Fundamentals of Working with Data

Monia Ranalli
monia.ranalli@uniroma2.it

Objectives

- give an answer to: **What is Statistics?**
- strategies for **collecting data**
- organize data by tables and use appropriate **graphical techniques** to describe various data sets
- identify **variables as categorical** (binary, ordinal, nominal) or **quantitative** (discrete, continuous)
- conceptualize **statistical inference**
- use appropriate **summary measures** to describe various data sets
- construct and use **box plots**
- explore **association** between two quantitative variables

Why Study Statistics?

- ...the annual report of a company printed that the sales next year are expected to be 11.50 million €with a standard deviation of 1.2 million €. $\rightarrow$ **To evaluate numerical facts**
- ...it has been asked to project the sales of a company for next year. $\rightarrow$ **To perform statistical data analysis or to interpret the results of sampling**
- ...a survey on the drinking habits of Italians estimated the percentage of adults across the country who drink beer, wine, or hard liquor, at least occasionally. Of the 1516 adults interviewed, 985 said that they drank at least occasionally. What can we say about the proportion of Italians that drink at least occasionally? Due to various constraints, for example, time or budget, one can only sample from the population instead of take a census of the population. We need a sample that closely represents the population. One way is to obtain a random sample. $\rightarrow$ **To make inference about the population through the sample**

What Do Statisticians Do?

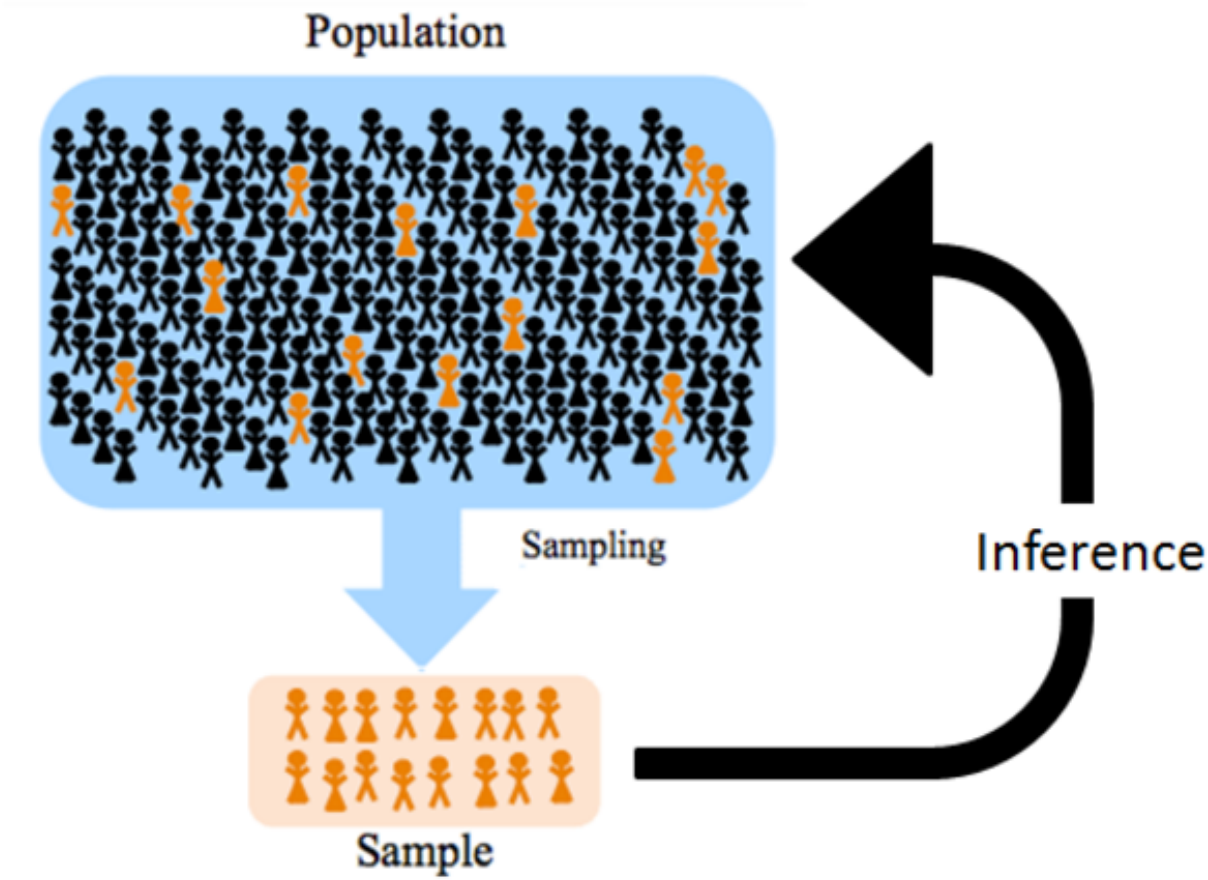
- **Gather data** $\rightarrow$ Draw a random sample of students, for example. The sample size depends on how accurate you need your inference to be and the margin of error you can tolerate.
- **Summarize data** $\rightarrow$ Summarize data from the sample (e.g. sample mean and sample standard deviation).
- **Analyze data** $\rightarrow$ Analyze the data through statistical techniques and make inference through confidence interval and hypothesis testing.
- **Draw conclusions and report the results of their analysis** $\rightarrow$ Write reports and support your conclusions including plots, tabular and numerical displays.

Collecting Data

- **Non-probability Methods**
  - **Convenience sampling:** for example, surveying passengers as they approach the ticket counter at the train station.
  - **Volunteer sampling:** for example, distribute a questionnaire to each first year-student requesting they complete the survey and return it at end of the semester.
- **Probability Methods**. [**Benefit**: making inference]
  - **Simple random sample:** select a group of subjects from a population where each subject in the population has an equal chance of being selected.
  - **Stratified random sample:** identify the population of interest, divide it into strata or groups based on some characteristic (e.g. age, sex, level of education), then perform simple random sample from each strata.
  - **Cluster sample:** take a random cluster of subjects from the population of interest. For example, stratify passengers of Trenitalia Frecciarossa trains by class they travel (Standard, Premium, Business and Executive) and randomly select such classes from various trains and survey each passenger in that class and train selected.

Sample & Population

- **Population:** the entire set of possible observations in which we are interested
- **Sample:** a subset of the population from which information is actually collected



Sample Statistics & Population Parameters

- A **parameter** is a numerical summary of the population → proportion of all teenagers in the United States who have smoked in the last month
- A **statistic** is a numerical summary of a sample taken from the population → proportion of teenagers who have smoked in the last month out of a sample of 200 randomly selected teenagers in the United States

Statistics are used to make inference about population parameters

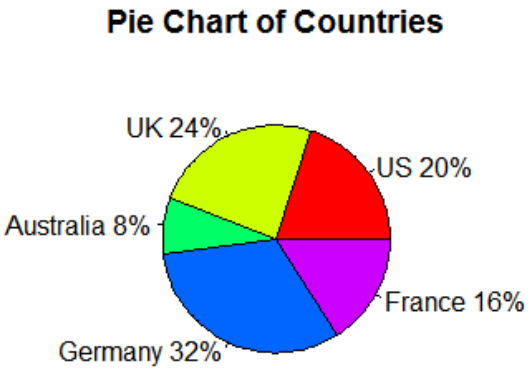
$$\Downarrow$$

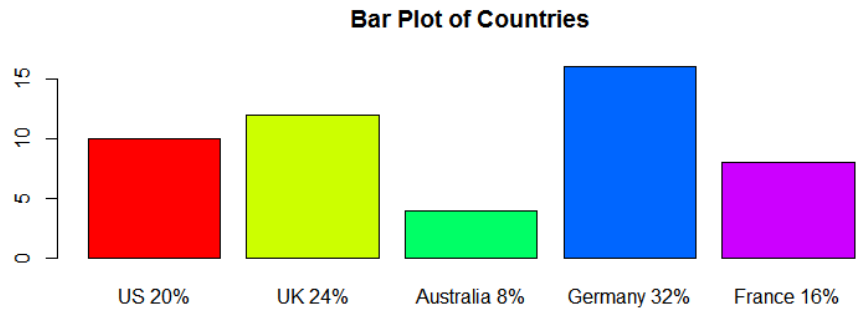
- **Descriptive statistics** → Techniques of describing data in ways to capture the essence of the information
- **Inferential statistics** → to draw conclusions from data about the population

Types of Variables

- **Qualitative (Categorical) variable** each observation belongs to one of a set of categories order between them. **Categorical values** may be:
  - **binary** where there are two choices, e.g. Gender (Male - Female);
  - **ordinal** where the categories imply levels with hierarchy or order of preference, e.g. Education (Primary, High school, College)
  - **nominal** where no hierarchy is implied, e.g. Religious Affiliation (Catholic, Jewish,...).
- **Quantitative variable:** observations on it take numerical values that represent different magnitudes of the variable. Quantitative values can be:
  - **discrete** if they are possible values form a set of separate numbers, such as 0,1,2,3,... The set of possible values is not dense. E.g., Number of children in a family or Number of foreign languages spoken by an individual;
  - **continuous** if they are possible values from an interval. The set of possible values is dense. E.g., Height/Weight, Age or Blood pressure

Single Categorical Variable - Pie chart and Bar Plot





Pie chart: area of the pie represents the percentage of that category

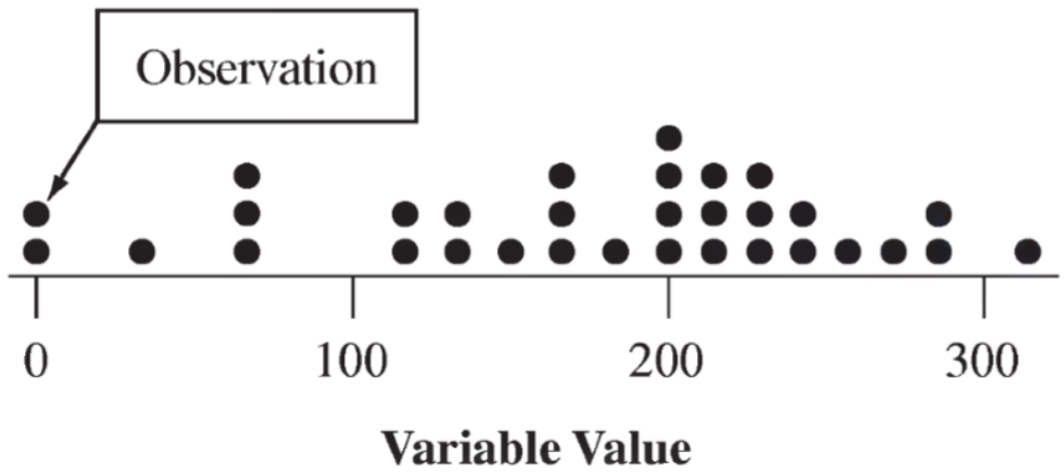
- It is easier to compare one category with the whole
- It may not be suitable for too many categories

Bar chart: the height of the bar for each category is equal to the frequency (number of observations) in the category.

- It is easier to compare categories

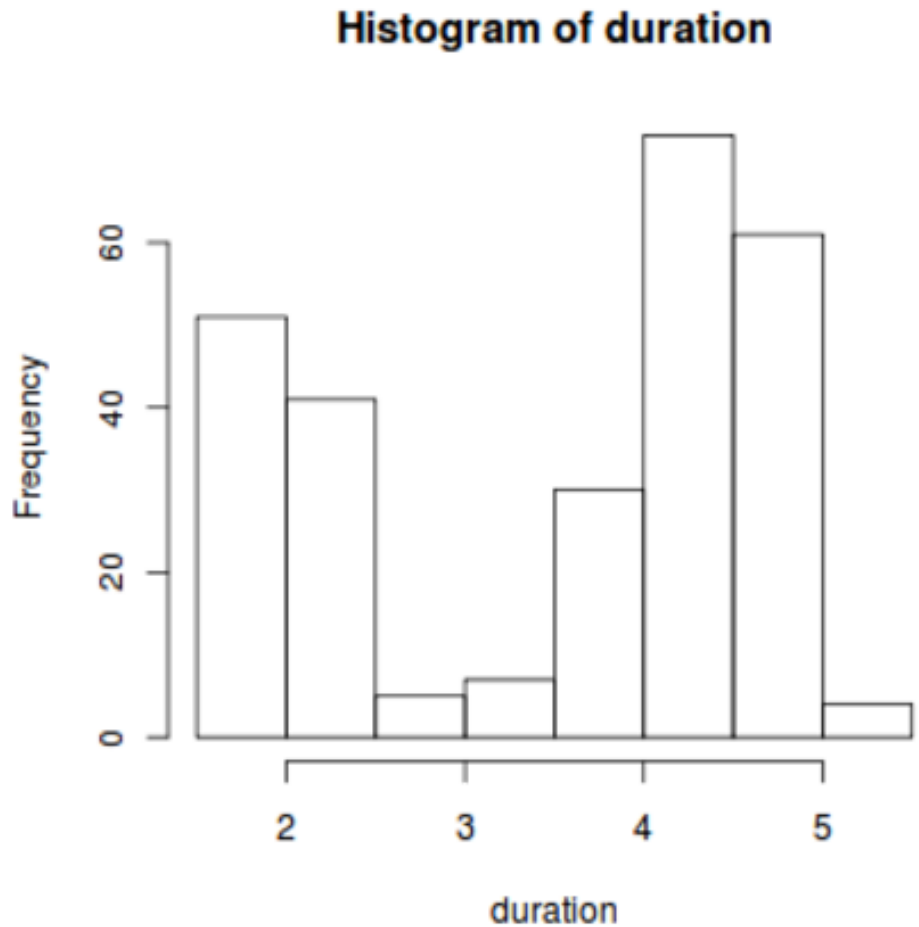
Single Quantitative Variable (Discrete)

- **Dotplot:** useful to show the relative positions of the data.



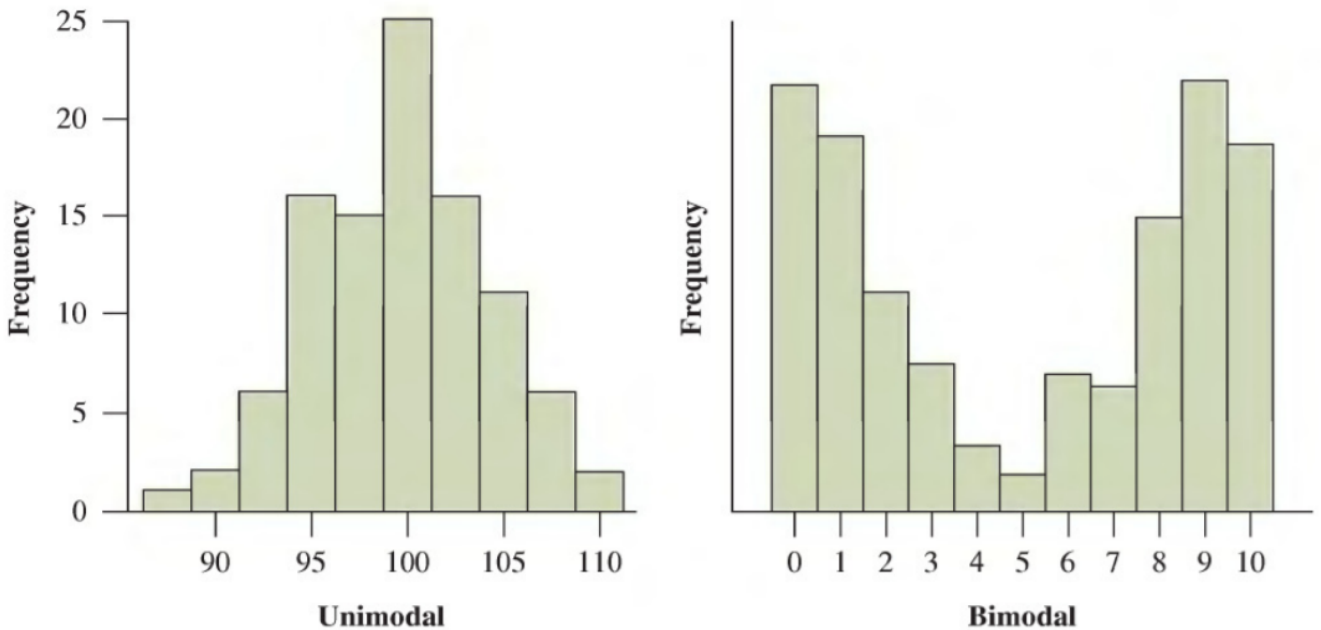
Single Quantitative Variable (Continuous)

- **Histogram:** it displays the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.



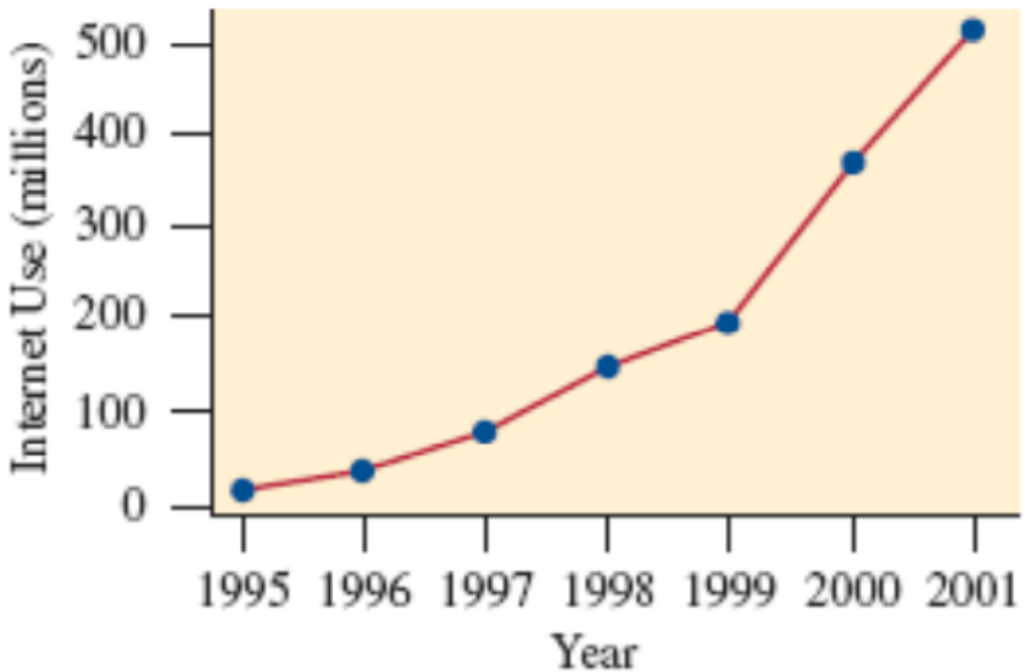
Shape of a distribution

- **Unimodality:** A distribution of data with a single mound is called **unimodal**. The highest point is at the **mode**.
- **Bimodality:** A distribution with two distinct mounds is called **bimodal**.



Single Time Series

- **Time Plot:** Used for displaying a time series, a data set collected over time. Points are usually connected. Common patterns in the data over time, known as **trends**, should be noted.



Measures of Central Tendency

- **Mean:** the average of the data. (**Note**: it can be computed only for quantitative variables!)

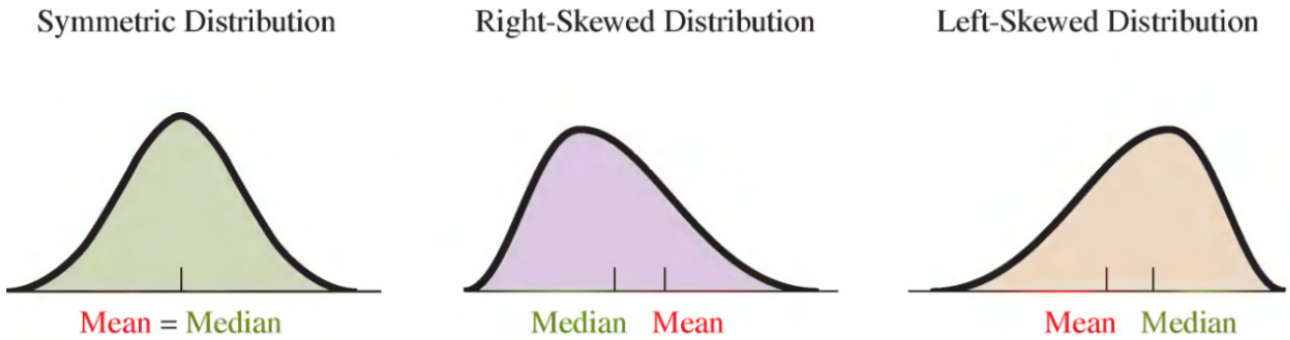
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_n}{n}$$

If $y_i = a + bx_i$ with $i = 1, \ldots, n$, then $\bar{y} = a + b\bar{x}$

- **Median:** the middle value of the ordered data.
  Steps to finding the median for a set of data:
  1. Order the observations
  2. Find the location of median in the ordered data by $(n+1)/2$
  3. The value that represents the location found in Step 2 is the median. **NOTE**: if the sample size is an odd number then the median is the middle observation. If sample size is an even number, then the median is the average of the two middle observations . The result may or may not be an observed value.
- **Mode:** the value that occurs most often in the data.

Some remarks

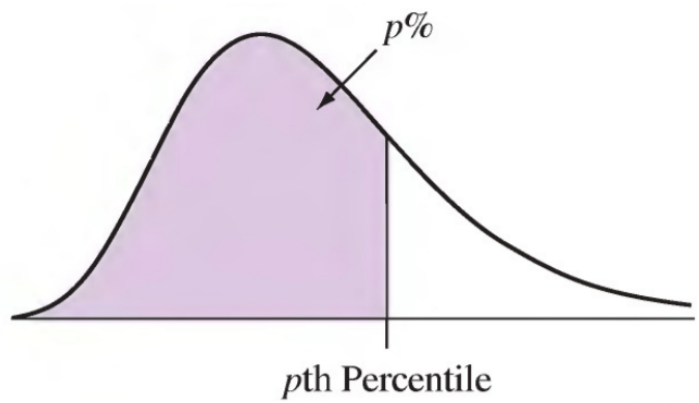
- Mean, median and mode are usually not equal



- If the data distribution is symmetric, then the mean is equal to the median
- Mean is affected by extreme values
  Given this data set, $95, 78, 69, 91, 82, 76, 76, 86, 88, 80$, the mean is $\bar{x} = (95 + 78 + 69 + 91 + 82 + 76 + 76 + 86 + 88 + 80)/10 = \mathbf{82.1}$. If the entry **69** is mistakenly recorded as **9**, the mean would be **76.1**, which is very different from 82.1.
- Median is resistant, i.e. it is not affected by extreme values
  Ordered original data set is: $69, 76, 76, 78, 80, 82, 86, 88, 91, 95$. With $n = 10$, the median is the average of the fifth (80) and sixth (82) ordered value and the median is **81**. Ordered new data set (with 69 coded as 9) is: $9, 76, 76, 78, 80, 82, 86, 88, 95$ where the median is still **81**.

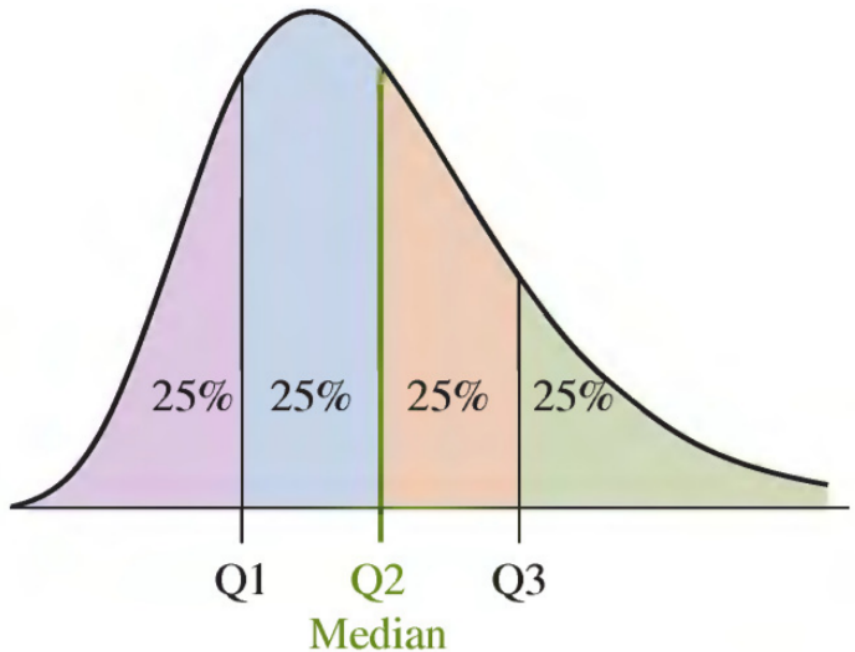
Measures of Variability

- **Range**: Maximum-Minimum
  Easy to calculate, but very much affected by extreme values (range is not a resistant measure of variability)
- **Interquartile range (IQR)**: we need to first talk about percentiles.
  The ***p*th percentile** is a value such that such that after the data are ordered from smallest to largest, at most, p% of the observations fall below or at that value, and at most, (100 - p)% above or at it.



Measures of Variability

- **Interquartile range (IQR)**:
  IQR$=Q_3 - Q_1$ = upper quartile - lower quartile
  = 75th percentile - 25th percentile
  IQR is not affected by extreme values. It is thus a resistant measure of variability



Measures of Variability

- **Population Variance**: it is the average squared distance from the mean. (μ is the population mean)

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

If $y_i = a + bx_i$ with $i = 1, \ldots, n$, then $\sigma_Y^2 = b^2\sigma_X^2$

- **Population Standard Deviation**: it can be found by $\sigma = \sqrt{\sigma^2}$; it has the same unit as x_i's $\rightarrow$ spread in terms of the original unit.
  If $y_i = a + bx_i$ with $i = 1, \ldots, n$, then $\sigma_Y = \mid b \mid \sigma_X$
- **Sample Variance and Sample Standard Deviation**: When the data set is a sample,

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

Why do we divide by $n-1$ instead of by n? For now...Since μ is unknown and estimated by $\bar{x}$, the x_i's tend to be closer to $\bar{x}$ than to μ. To compensate, we divide by a smaller number, $n-1$. It will be explained better later in the course!

Example

By inserting a quarter into two candy vending machines A and B, the number of pieces dropped is random.

Same center, but what about their spreads?

- **Machine A**: $1, 2, 3, 3, 5, 4 \Rightarrow \bar{x} = 3,\ Q_2 = 3,\ \text{mode} = 3,$
  $s^2 = \dfrac{(1-3)^2 + (2-3)^2 + 2 \times (3-3)^2 + (5-3)^2 + (4-3)^2}{6-1} = \mathbf{2},$
  $s = \sqrt{2} = \mathbf{1.414}$
- **Machine B**: $2, 3, 3, 3, 3, 4 \Rightarrow \bar{x} = 3,\ Q_2 = 3,\ \text{mode} = 3,$
  $s^2 = \dfrac{(2-3)^2 + 4 \times (3-3)^2 + (4-3)^2}{6-1} = \mathbf{0.4},\ s = \sqrt{0.4} = \mathbf{0.6325}$

Z-score

- Z-value or Z-score or simply Z, represents **the number of standard deviations an observation is from the mean**. The z-score for a particular observation is calculated as

$$z_i = \frac{x_i - \bar{x}}{s}.$$

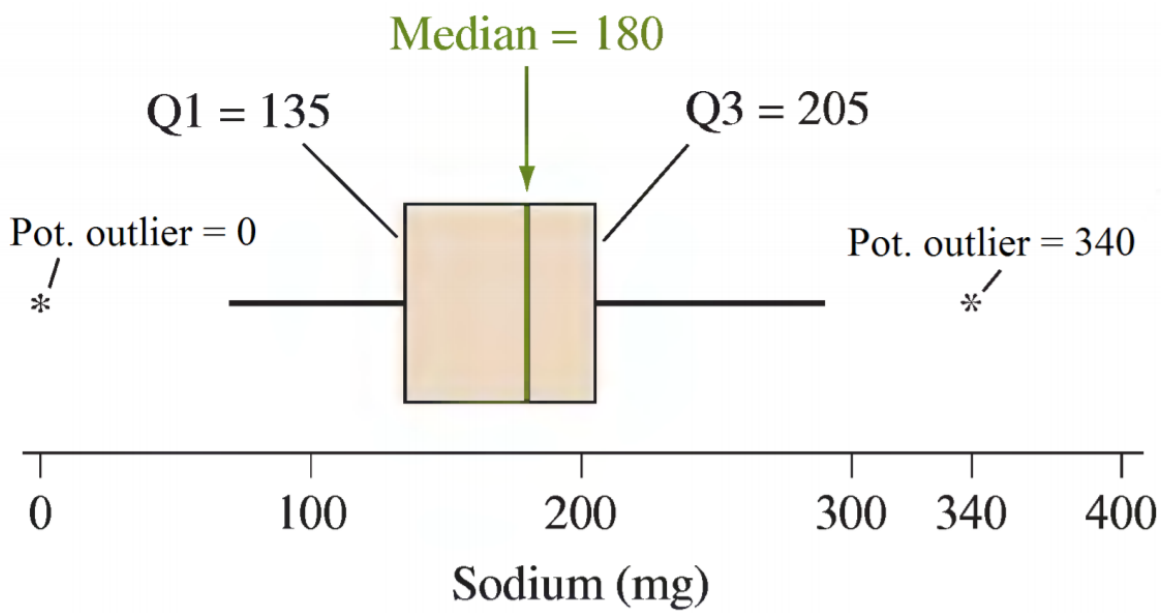
- The z-scores have **mean 0 and standard deviation 1**.
- **Positive z-score** $\rightarrow$ the observation is above the mean.
- **Negative z-score** $\rightarrow$ the observation is below the mean.
- An observation from a bell-shaped (or nearly symmetric) distribution is a **potential outlier if its z-score is beyond** ± 3.

Example: For a recent final exam the mean was 68.55 with a standard deviation of 15.45. Student A scored an 80%:
$z_A = (80 - 68.55)/15.45 = \mathbf{0.74}$, which means the score of 80 was 0.74 standard deviation above the mean. Student B scored a 60%:
$z_B = (60 - 68.55)/15.45 = \mathbf{-0.55}$, which means the score of 60 was 0.55 standard deviation below the mean.

Box Plots

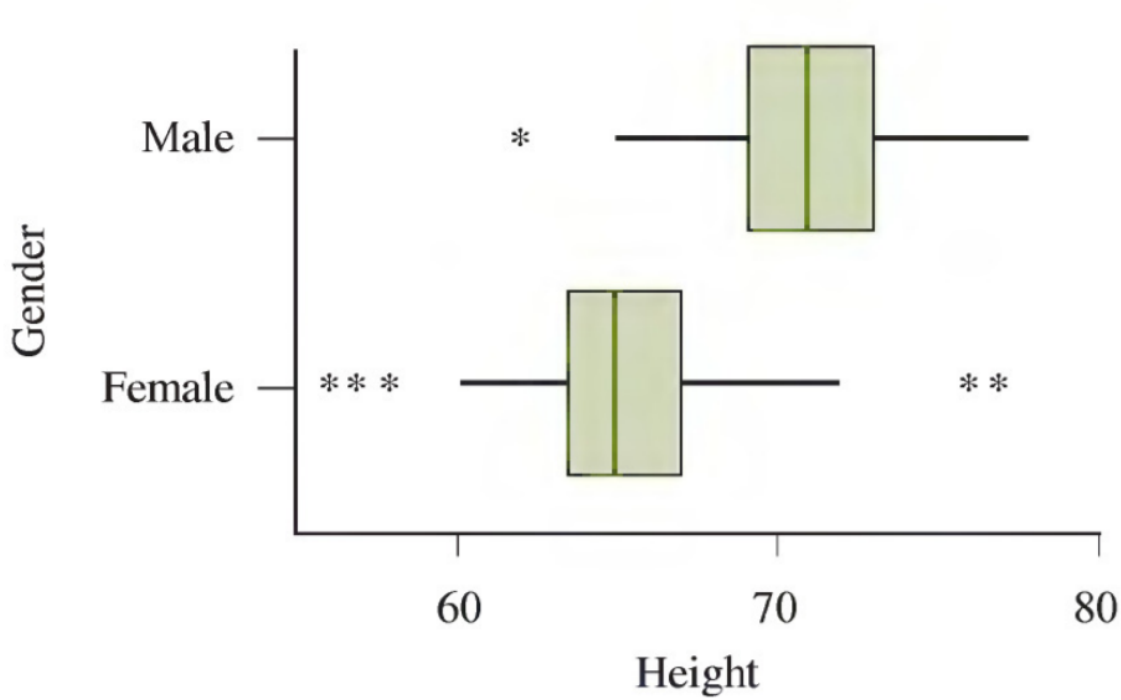
- It shows: the **skewness** of the distribution, the **central location** and the **variability**
- To create it we need the **five number summary**: minimum value, Q_1, Q_2, Q_3, and maximum value.
- It identifies **potential outliers**: Lower Limit$=Q_1 - 1.5 \times IQR$, Upper Limit$=Q_3 + 1.5 \times IQR$.

Example: Boxplot for Cereal Sodium Data

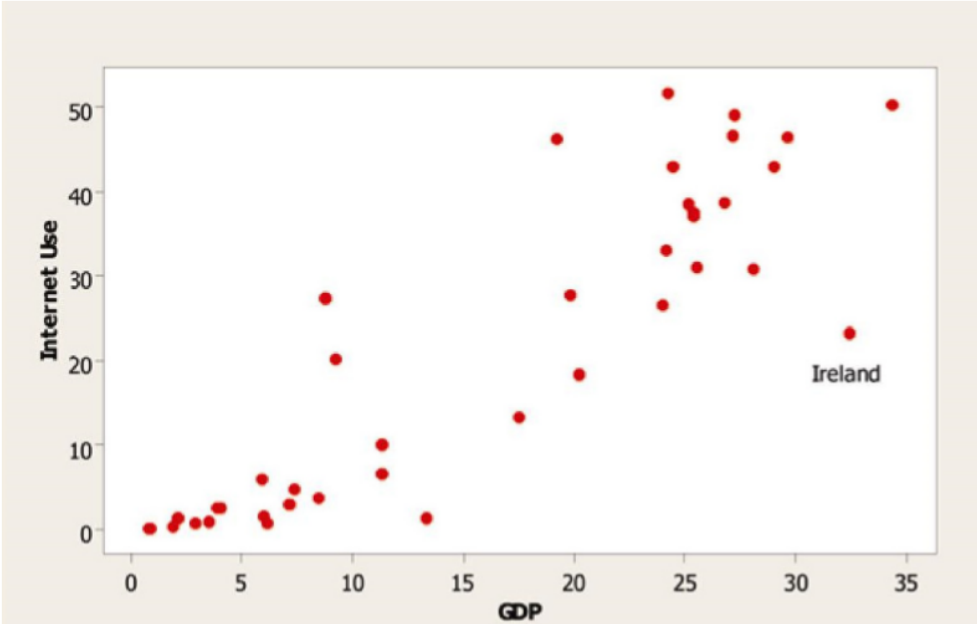


Side-by-Side Box Plots

- **Quantitative data** can be broken down **by levels of a categorical variable**
- The side-by-side boxplot produces an **excellent visual comparison** for **shape**, **outliers**, **variability**, etc.



Association between Two Quantitative Variables



Two quantitative variables x and y are

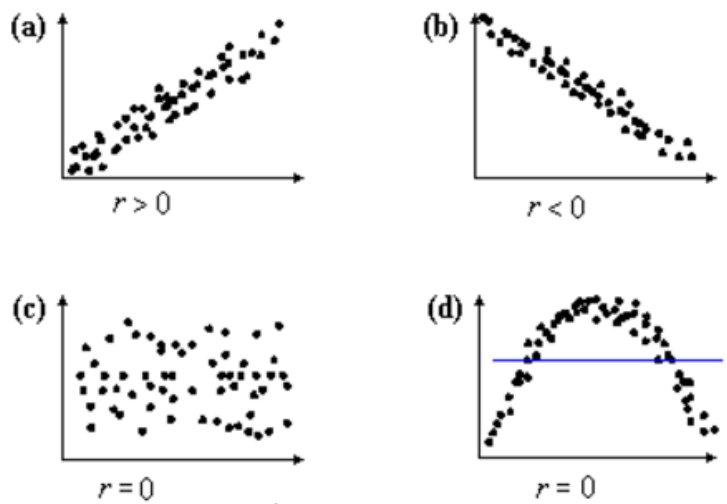
- **Positively associated**: high values of x tend to occur with high values of y; low values of x tend to occur with low values of y
- **Negatively associated**: high values of one variable tend to pair with low values of the other variable (high values of x tend to occur with low values of y; low values of x tend to occur with high values of y)

Correlation

- Correlation measures **the strength and direction of the linear association** between two quantitative variables

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- A positive r value indicates a positive association
- A negative r value indicates a negative association
- The closer r is to ± 1 the closer the data points fall to a straight line, and the stronger the linear association is.
- The closer r is to 0, the weaker the linear association is.



Correlation

The following table contains spending (X) and income (Y) amounts in thousands of euro for 5: families

	x_i	y_i	$sx_i = x_i - \bar{x}$	$sy_i = y_i - \bar{y}$	$sx_i \times sy_i$	sx_i^2	sy_i^2
	8	15	2.6	7.2	18.72	6.76	51.84
	5	5	-0.4	-2.8	1.12	0.16	7.84
	7	9	1.6	1.2	1.92	2.56	1.44
	1	3	-4.4	-4.8	21.12	19.36	23.04
	6	7	0.6	-0.8	-0.48	0.36	0.64
Totale	27	39	0	0	42.40	29.20	84.80

- $\bar{x} = \dfrac{27}{5} = 5.4$
- $\bar{y} = \dfrac{39}{5} = 7.8$
- $r = \dfrac{42.40}{\sqrt{29.20 \times 84.80}} = 0.8520$