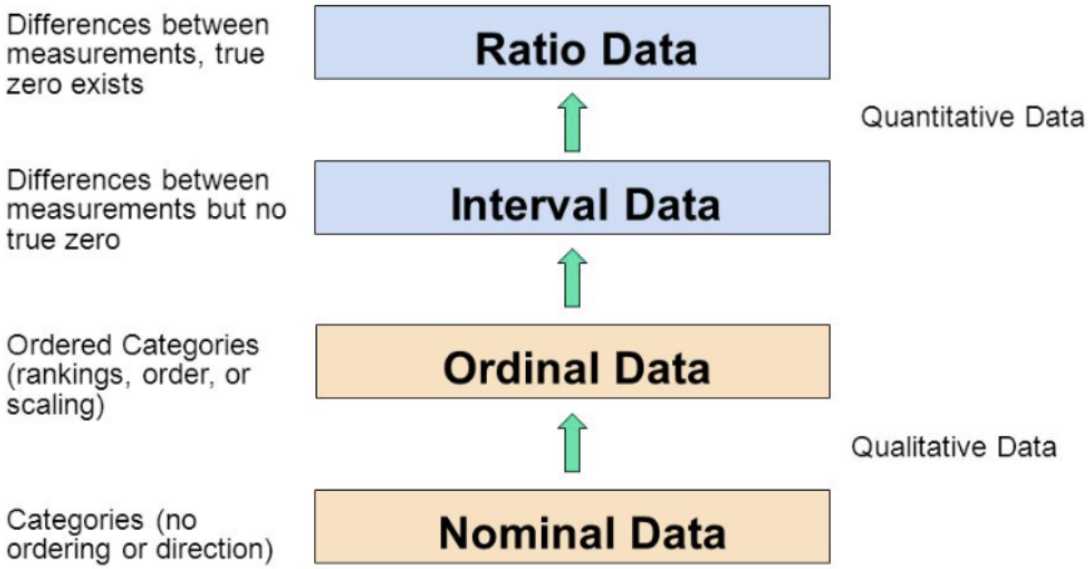
# Les niveaux de mesures

Remarques:

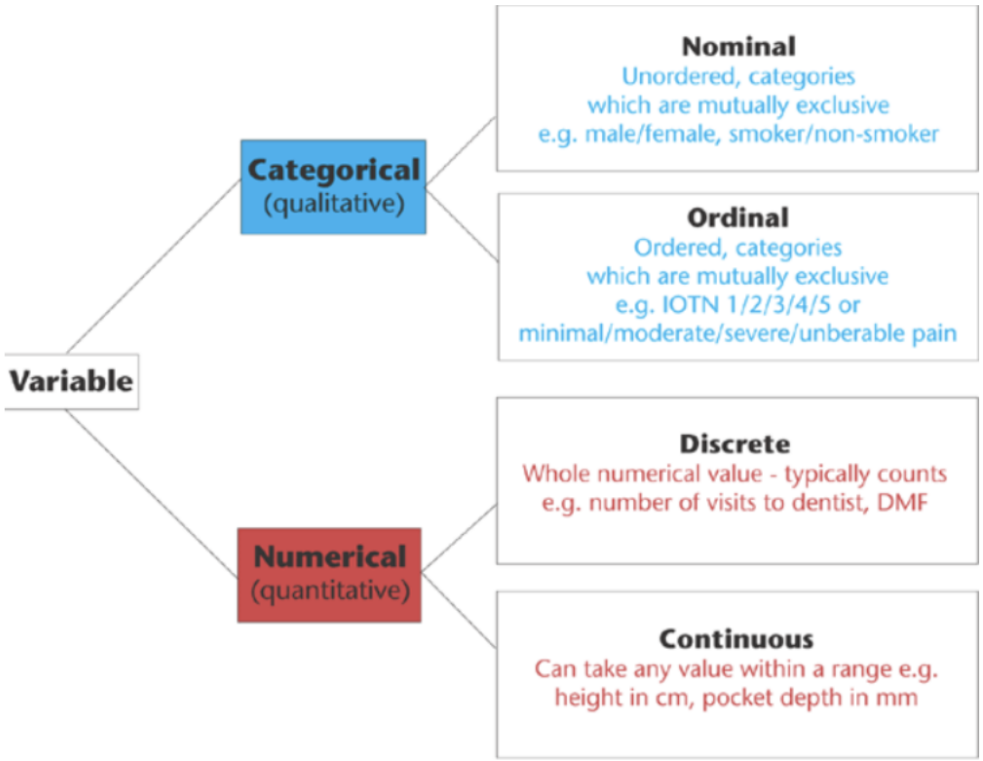
- Variable de ratio car permettent de calculer des ratios
- variable d'intervalle + variable de ratio == variable quantitative
- Peuvent être continues quand elles peuvent prendre une infinité de valeurs ou discrètes quand elles prennent un nombre fini de valeurs.

# Résumé

https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/



# Résumé


Variable
Categorical
(qualitative)
Nominal
Unordered, categories
which are mutually exclusive
e.g. male/female, smoker/non-smoker
Ordinal
Ordered, categories
which are mutually exclusive
e.g. IOTN 1/2/3/4/5 or
minimal/moderate/severe/unberable pain
Numerical
(quantitative)
Discrete
Whole numerical value - typically counts
e.g. number of visits to dentist, DMF
Continuous
Can take any value within a range e.g.
height in cm, pocket depth in mm

# Les niveaux de mesures

- Il est capital de distinguer le type de variable que vous utilisez pour être en mesure d'utiliser la bonne statistique.
- On verra l'application dans R.

# Catégories mutuelllement exclusives et collectivement exhaustives

Les valeurs des variables que vous définissez doivent être :

- **mutuellement exclusives** : pas de chevauchement
- **collectivement exhaustives** : comprend l'ensemble des catégories

Remarque:

- Quand vous créez de nouvelles variables, vous devez toujours tenir compte de cela.

# Validité et fiabilité (validity and reliability)

# Validité / fiabilité

- Validité : Le degré auquel une variable mesure ce que nous pensons qu'elle mesure. Est-ce que la variable reflète le concept ?
- Fiabilité: Le degré auquel une variable donne des résultats cohérents. D'autres chercheurs doivent être en mesure d'effectuer exactement la même expérience dans les mêmes conditions en obtenant les mêmes résultats.

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Lorsque vous essayez de mesurer des concepts difficiles comme le statut social, il est souvent préférable

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Lorsque vous essayez de mesurer des concepts difficiles comme le statut social, il est souvent préférable
- d'utiliser une variété de mesures qui peuvent être analysées indépendamment ou

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Lorsque vous essayez de mesurer des concepts difficiles comme le statut social, il est souvent préférable
- d'utiliser une variété de mesures qui peuvent être analysées indépendamment ou
- combinées en une seule mesure globale (ou composite).

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Lorsque vous essayez de mesurer des concepts difficiles comme le statut social, il est souvent préférable
- d'utiliser une variété de mesures qui peuvent être analysées indépendamment ou
- combinées en une seule mesure globale (ou composite).
- Par exemple, de nombreux sociologues utilisent une combinaison de revenu, d'éducation et de profession pour déterminer le statut socio-économique global du répondant.

# Validité / fiabilité

- Il est important de se rappeler que les mesures peuvent être valables/valides sans être fiables, de même, les mesures peuvent être fiables sans être valides.
- Le but est de viser des niveaux élevés de validité et de fiabilité pour éviter le problème du “garbage in, garbage out”.

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Quel est ton revenu l'année passée?

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Quel est ton revenu l'année passée?
- Quel est ton revenu la semaine passée?

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Quel est ton revenu l'année passée?
- Quel est ton revenu la semaine passée?
- Les questions ne sont pas valides: Qu'en est-il des revenus des parents

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Diras-tu que tu fais partie de la classe ouvrière, classe moyenne, ou de la classe supérieure?

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Diras-tu que tu fais partie de la classe ouvrière, classe moyenne, ou de la classe supérieure?
- Majorité vont dire classe moyenne

# Validité / fiabilité

- Exemple : Supposons que nous voulons mesurer la position sociale d'un groupe d'étudiants. Quelles questions pouvons nous poser pour la mesurer?
- Diras-tu que tu fais partie de la classe ouvrière, classe moyenne, ou de la classe supérieure?
- Majorité vont dire classe moyenne
- Es-tu financièrement en sécurité? très subjective donc pas fiable

# Les données individuelles et les données agrégées

- Dans la plupart des cas, nous travaillons avec des données individuelles.
- Dans certains cas, nos données sont agrégées (PIB par pays par exemple)
  - Données écologique si l'unité d'agrégation est l'espace
  - Donne lieu à des erreurs écologiques: inférer sur les individus les résultats agrégés.

# Statistiques descriptives

# Objectifs

Les objectifs de la statistique descriptive sont de :

- définir le ou les groupes étudiées (population ou échantillon)
- définir le codage des observations
- définir la présentation des données : numérique et/ou graphique
- réduire les données à quelques indicateurs statistiques synthétiques

# Objectifs

La description des données

- souvent la première approche dans la compréhension d'un phénomène
- réduction des données à quelques indices numériques permettant de manipuler les données
- permettra la formulation d'hypothèses qui pourront être vérifiées à l'aide de tests statistiques lors d'études organisées ultérieurement

# Les distributions de fréquences (Chap 2 : Fox)

- Fox fait référence au livre du bac
- Fox, W. 1999. Statistiques sociales. Les Presses de l'Université Laval. Traduit de l'Anglais et adapté par L.M. Imbeau.
- Façon simple et directe de résumer les informations d'une variable
- Il s'agit de compter le nombre de cas pour chaque valeur ou modalité
- Ce résumé s'appelle **distribution de fréquence**
- Approprié pour les cas avec peu de modalité (donc pour les variables **ordinales** ou **nominales**)

# Les distributions de fréquences (Chap 2 : Fox)

Tableau 2.5. Niveau d'instruction atteint (en fréquences)

| Niveau d'instruction | f |
|---|---|
| Universitaire avancé | 4 |
| Premier cycle universitaire | 9 |
| Collège | 3 |
| Secondaire | 24 |
| Pas de secondaire | 10 |
| Total | 50 |

# Les distributions de fréquences

- **Avantages**
  - Facile à calculer
  - Donne plusieurs indications sur :
    - les cas fréquents (mode)
    - là où la distribution est coupée en deux (médiane, avec les fréquences ou pourcentages cumulés)
    - les cas rares (donc besoin de regroupement)
    - les cas déviants (cas d'une variable d'intervalle/ratio)
  - Permet de détecter les données manquantes (cependant, il faut enlever ces cas avant de calculer les fréquences)
- **Désavantages**
  - Difficile à interpréter surtout avec les grands nombres
  - Ne permet pas les comparaisons

# Les distributions de pourcentages

- Pour remédier à cela, on va calculer les pourcentages ou proportions
- Se calcule par le rapport entre le nombre de cas et le nombre de cas total (proportion)
- Si multiplié par 100, cela devient des pourcentages

# Les distributions de pourcentages

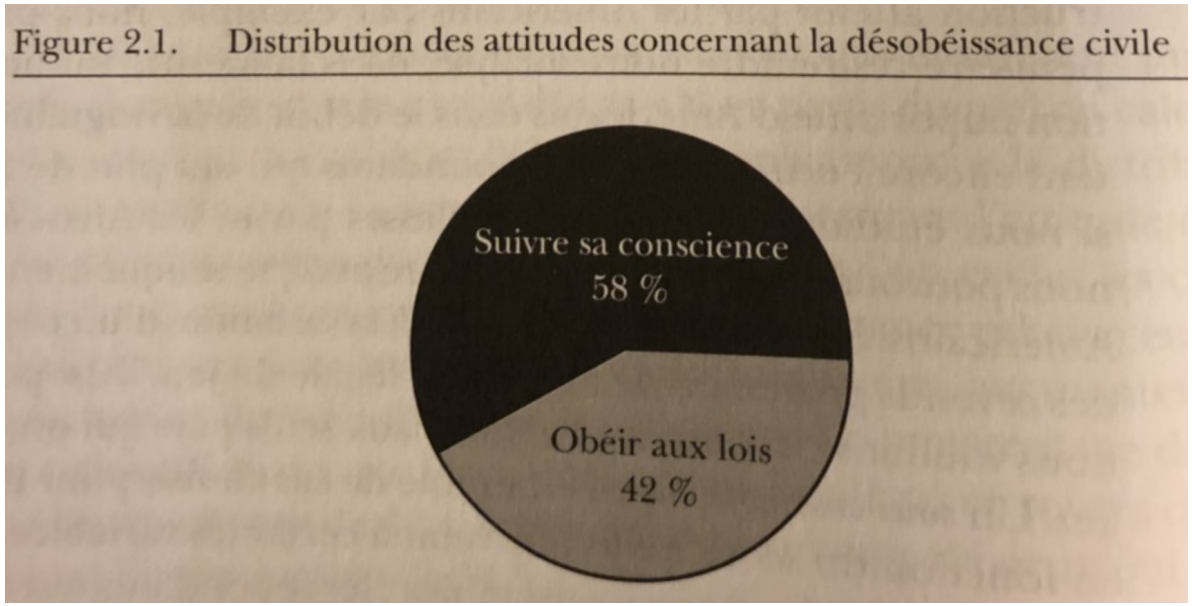
Tableau 2.7. Niveau d'instruction atteint (en pourcentages)

| Niveau d'instruction | Pourcentages |
|---|---|
| Universitaire avancé | 8 |
| Premier cycle universitaire | 18 |
| Collège | 6 |
| Secondaire | 48 |
| Pas de secondaire | 20 |
| Total | 100 |
| (N) | (50) |

# Représentation (voir cours prochain)

Se représente graphiquement par :

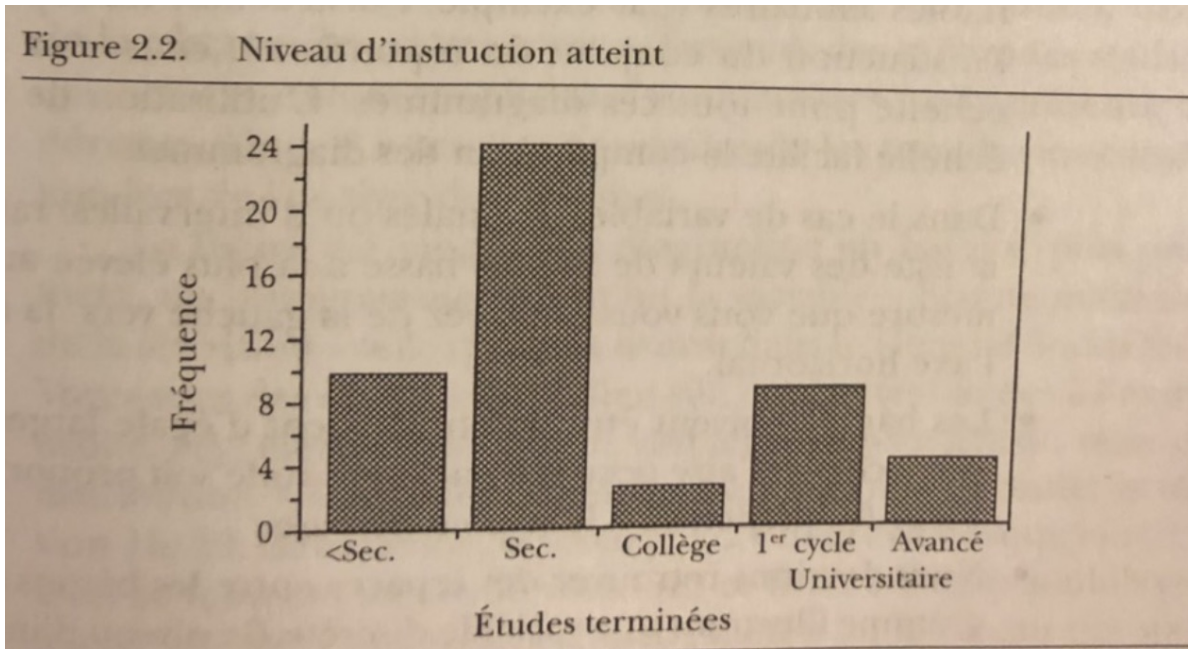
1. les diagrammes circulaires



# Représentation

Se représente graphiquement par :

2. les diagrammes en bâtons (ou diagrammes de barres - barplot)



# Description

- Lire page 65-68 dans Fox.

# Paramètres de position (Chap 3, Fox)

Les variables continues (intervalle/ratio) sont décrites numériquement par :

1. des **paramètres de position**

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

# Paramètres de dispersion (Chap 4, Fox)

Mais aussi par :

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

**La Moyenne (arithmétique)** = Somme des valeurs divisée par l'effectif de la série - Exemple : moyenne de 4, 7, 6. M = (4+7+6)/3

**La Médiane** = valeur telle que la moitié des observations lui sont inférieures et donc la moitié lui sont supérieures. - Exemple 1: médiane de 3, 6, 4, 8, 9 est **6** - Exemple 2: médiane de 5, 7, 8, 9 est (7+8)/2 = **7.5**

# Paramètres de position

**Le Mode** = Encore appelée valeur dominante: valeur observée de fréquence maximum telle que la moitié des observations lui sont inférieures et donc la moitié lui sont supérieures.

```
- Exemple : 1, 2, 3, 3, 3, 3, 4, 5, 6, 6, 6, 6, 7, 15 : mod
```

```
- On parle de distribution bimodale
- On peut penser que l'échantillon est en réalité issu de d
- Si toutes les valeurs sont différentes, autant de modes q
- 1, 2, 3, 5, 6, 9, 14, 16 --> chaque valeur = mode
```

- la seule pour laquelle il n'y a pas de formule

# Paramètres de position

**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

**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.

**Quantiles / Fractiles** Le quantile d'ordre *k* est la valeur qui sépare la distribution en k classes de même efféctif (au moins approximativement). - déciles, - quartiles, - quintiles, - tiertiles, - centiles, etc.

# 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.
- De même qu'il existe différentes mesures de paramètres de position, on trouve de nombreuses mesures de la dispersion.
- Deux d'entre elles sont généralement utilisées:
  - l'**intervalle interquartile** et
  - l'**écart type**

# Paramètres de dispersion

- L'**étendue** (ou *range* ou *amplitude*) est simplement la différence entre la plus grande et la plus petite valeur de la variable.
- 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 (EIQ)**, ou **dispersion centrale**.
  - EIQ = 3ème quartile - 1er quartile

# Paramètres de dispersion

- 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:

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

# Paramètres de dispersion

- la variance est elle aussi très sensible aux valeurs extrêmes
- 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**
  - écart type = **sqrt**(variance)

# Pour la semaine prochaine

# Pour la semaine prochaine

1. **Lectures obligatoires**
   - Fox(p123-172)
2. **Lectures Facultatives (important pour mieux assimiler le cours)**
   - Kieran (https://socviz.co/lookatdata.html#what-makes-bad-figures-bad
   - Wickham (https://r4ds.had.co.nz/data-visualisation.html)
   - Wickham (https://r4ds.had.co.nz/exploratory-data-analysis.html)
3. **Datacamp**
   - Introduction à Tidyverse