

# **Week 2:**

# **Look, understand and make sense of your data!**

Phase 1

# Key ideas of this lecture

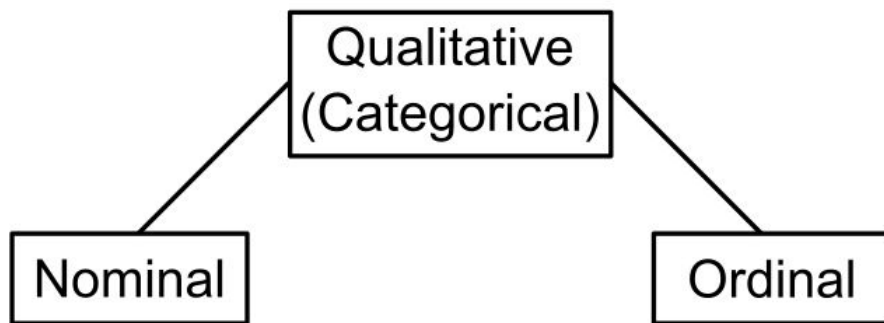
- Descriptive statistics: first step after collecting the data.
- Visualization and summary: better understanding and communication.
- Key concepts: Tendency, variability and shape.

# Understand the data: types of variables

It is important to have clear what kind of variables we have in our data.

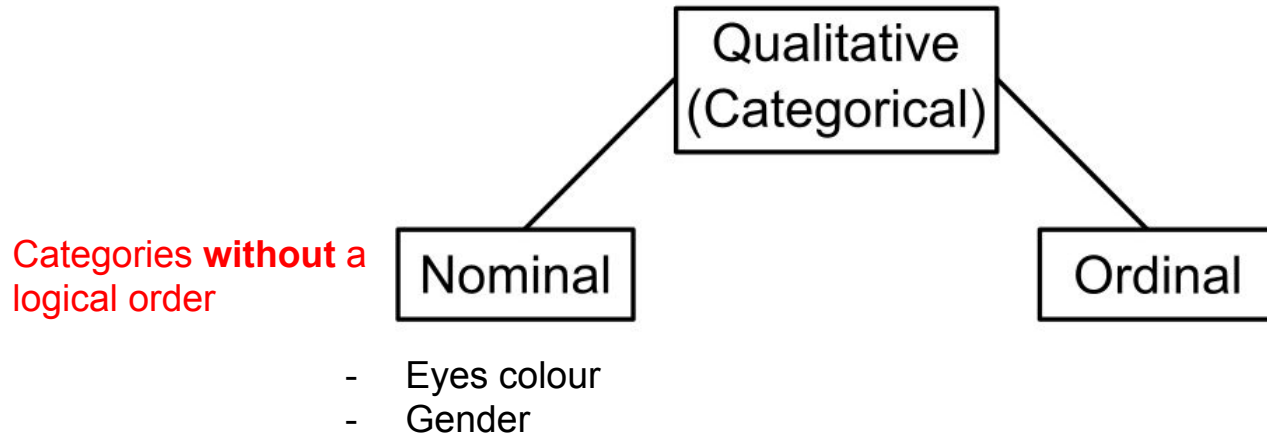
# Understand the data: types of variables

Do variables contain only categories or groups for which adding and averaging do not make sense? Then we have **qualitative** data.



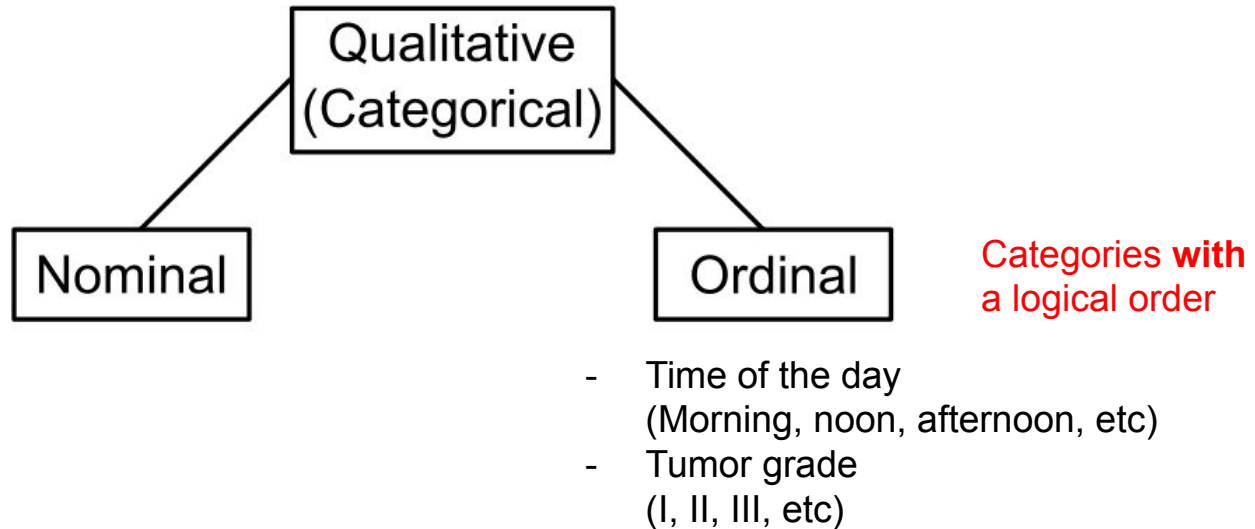
# Understand the data: types of variables

Do variables contain only categories or groups for which adding and averaging do not make sense? Then we have **qualitative** data.



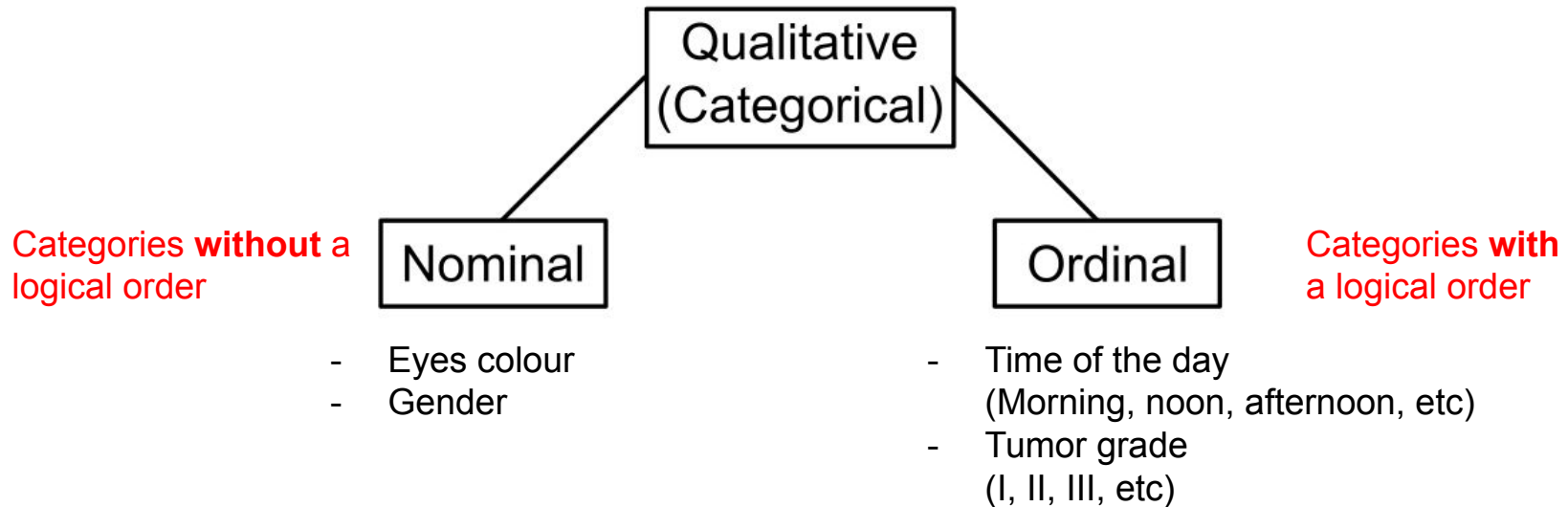
# Understand the data: types of variables

Do variables contain only categories or groups for which adding and averaging do not make sense? Then we have **qualitative** data.



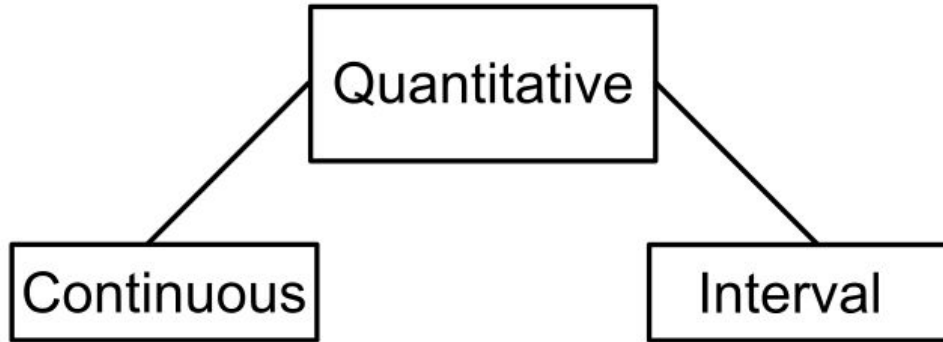
# Understand the data: types of variables

Do variables contain only categories or groups for which adding and averaging do not make sense? Then we have **qualitative** data.



# Understand the data: types of variables

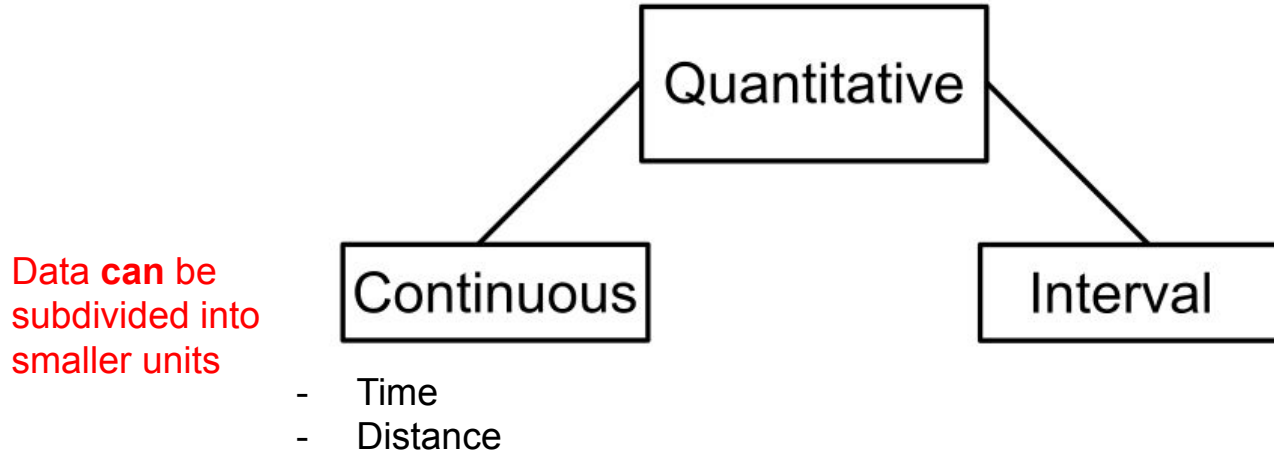
Do variables take numerical values for which arithmetic operations such as adding and averaging make sense? Then we have **quantitative** data.





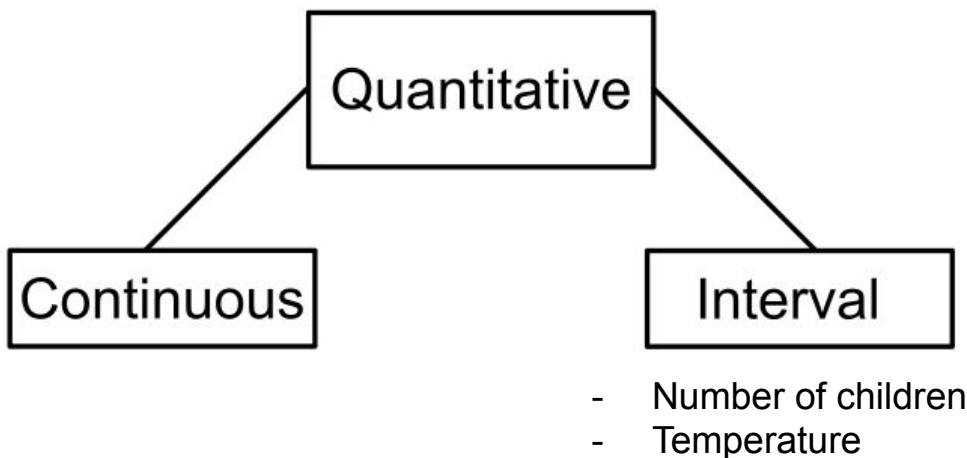
# Understand the data: types of variables

Do variables take numerical values for which arithmetic operations such as adding and averaging make sense? Then we have **quantitative** data.



# Understand the data: types of variables

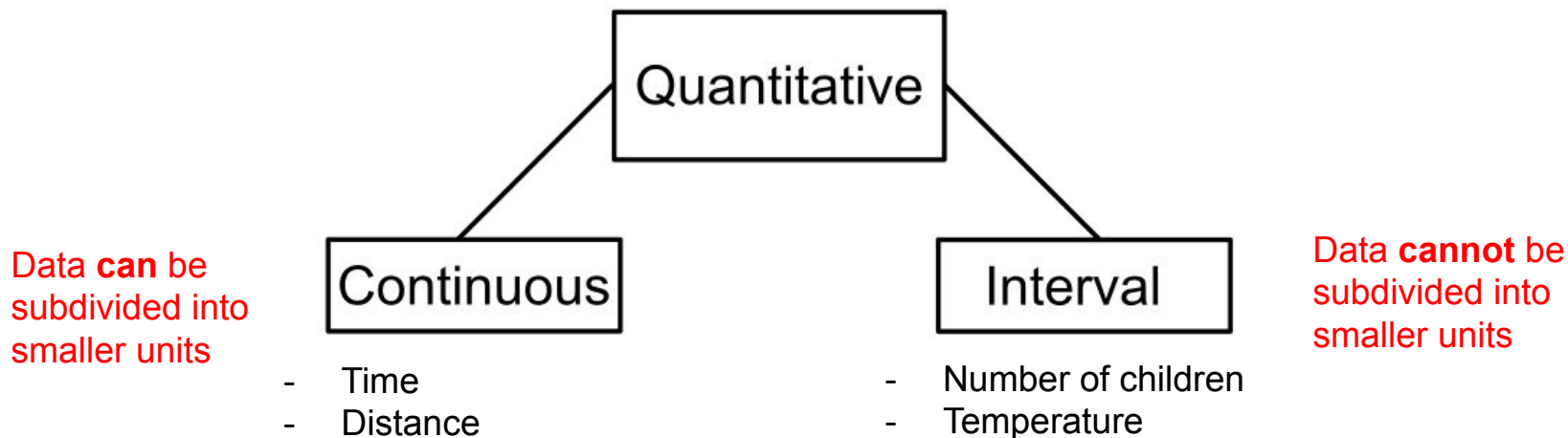
Do variables take numerical values for which arithmetic operations such as adding and averaging make sense? Then we have **quantitative** data.



Data **cannot** be subdivided into smaller units.

# Understand the data: types of variables

Do variables take numerical values for which arithmetic operations such as adding and averaging make sense? Then we have **quantitative** data.



# Let's practice all together!

In a company, what type of data is the following?

- Department?
- Number of years with the company?
- Salary?
- Education (coded as high school, some college, or college degree)?

# Understand the data: describe them!

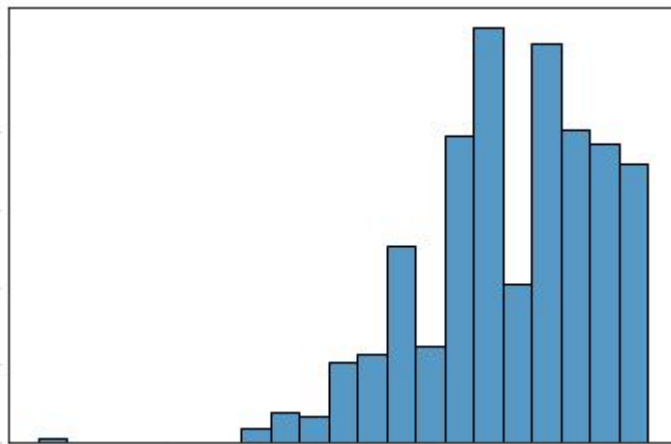
- We usually start with visualizing our data to understand them and get a flavour about our research question.

# Understand the data: describe them!

- We usually start with visualizing our data to understand them and get a flavour about our research question.

e.g. **Histograms** to visualize one **continuous** variable.

In ggplot: `geom_histogram()`



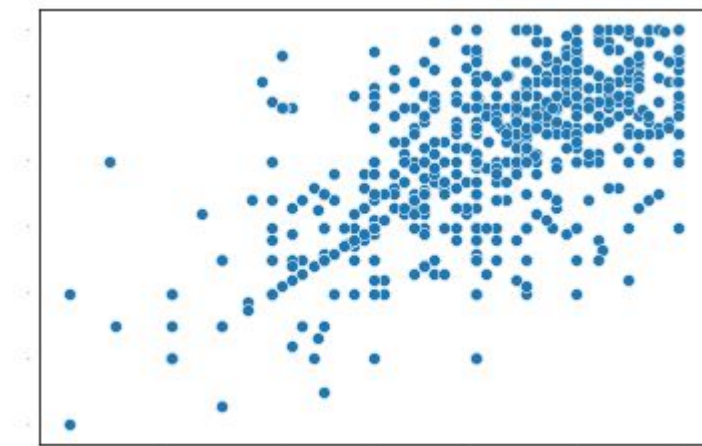


# Understand the data: describe them!

- We usually start with visualizing our data to understand them and get a flavour about our research question.

e.g. **Scatterplots** to  
visualize one  
**continuous** variable  
against another  
**continuous** variable

In ggplot: `geom_point()`





# Understand the data: describe them!

- We usually start with visualizing our data to understand them and get a flavour about our research question.
- However, we need to make sense of what we see. This is done by summarizing the data in a compact, easily-understood way. This is usually called **descriptive statistics** (or exploratory data analysis, EDA).

# Understand the data: describe them!

- We usually start with visualizing our data to understand them and get a flavour about our research question.
- However, we need to make sense of what we see. This is done by summarizing the data in a compact, easily-understood way. This is usually called **descriptive statistics** (or exploratory data analysis, EDA).

In contrast, **inferential statistics** is a follow-up step, aiming at drawing conclusions about what we do not know.

# Understand the data: describe them!

- We usually start with visualizing our data to understand them and get a flavour about our research question.
- However, we need to make sense of what we see. This is done by summarizing the data in a compact, easily-understood way. This is usually called **descriptive statistics** (or exploratory data analysis, EDA).
- When summarizing the data, we generally talk about **tendency, variability and shape**.

# Central Tendency

- In most situations, the first thing that you'll want to calculate is a measure of central tendency. That is, you'd like to know something about the “average” or the “middle” of where your data lies.
- The most commonly used measures are the **mean**, **median** and **mode**; occasionally people will also report a trimmed mean.

# A measure of central tendency: The mean

- For a sample  $X$  consisting of  $N$  observed values  $(X_1, X_2, \dots, X_N)$ , the **sample mean**, denoted as  $\langle X \rangle$ , is calculated as

$$\langle X \rangle = \frac{1}{N}(X_1 + X_2 + \dots + X_N) = \frac{1}{N} \sum_{i=1}^N X_i$$

- The **population mean** is also computed the same way but is denoted as  $\mu$ . It is often not possible to calculate  $\mu$  since population data are rarely available.
- The sample mean is a **sample statistic**, i.e. a quantity computed from the sample. It serves as an estimate of the population mean, which, yet not perfect, is usually a pretty good estimate (unless the sample is not a good representative of the population).

# A measure of central tendency: The mean

- For a sample  $X$  consisting of  $N$  observed values  $(X_1, X_2, \dots, X_N)$ , the **sample mean**, denoted as  $\langle X \rangle$ , is calculated as

$$\langle X \rangle = \frac{1}{N} (X_1 + X_2 + \dots + X_N) = \frac{1}{N} \sum_{i=1}^N X_i$$

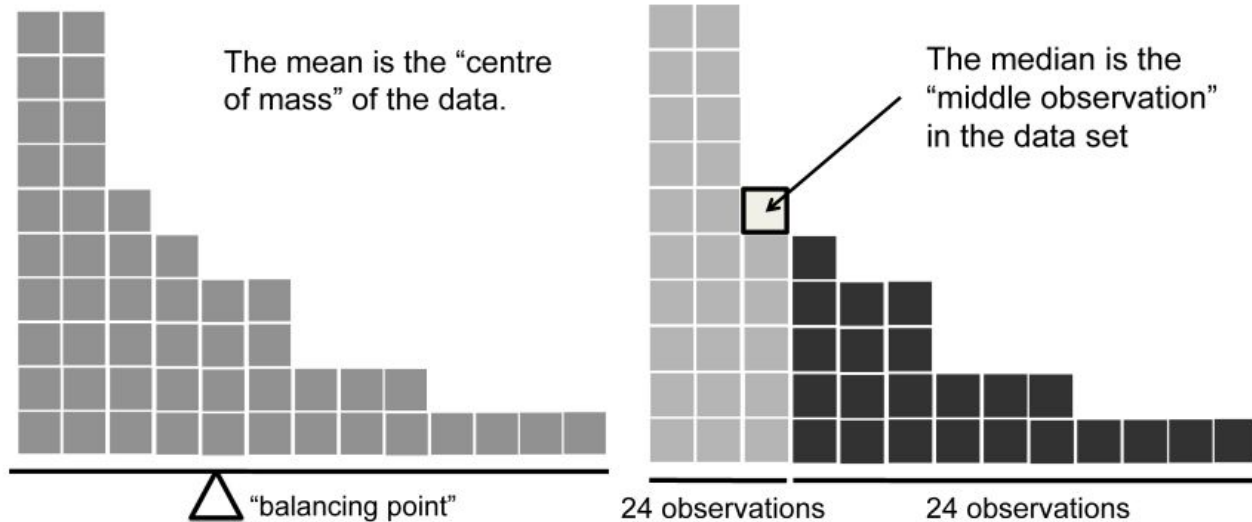
- The **population mean** is also computed the same way. It is often not possible to calculate the population mean  $\mu$ . It is often not possible to calculate the population mean  $\mu$ .
- The **sample mean** is a good estimate of the population mean, which, yet not perfect (unless the sample is not a good representation of the population).

**We'll come back in week 4 to a more thorough explanation about the difference between sample and population!**

# A measure of central tendency: The median

- Another widely used measure of tendency is the median.
- The median of a sample of observations is just the **middle value**. For example:
  - In 1, 2, **3**, 4, 5, the median is 3
  - In 1, 2, **3, 4**, 4, 5, the median is  $(3+4) / 2 = 3.5$

# Mean vs Median





# Other central tendency measures

- Trimmed mean:
  - It is the mean calculated by dropping extreme values (on both ends).
  - It is useful when **outliers** are presented in the data.
  
- Mode:
  - It is the most frequent value.
  - It is particularly important with **categorical** data.

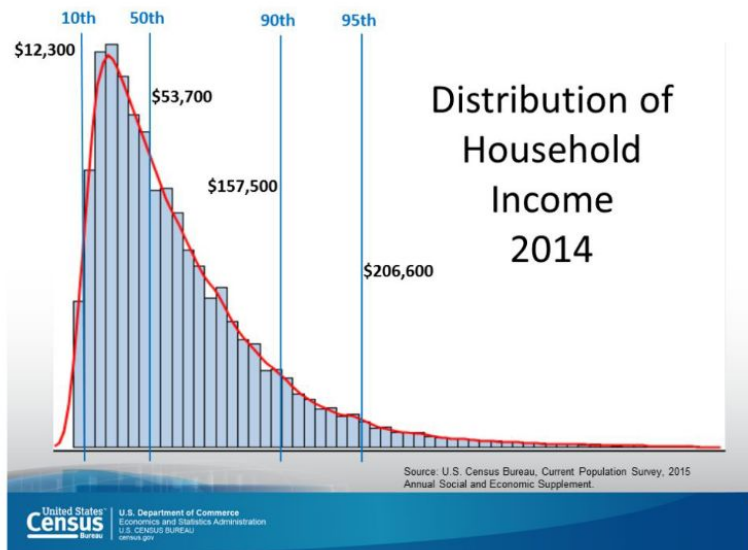
# Which one to use?

It depends a little on what type of data you've got and what you're trying to achieve. As a rough guide:

- **Nominal categorical** data: The mode.
- **Ordinal categorical** data: The median.
- **Quantitative (discrete and continuous)** data: The mean and/or the median.

# Which one to use? US Household income example

To estimate the typical household income in the US, would you be more interested in the mean or median income?

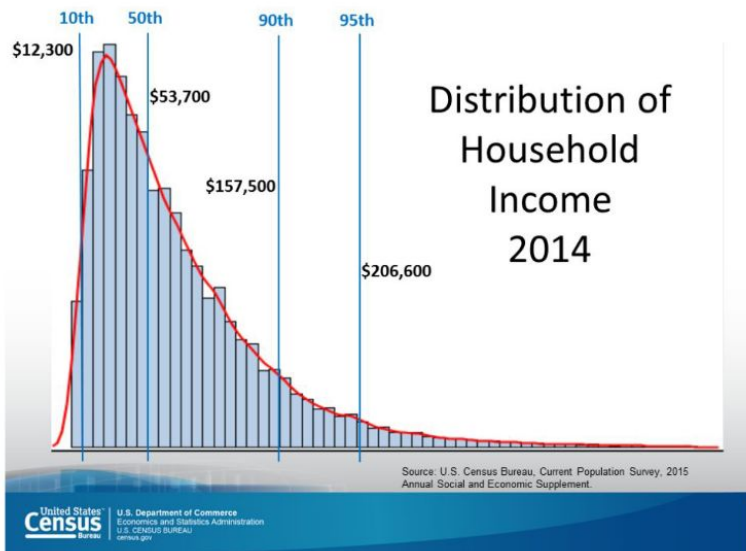


Median: \$53,700

Mean: \$75,738

# Which one to use? US Household income example

To estimate the typical household income in the US, would you be more interested in the mean or median income?



Median: \$53,700

Mean: \$75,738

Sometimes, the mean is not the best tendency measure (e.g. when distributions are **asymmetrical**, i.e. **always** visualize your data to decide!!!!)

# Variability of the data

When we talk about variability in the data, we mean to address the following questions:

- That is, how “**spread out**” are the data?
- How “**far**” away from the **mean or median** do the observed values tend to be?

# Variability of the data

It **complements** the information provided by the central tendency:

	# people at Sally's book club	# people at Maria's book club
Week 1	8	1
Week 2	10	18
Week 3	11	10
Week 4	9	2
Week 5	12	19
<b>Mean</b>	$= \frac{8 + 10 + 11 + 9 + 12}{5} = 10$	$= \frac{1 + 18 + 10 + 2 + 19}{5} = 10$
<b><i>Standard Deviation</i></b>	$= \sqrt{\frac{(8-10)^2 + (10-10)^2 + (11-10)^2 + (9-10)^2 + (12-10)^2}{4}} \approx 1.6$	$= \sqrt{\frac{(1-10)^2 + (18-10)^2 + (10-10)^2 + (2-10)^2 + (19-10)^2}{4}} \approx 8.5$

# A measure of variability: Variance

The **variance** is sometimes referred to as the “mean square deviation”; that is, the average of all points with respect to the mean sample in square units:

$$s^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \langle X \rangle)^2$$

# A measure of variability: Standard Deviation

- Sometimes also called the “root mean squared deviation” or RMSD, it is the **square root of the variance**:

$$s = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \langle X \rangle)^2}$$

- As a result, in contrast to the variance, the standard deviation is in **the same units** as the data themselves.
- (One) **interpretation**: 68% of the data fall within 1 standard deviation of the mean, 95% within 2 standard deviation of the mean, and 99.7% within 3 standard deviations of the mean. (Only for Gaussian distributions!)



# Other variability measures

- Average absolute deviation (AAD):

$$AAD = \frac{1}{N} \sum_{i=1}^N |X_i - \langle X \rangle|$$







- Median absolute deviation (MAD):

$$MAD = \text{median}(|X_1 - \langle X \rangle|, |X_2 - \langle X \rangle|, \dots, |X_N - \langle X \rangle|)$$

- Range: Biggest value minus the smallest value.
- Interquartile range: Difference between the **25th quantile** and the **75th quantile**. (What is a quantile? We'll see this later)

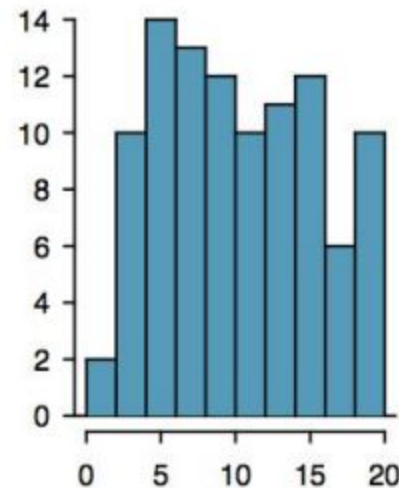
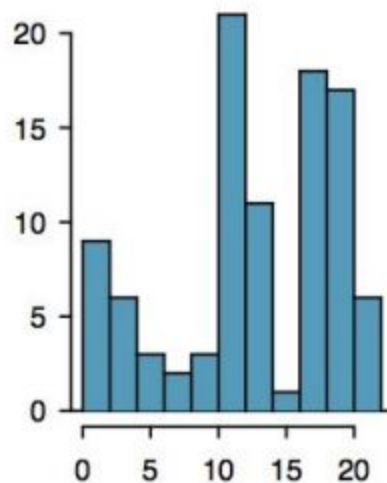
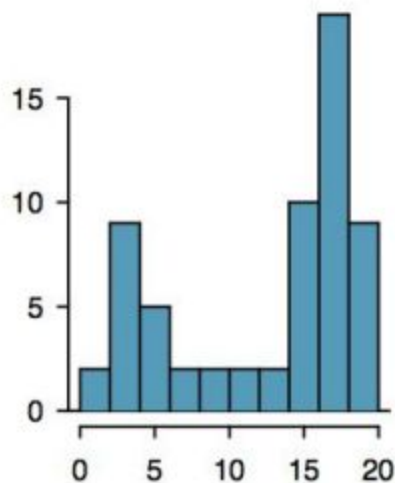
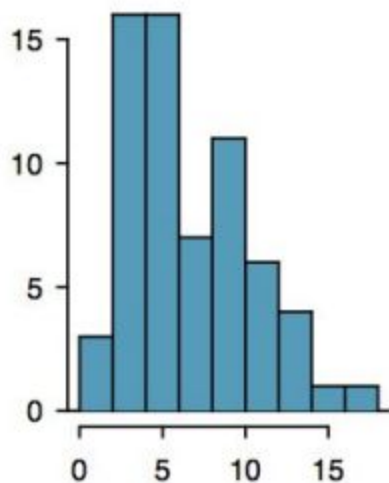
# Which one to use?

(Personal) ranking of use:

- Standard Deviation    
- IQR   
- AAD, MAD  
- Variance 
- Range 

# Shape of the data: Modality

Do our data exhibit a single prominent peak (unimodal), several prominent peaks (bimodal/multimodal), or no apparent peaks (uniform)? ✓✓✓✓



# Shape of the data: Skewness

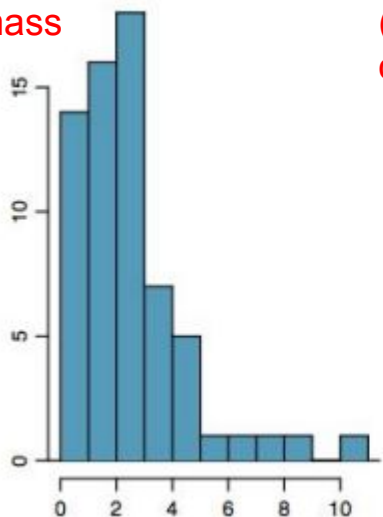
Skewness is a measure of **asymmetry** of the data

$$skewness(X) = \frac{1}{N\hat{\sigma}^3} \sum_{i=1}^N (X - \langle X_i \rangle)^3$$



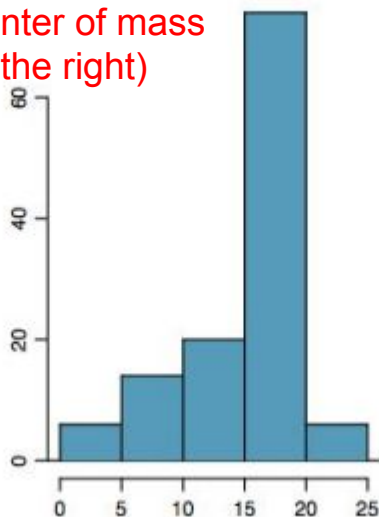
Positive

(center of mass  
on the left)

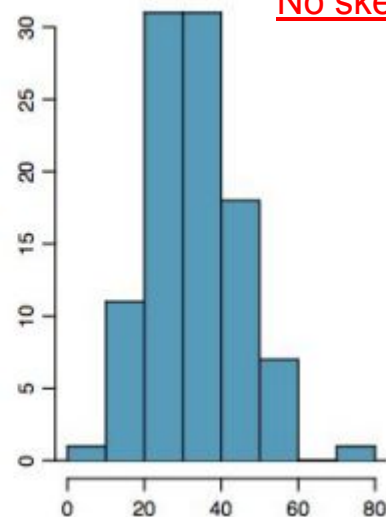


Negative

(center of mass  
on the right)



No skew

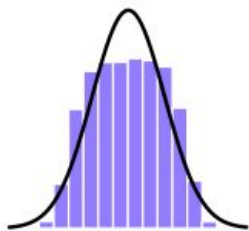


# Shape of the data: Kurtosis

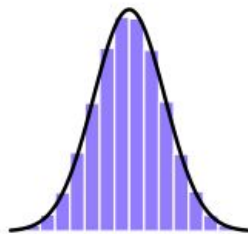
Kurtosis is a measure of the  
“**pointiness**” of a the data ✓✓

$$kurtosis(X) = \frac{1}{N\hat{\sigma}^4} \sum_{i=1}^N (X - \langle X_i \rangle)^4 - 3$$

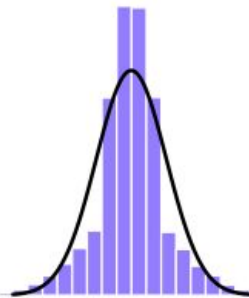
Platykurtic  
("too flat")



Mesokurtic

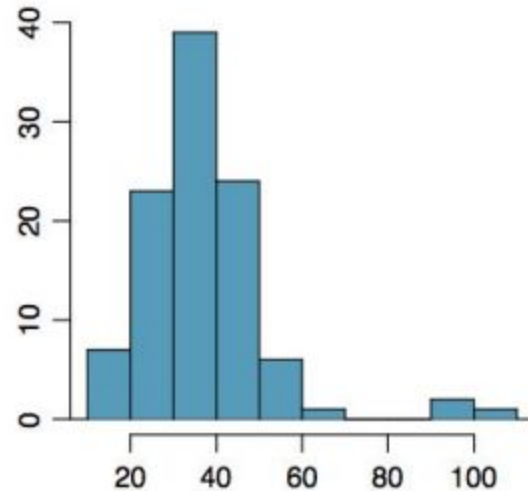
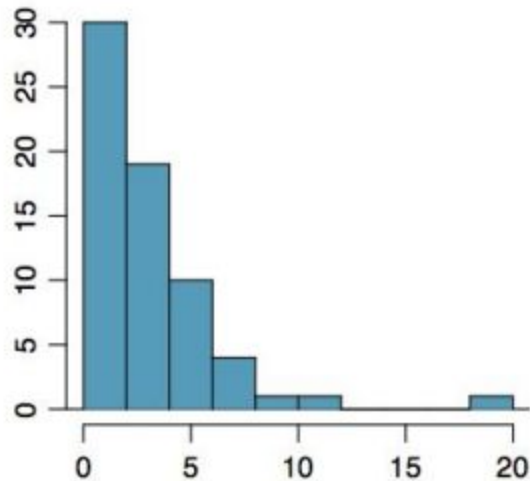


Leptokurtic  
("too pointy")



# Shape of the data: Outliers

Do we suspect to have unusual observations or potential outliers in our data?



# Recap

- **Always start by understanding your data!**
- One way of doing this is by **visualizing** (e.g. using R ggplot) and **summarizing** (e.g. using R dplyr) them.
- Descriptive statistics compress data to make it easier to understand and communicate (crucial in research!!!!)
- We generally want to talk about measures of tendency, variability and shape, but be **aware** of your data properties before selecting a specific measure!