Chapter 1: Univariate Analysis

Introduction to Quantitative Methods

Maxime Chabriel

2024

Introduction

- **Univariate analysis :** When we study only one variable
- Distribution : How the values of a variable on multiple observations are spread
- This chapter will be about understanding what is a distribution and how to simplify / study it
- Can you survive this class without understanding what is a distribution?
  - Yes
  - But if you do understand, everything else becomes easy to learn

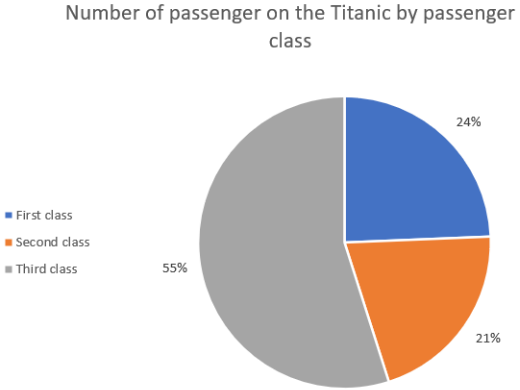
Representing a distribution

(NB this is by all means not an exhaustive list)

A Galton board

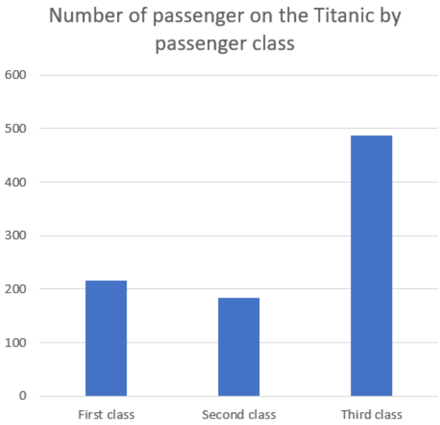
Categorical / Discrete variables

- The pie chart (circular graph, camembert). Good to represent proportions



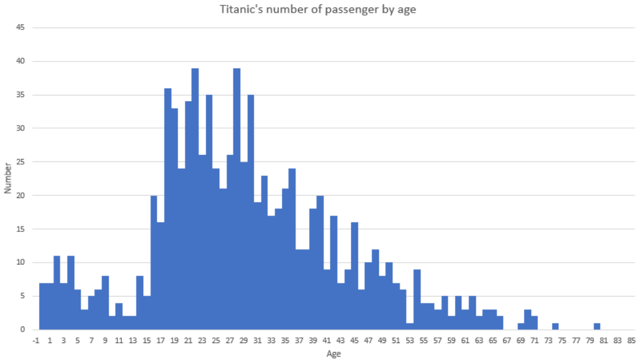
Categorical / Discrete variables

- The bar chart



Categorical / Discrete variables

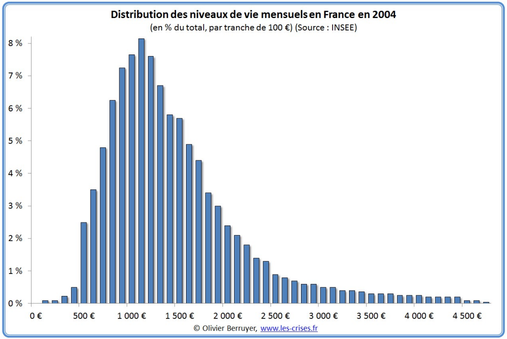
- The bar chart. Good to represent ordinal categorical variables or categorical variables with many modalities



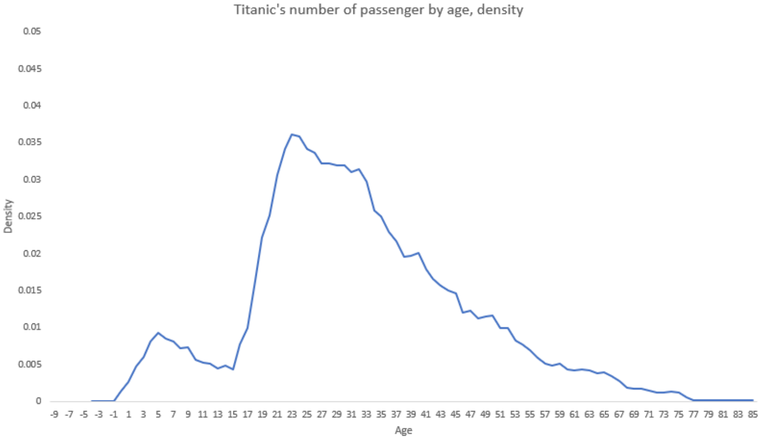
- An age pyramid: https://www.insee.fr/fr/outil-interactif/5014911/pyramide.htm!y=1991v=2l=enc=0

Continuous values

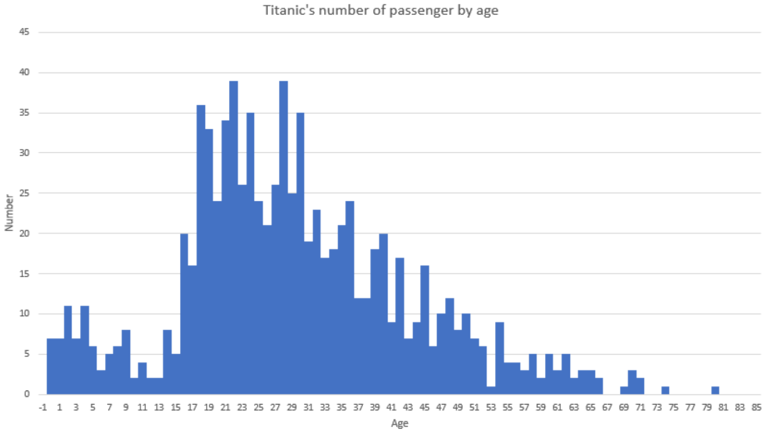
- Bar charts work fine on continous variables too (we then call them histograms). You just need to create some intervals (connoisseurs call them bins)



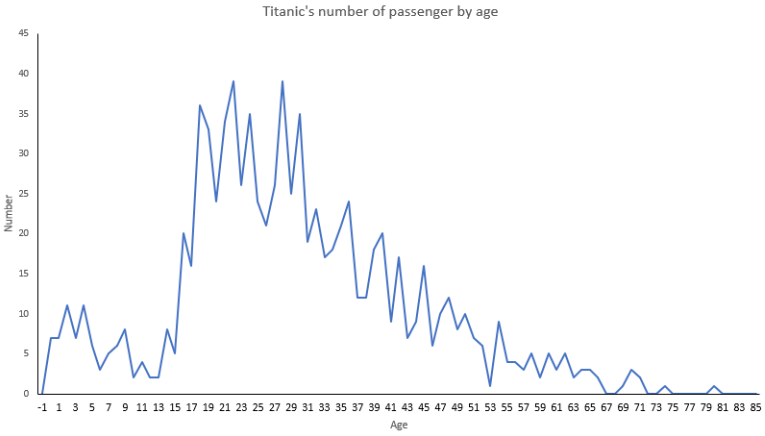
- But otherwise, in practice we use a **density**



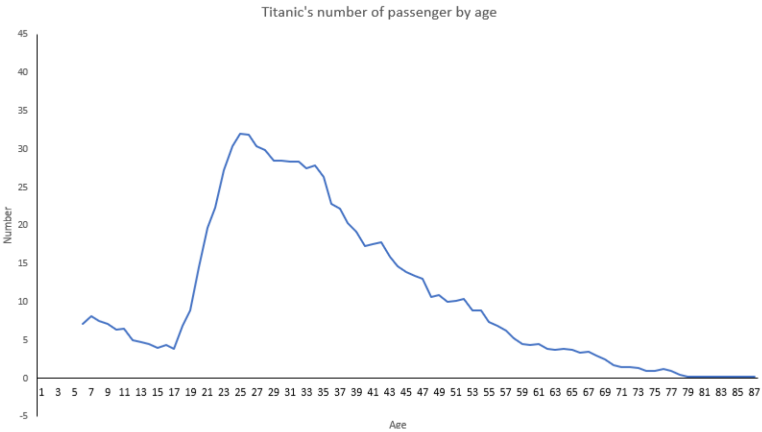
- How it is built (intuitive answer): We start from histogram bars



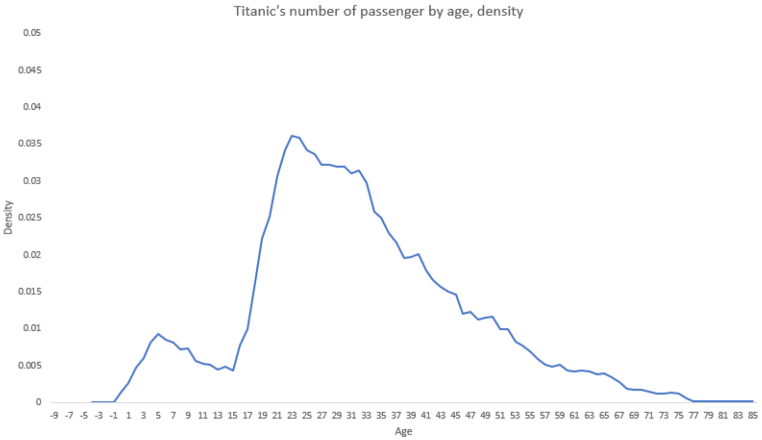
- How it is built (intuitive answer): We link the datapoints to create a line plot



- How it is built (intuitive answer): We smooth the line plot



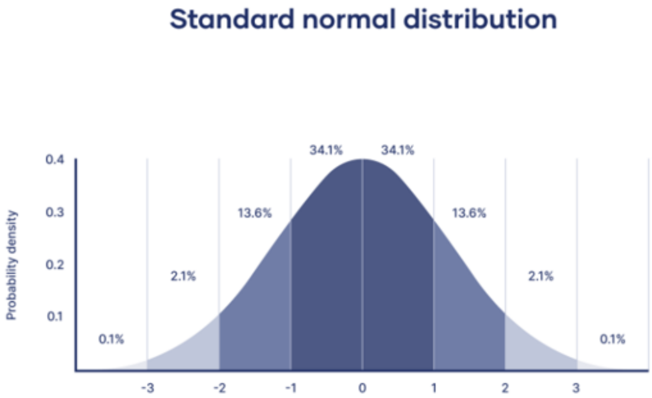
- How it is built (intuitive answer): We rescale the y axis and stretch the plot so that is begins and ends by 0

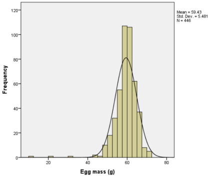


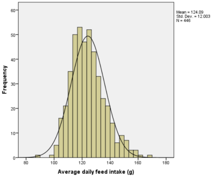
- But otherwise we use a **density plot**
- How it is built (easy answer): smooth interpolation of the rescaled histogram bars
  - Afik you cannot implement a density on Excel. We will do it in R in the final classes
  - NB There are many algorithms out there to build a density, you do not need to know any. But you need to know how to read one
- How to read it:
  - A bit like a histogram (although the y axis is not directly readable)
  - The area between a point a and b on the x axis represents the % of the observations that are greater than a and smaller than b

Example with the gaussian / normal density:

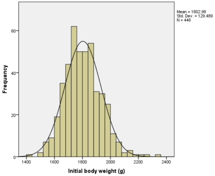
- This density is very famous because it follows a lot of mathematical properties
    - It is symetric
    - It has infinite but thin tails
- And it is frequently found in nature!
    - The height of people of similar age and sex
    - The measurement error of the weight of a supermarket pasta bag
    - By how much you miss when you play darts
    - The random movements of atoms, particles, fluids...
    - Some financial identicators (or are they?)
    - etc.

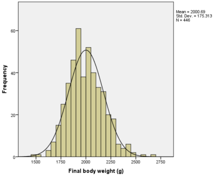

Standard normal distribution
Probability density
0.4
0.3
0.2
0.1
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
-3
-2
-1
0
1
2
3


Mean = 59.43
Std. Dev. = 5.481
N = 446
Frequency
Egg mass (g)

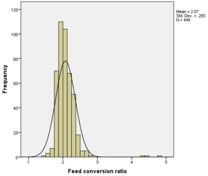

Mean = 124.09
Std. Dev. = 12.003
N = 446
Frequency
Average daily feed intake (g)

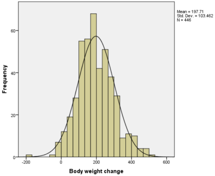
(c) **(d)**

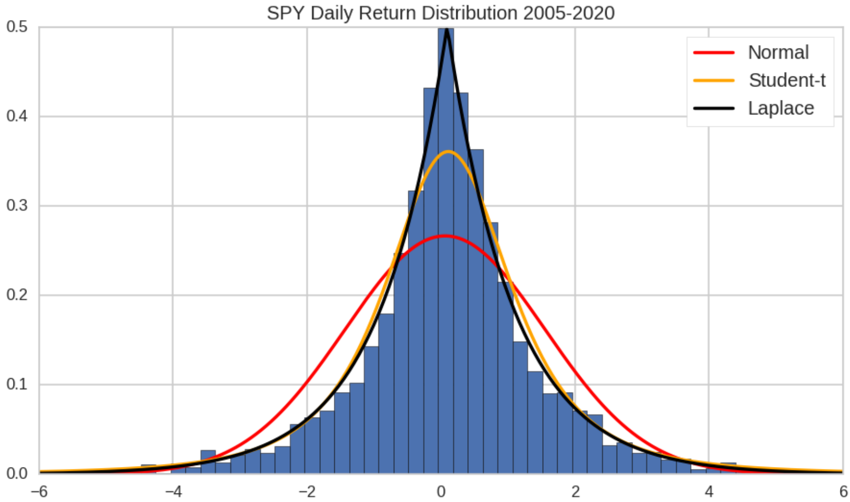

Mean = 1802.98
Std. Dev. = 129.489
N = 446
Frequency
Initial body weight (g)


Mean = 2000.69
Std. Dev. = 175.313
N = 446
Frequency
Final body weight (g)

(e) **(f)**


Mean = 2.07
Std. Dev. = .285
N = 446
Frequency
Feed conversion ratio


Mean = 197.71
Std. Dev. = 103.462
N = 446
Frequency
Body weight change


SPY Daily Return Distribution 2005-2020
Normal
Student-t
Laplace
0.5
0.4
0.3
0.2
0.1
0.0
−6
−4
−2
0
2
4
6

Statistics

(Congrats for making it this far, now the interesting stuff)

- a **statistic**: A value built from a distribution. A good statistic gives interesting information on the distribution. Everything below *are* statistics...

- **maximum/minimum**

- **mean**: What you would get if you shared all the values equally betwen each observation

$$E(X) = \sum_{i=1}^{N} \frac{X_i}{N}$$

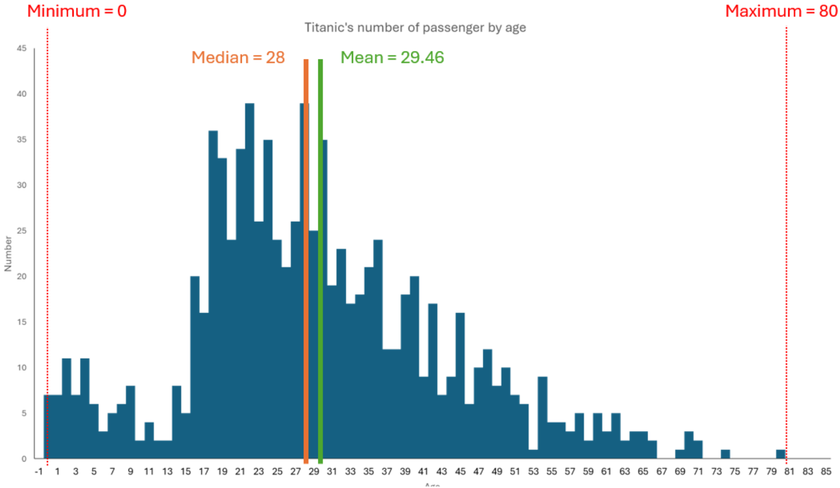
X is the distribution of N values and X_i is the ith value of X. The mean is equal to the sum of the values of the distribution divided by the number of values in the distribution

- **median**: The value greater than or equal to 50% of the values of the distribution *and* the value smaller than or equal to 50% of the values of the distribution. When there are competing choices, we take the average of these choices

Name	Age
Gus	3
Alice	10
Bob	15
Harry	23
Denis	28
Clara	37
Francis	42
Erwan	59
Irene	62

Average: $\frac{3 + 10 + 15 + 23 + 28 + 37 + 42 + 59 + 62}{10} = 31$

Median


Minimum = 0
Maximum = 80
Titanic's number of passenger by age
Median = 28
Mean = 29.46
Number
0
5
10
15
20
25
30
35
40
45
-1 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85

- **mode(s)**: Where the density gets its highest value. For discrete variables, this is the most frequent value in a distribution. NB There can be several modes

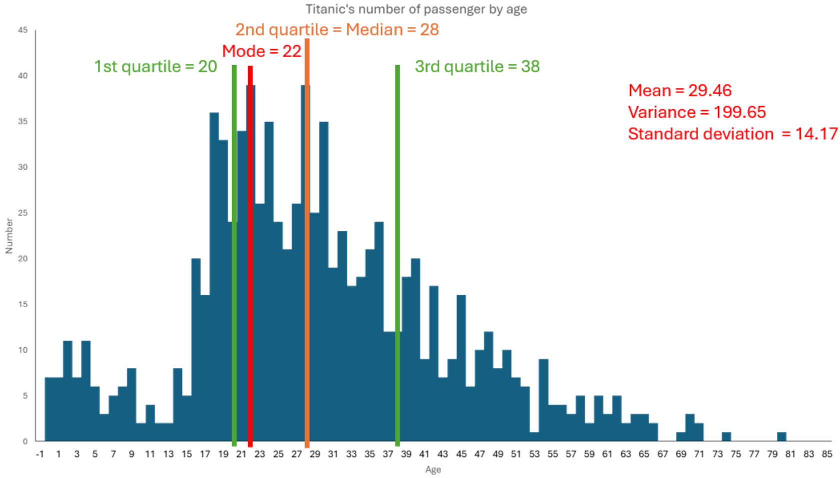
- **quantiles**: Values that divide a distribution in equal groups
  - Example: The median divides the distribution into 2 equal groups. The median *is* a quantile
  - Other types of quantiles: quartiles (divide in 4), quintiles (in 5), deciles (in 10), centiles (in 100)...
  - *median = 2nd quartile = 5th decile = 50th centile !*

- **variance**: How much the values of a distribution are spread out
  - You might also encounter the **standard deviation**. This statistic can be *very approximately* understood as the average distance between any value and the distribution's mean

$$V(X) = \sum_{i=1}^{N}(X_i - E(X))^2$$

$$V(X) = \sigma(X)^2$$

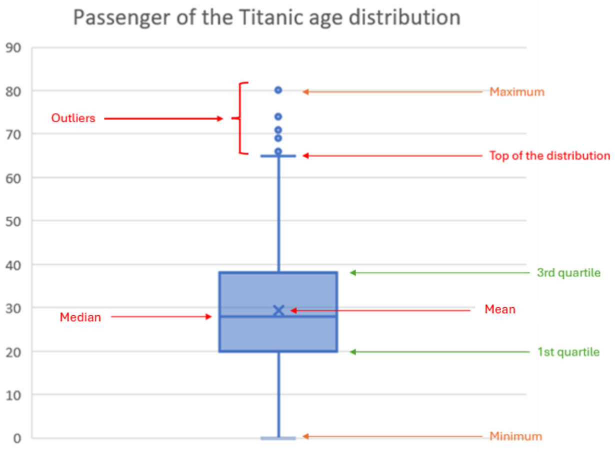
Example on Excel


Titanic's number of passenger by age
2nd quartile = Median = 28
Mode = 22
1st quartile = 20
3rd quartile = 38
Mean = 29.46
Variance = 199.65
Standard deviation = 14.17
Number
Age

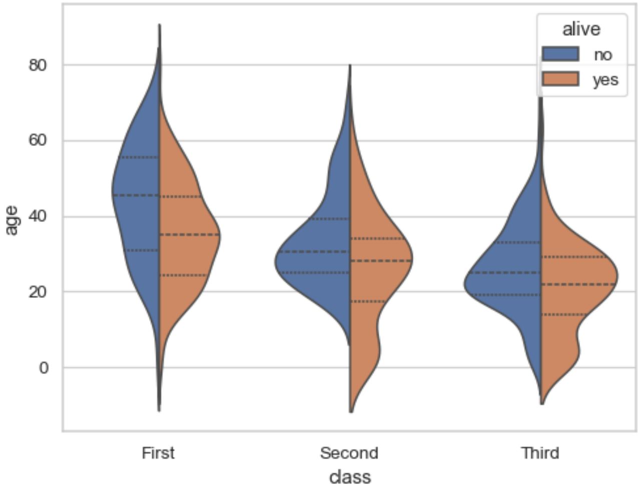
- **Outlier**: Values abnormally high or low. There is no absolute criterion to define an outlier. You can set them aside in your analysis **if and only if** you justify it properly and **always** say it when you do. You are that close from being the new Lysenko or Didier Raoult. Examples of what you are allowed to do:
  - The outliers exist because of an error in your data. Someone cheated at all his final exams and got better grades than everyone else
  - The outliers exist because your observations are not comparable. If you study sociodemographic values of some countries, you may set aside microstates
  - The outliers exist because of a random event. You compare currency trends and one country in your dataset is experiencing hyperinflation

- But beware if:
  - The outliers are throwing off the scales of your graphs but they are important to your analysis. You study wealth distribution in a country and there are a few billionaires in your dataset
  - Some softwares/algorithms might automatically detect outliers using probabilistic criteria (units of standard deviation). These will always be mere suggestions, you make the final say!
  - When in doubt, don't

All the main statistics can be condensed into a **boxplot** (aka whisker plot, boîte à moustache):



Variations of the boxplot do exist. The only limit is your imagination! "Violin plots" done in R:



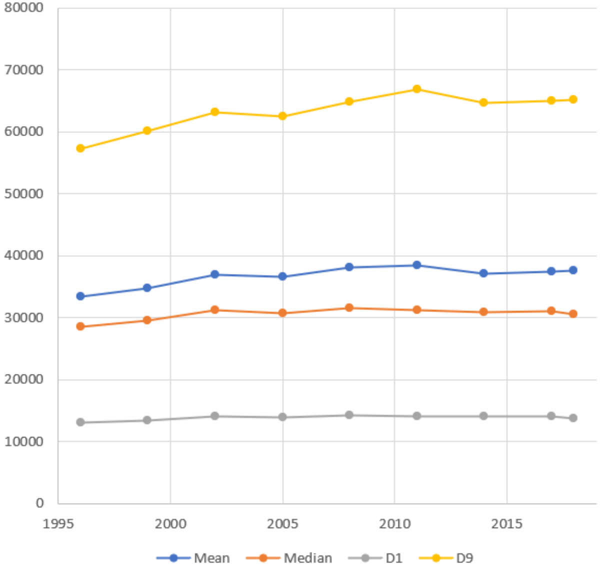
Choosing a statistic

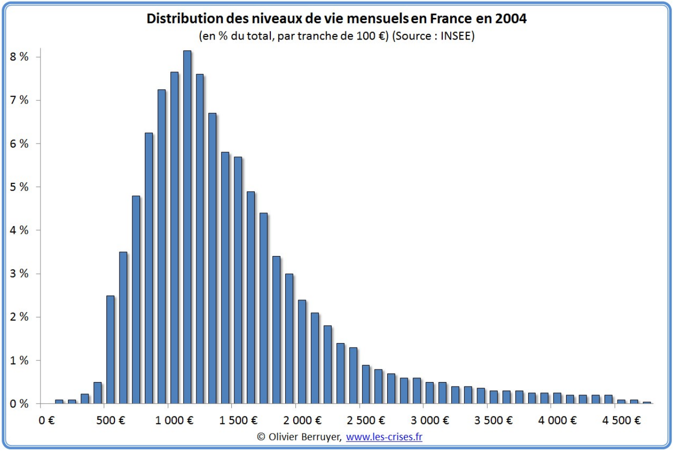
In practice, what is the difference between the mean and the median? Why do their values differ?

In practice, what is the difference between the mean and the median? Why do their values differ?

- The mean is sensible to extreme (very high or very low) values
- The median is not very useful when there is a lot of dispersion
- If the mean is greater than the median, there are some extreme upper values. We say that the distribution is *right-skewed*
- When someone only shows a mean, a median, or any other statistic, ask yourself, what about the rest of the data?

Evolution of available income in French households (INSEE, euro 2018):




Distribution des niveaux de vie mensuels en France en 2004
(en % du total, par tranche de 100 €) (Source : INSEE)
8 %
7 %
6 %
5 %
4 %
3 %
2 %
1 %
0 %
0 €
500 €
1 000 €
1 500 €
2 000 €
2 500 €
3 000 €
3 500 €
4 000 €
4 500 €
© Olivier Berruyer, www.les-crises.fr

Other famous statistics/inequality metrics:

- **Gini coefficient**: Only for positive values (income for instance). 0 represents perfect equality (all the values of the distribution are the same). 1 means that all values except one equal 0 (1 observation has 100% of the wealth)
- **Interdecile ratio**: D9 divided by D1. How many times the 90th percentile is bigger than the 10th percentile
- **Herfindhal index**: To measure how much a sum (a cake) is being divided. Mostly used in market competition analysis. 0 means that there are infinite values with equal shares (atomistic competition). 1 means there is only 1 observation with 100% of the share (there is a monopoly)

So how to choose a good statistic?

- It depends on what you want to show. Or which measure has the most striking value (whether you like it or not, statistics are politics)
- You do not need to choose only one. The more you include statistics in your analysis, the more your analysis will be convincing. Nuance is important!
- Some measures are easier to understand than others. Think about your audience. Ex: Variance vs Interdecile ratio

Comparing the distribution of variables

Some statistics can be directly used to compare variables:

- Gini index
- Interdecile ratio

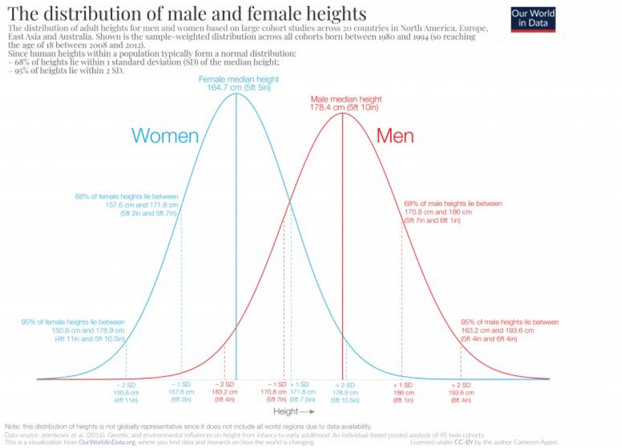
Most are not, because they are **scale dependent**:

- Mean, median
- Variance

These are statistics that, if you changed the unit of the variable (km to m, € to $, min to s...), they would change value! For these, you need to put the variables you compare in **the same unit**.

Depending on the context, you may also want to answer these questions:

- Is the distribution **symetric**?
- The **tails** of the distribution. Do they exist? Are they fat?
- The **skewness**. Are the values more concentrated on the right (left skew)? Or the left (right skew)?
- Does the distribution looks like a **normal** (gaussian) distribution?



The distribution of male and female heights
The distribution of adult heights for men and women based on large cohort studies across 20 countries in North America, Europe, East Asia and Australia. Shown is the sample-weighted distribution across all cohorts born between 1980 and 1994 (so reaching the age of 18 between 2008 and 2012).
Since human heights within a population typically form a normal distribution:
– 68% of heights lie within 1 standard deviation (SD) of the median height;
– 95% of heights lie within 2 SD.
Our World in Data
Female median height 164.7 cm (5ft 5in)
Male median height 178.4 cm (5ft 10in)
Women
Men
68% of female heights lie between 157.6 cm and 171.8 cm (5ft 2in and 5ft 7in)
68% of male heights lie between 170.8 cm and 186 cm (5ft 7in and 6ft 1in)
95% of female heights lie between 150.6 cm and 178.9 cm (4ft 11in and 5ft 10.5in)
95% of male heights lie between 163.2 cm and 193.6 cm (5ft 4in and 6ft 4in)
– 2 SD 150.6 cm (4ft 11in)
– 1 SD 157.6 cm (5ft 2in)
– 2 SD 163.2 cm (5ft 4in)
– 1 SD 170.8 cm (5ft 7in)
+ 1 SD 171.8 cm (5ft 7.5in)
+ 2 SD 178.9 cm (5ft 10.5in)
+ 1 SD 186 cm (6ft 1in)
+ 2 SD 193.6 cm (6ft 4in)
Height
Note: this distribution of heights is not globally representative since it does not include all world regions due to data availability.
Data source: Jelenkovic et al. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts.
This is a visualization from OurWorldinData.org, where you find data and research on how the world is changing.
Licensed under CC-BY by the author Cameron Appel.

Finally (reminder from the introduction), **always** ask yourself:

- Are there some missing values?
- Is my sample representative of the population I am studying?
- Always question the quality / source of your data

Your turn, practice time