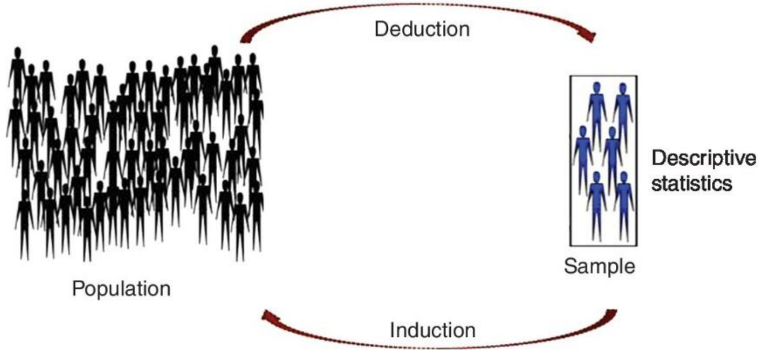
Chapter 2

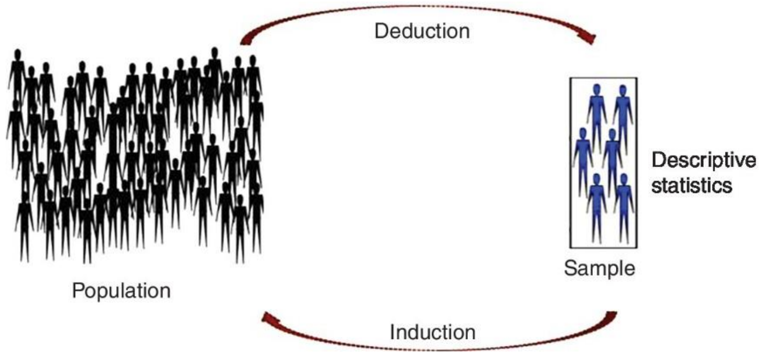
Descriptive statistics

Statistics



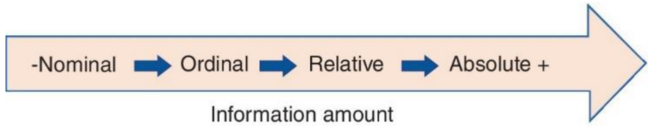
- **Population**
  - A set of similar instances/objects or events which is of interest for some question or experiment
  - *E.g. all students of my school, all nails produced by a machine*
- **Sample**
  - A set of a data collected and/or selected from a population by a defined procedure
  - *E.g. a subset of the students of my school that answered to a survey, a subset of randomly selected nails produced by a machine*

Statistics



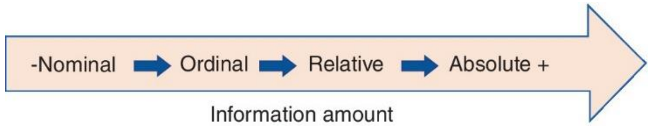
- Deduction
  - Reasoning about the sample extracted from that population
  - Probabilities in about deduction
- Induction
  - Concerns reasoning about the population given a sample
- Descriptive statistics
  - Descriptive statistics are methods / techniques to describe or summarize samples in order to help humans to understand it

Scale Types



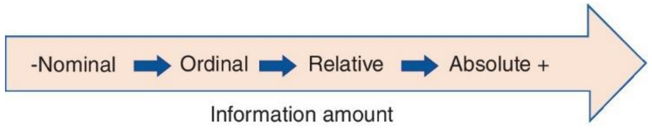
- **Qualitative** scales
  - **Nominal**: categorize data in a non-ordinal way
    - Operations: = and ≠
    - *E.g. friend's name and gender*
  - **Ordinal**: categorize data in a ordinal way
    - Operations: =, ≠, <, >, ≤, and ≥
    - *E.g. homestead, village, town, city, metropolis*
- **Quantitative** scales
  - **Relative**: does not have an absolute zero
    - Operations: =, ≠, <, >, ≤, ≥, - and +
    - *E.g. temperature*
  - **Absolute**: has an absolute zero
    - Operations: =, ≠, <, >, ≤, ≥, -, +, / and ×
    - *E.g. weight and height*

Scale Types Example



Friend	Max temp	Weight	Height	Gender	Company
Andrew	25	77	175	M	Good
Bernhard	31	110	195	M	Good
Carolina	15	70	172	F	Bad
Dennis	20	85	180	M	Good
Eve	10	65	168	F	Bad
Fred	12	75	173	M	Good
Gwyneth	16	75	180	F	Bad
Hayden	26	63	165	F	Excellent
Irene	15	55	158	F	Bad
James	21	66	163	M	Good
Kevin	30	95	190	M	Bad
Lea	13	72	172	F	Good
Marcus	8	83	185	F	Bad
Nigel	12	115	192	M	Good

Scale Types



- *Using as example the attribute weight expressed in an absolute scale in kg we can convert it in any other scale type:*
  - *Relative: by subtracting, for instance, to the values 10: the old zero is now -10 and the new zero is the old 10; and the new 80 kg is no more the double of the new 40 kg*
- *Ordinal: we can define, for instance, the levels of fatness: fat when the weight is larger than 80 kg; normal when the weight is larger than 65 kg but lower or equal than 80 kg; and thin when the weight is lower or equal than 65 kg*
- *Nominal: We can transform the previous classification of fat, normal and thin in B, A and C, respectively*

Scales vs data types

- In software packages we must choose the data type for each attribute
  - Common types are text, character, factor, integer, real, float, timestamp, date or several others
  - A scale and a data type are different concepts despite related
  - For instance, a quantitative scale implies the use of numeric data types
- However, an attribute can be expressed as a number but the scale type can be qualitative
  - Think about an identity card you have with a numeric code
    - what kind of quantitative information does it have?
  - A code with letters could contain the same information

Descriptive Univariate Analysis: frequencies

- A frequency is basically a counter
- **Absolute frequency** counts how many times a value appears.
- **Relative frequency** counts the percentage of times that value appears.
- The **absolute cumulative frequency** is the number of occurrences less or equal than a given value
- The **relative cumulative frequency** is the percentage of occurrences less or equal than a given value

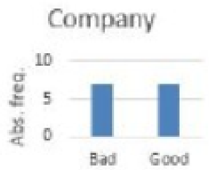
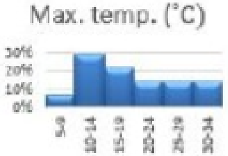
Descriptive Univariate Analysis: frequencies

Height	Abs. freq.	Rel. freq.	Abs. cum. freq.	Rel. cum. freq.
158	1	1/14=7.14%	1	1/14=7.14%
163	1	1/14=7.14%	2	2/14=14.29%
165	1	1/14=7.14%	3	3/14=21.43%
168	1	1/14=7.14%	4	4/14=28.57%
172	2	2/14=14.29%	6	6/14=42.86%
173	1	1/14=7.14%	7	7/14=50.00%
175	1	1/14=7.14%	8	8/14=57.14%
180	2	1/14=14.29%	10	10/14=71.43%
185	1	1/14=7.14%	11	11/14=78.57%
190	1	1/14=7.14%	12	12/14=85.70%
192	1	1/14=7.14%	13	13/14=92.86%
195	1	1/14=7.14%	14	14/14=100.00%

Descriptive Univariate Analysis: frequencies

- The relative frequencies define distribution functions, i.e., they describe how data are distributed
- Distribution functions are said empirical when they are obtained from a sample
- A discrete attribute, like one of the integer data type, has a probability mass function
- A continuous attribute, like one of the real data type, has a density probability function

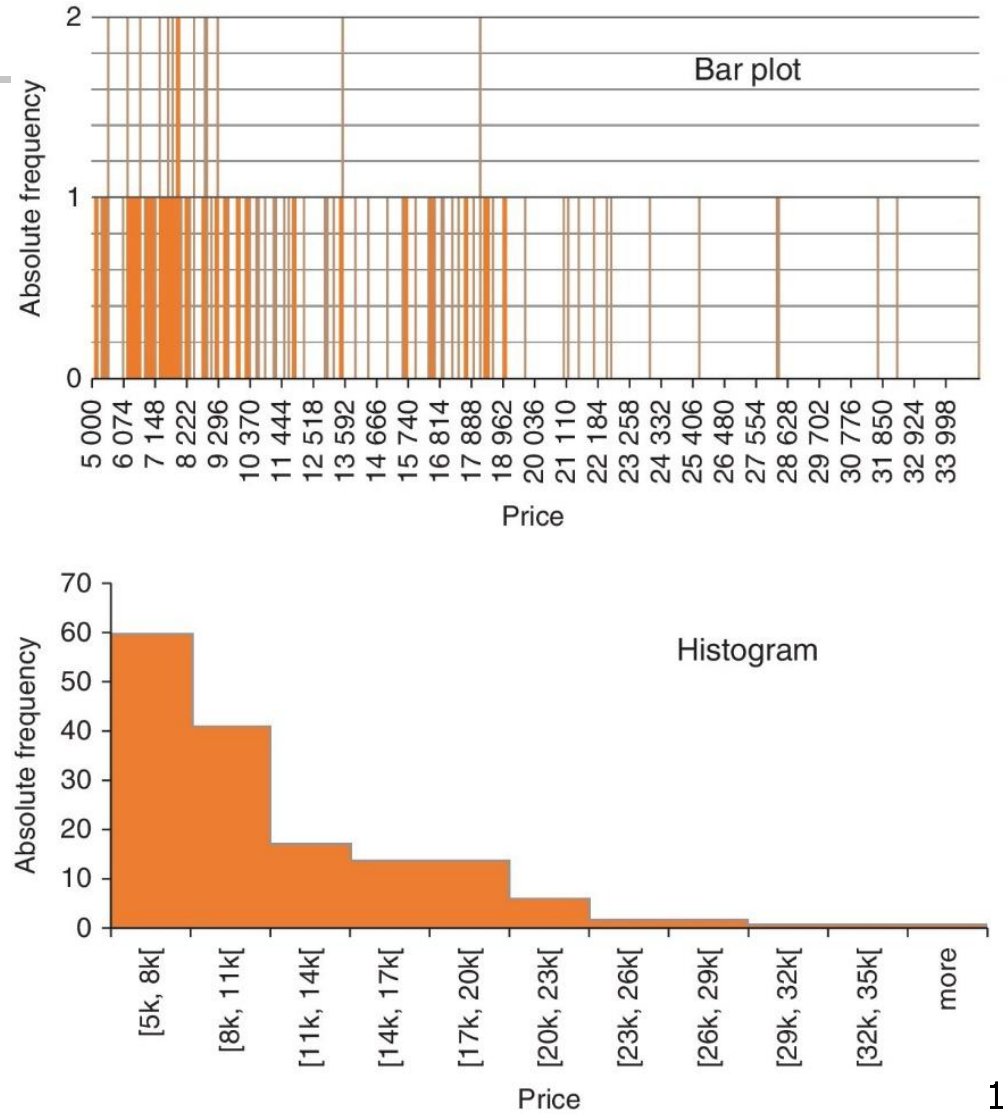
Descriptive Univariate Analysis: data visualization

Plot	Qualitative	Quantitative	Observation	Plot draft
Pie	Yes	No	Company relative frequency	
Bar	Yes	Not always	Company absolute frequency	
Line	No	Yes	Andrew's 5-day max. temperatures	
Area	No	Yes	Andrew & Eve 5-day max. temperatures	
Histogram	No	Yes	Max. last day temperatures of the 14 friends	

- Pie chart: it is used typically for nominal scales
- Bar chart: It is used typically for qualitative scales or quantitative scales with a limited number of values
- Line chart: they are specially used to deal with the notion of time
- Area charts: are specially used to compare time series and distribution functions
- Histograms: are used to represent empirical distributions for attributes with a quantitative scale

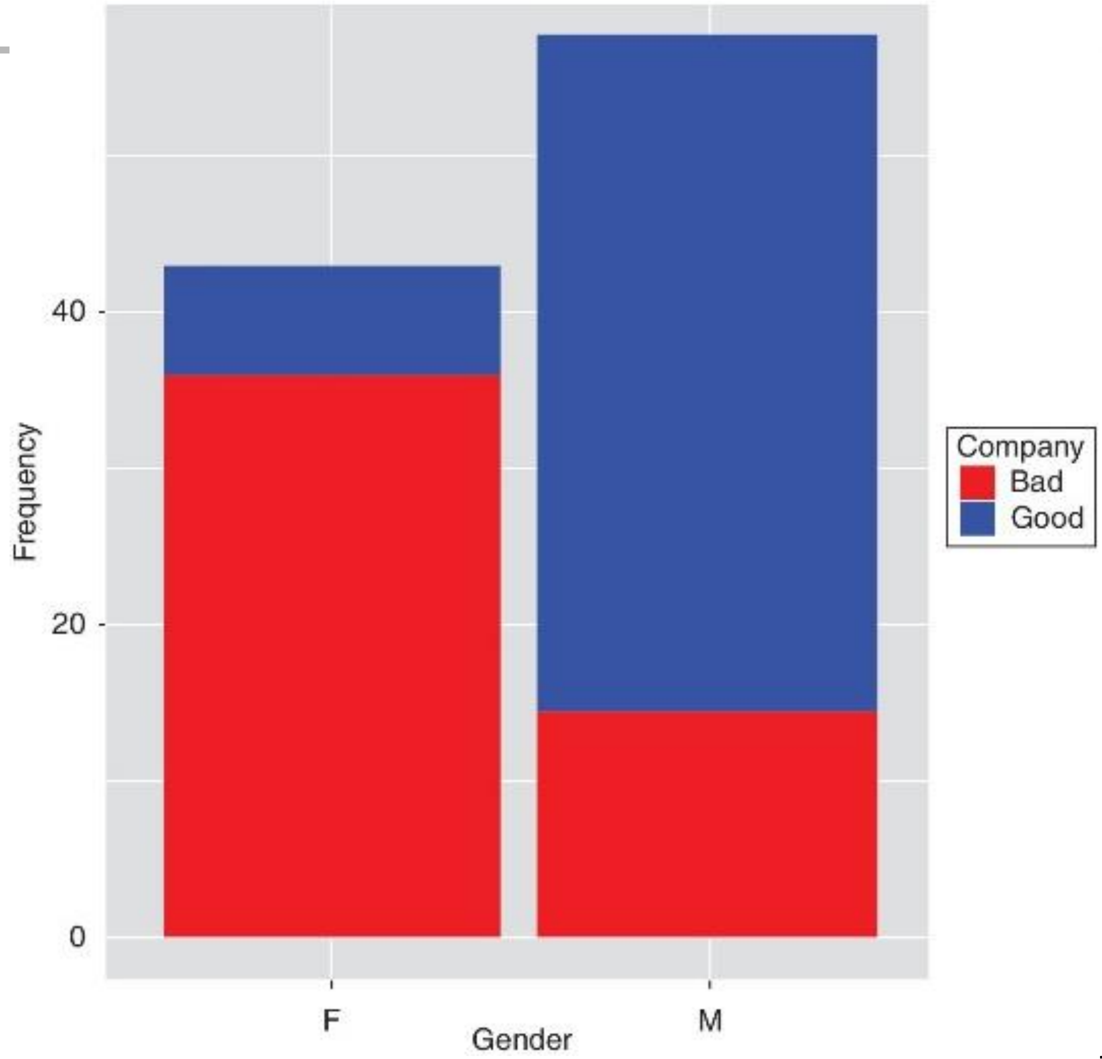
Descriptive Univariate Analysis: data visualization

- An important decision to draw a histogram is to define the number of cells
- The most advisable value is problem dependent
- As rule of thumb you can use a number around the square root of the number of values



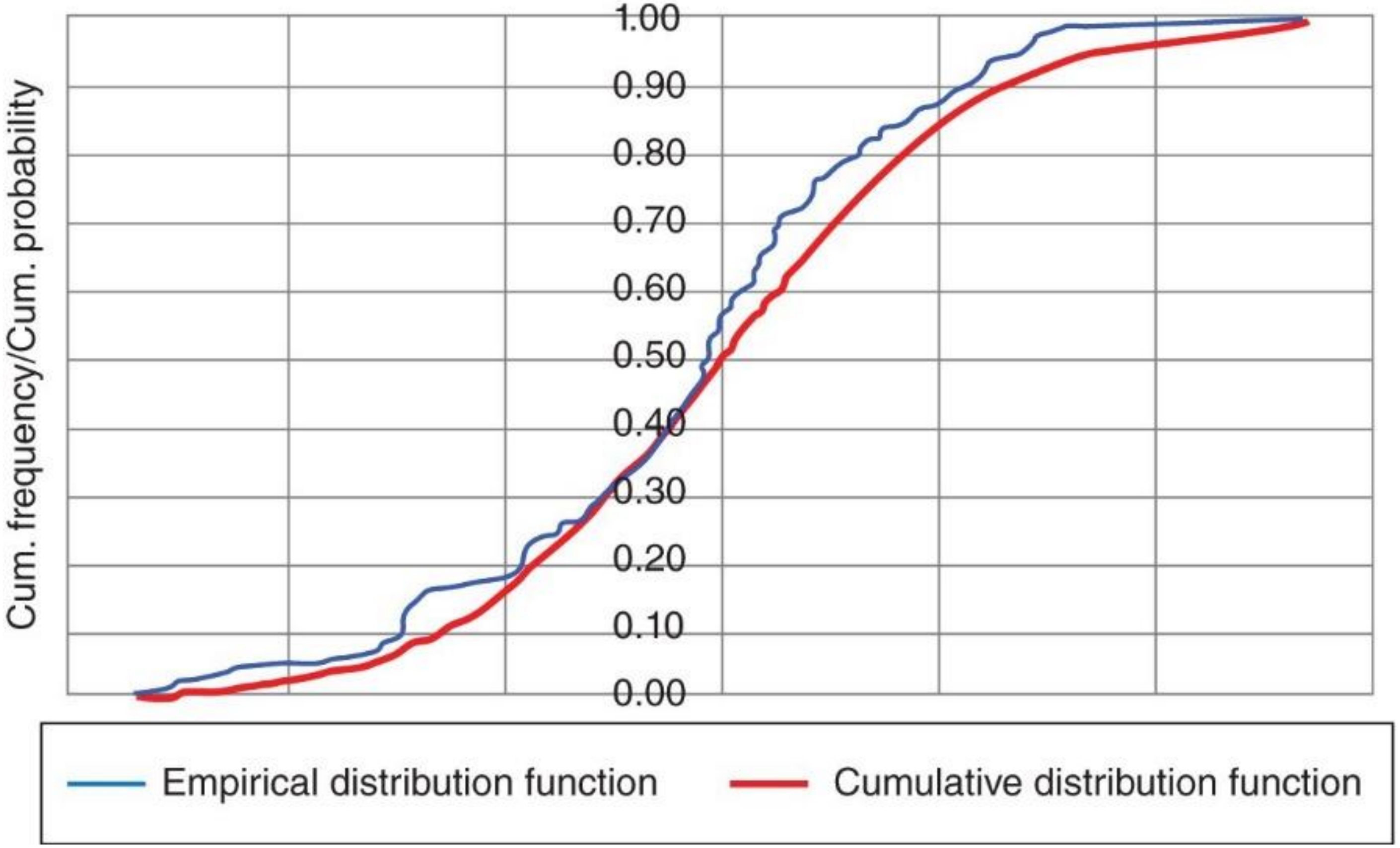
Descriptive Univariate Analysis: data visualization

- In a histogram, we can also separate the distributions for the values of some other attribute
- This is illustrated in the figure where the frequencies for the target value of "company" is split by gender



Descriptive Univariate Analysis: data visualization

- **Empirical distributions** are based in samples
- **Probability distributions** are about populations



Descriptive Univariate Analysis: statistics

- A statistic is a descriptor
- It describes numerically a characteristic of the sample or the population
- There are two main groups of univariate statistics:
  - **Location statistics**
  - **Dispersion statistics**

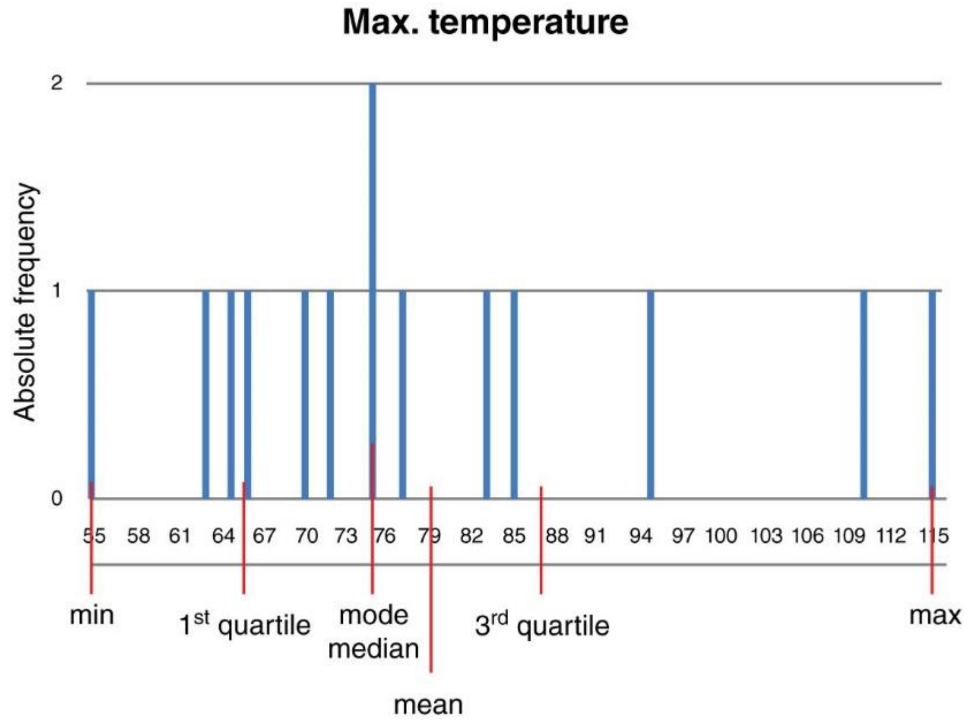
- **Location statistics**:
  - Minimum: is the lowest value
  - Maximum: is the largest value
  - Mean: is the average value
  - Mode: is the most frequent value
  - The value that is larger than:
    - 25% of all values is the 1st quartile
    - 50% of all values is the median or 2nd quartile
    - 75% of all values is the 3rd quartile

Example Location statistics

- *Let us use as example the attribute weight from our data set*

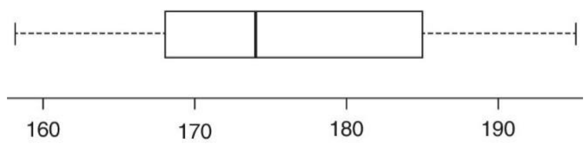
Location statistic	Weight (kg)
Min	55.00
Max	115.00
Mean or average	79.00
Mode	75.00
1st quartile	65.75
2nd quartile or mode	75.00
3rd quartile	87.50

- *Graphical representation of the statistics*



Descriptive Univariate Analysis: statistics

- **Box-plots** present the minimum, the 1st quartile, the median, the 3rd quartile and the maximum statistics, by this order, bottom-up or from left to right
  - *The attribute height*

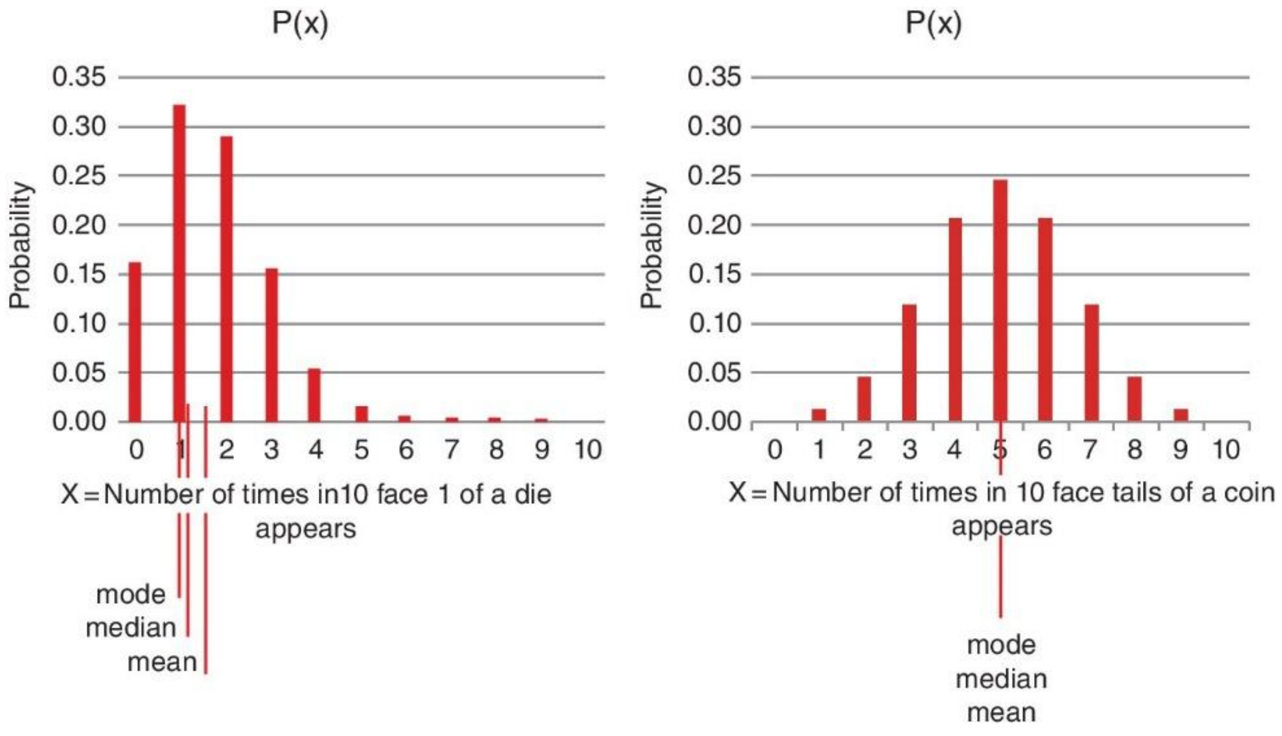


- Mean (or average), median and mode are known as **measures of central tendency**, because return a central value from a set of values

Location statistic	Nominal	Ordinal	Quantitative
Mean	No	Eventually	Yes
Median	No	Yes	Yes
Mode	Yes	Yes	Yes

Descriptive Univariate Analysis: statistics

- Box-plots can also be used to describe the symmetry/ skewness of an attribute
- The median or the mode are more robust as a central tendency statistic than the mean in the presence of extreme values or strongly skewed distributions



Descriptive Univariate Analysis: statistics

- Can the mean be used in ordinal scales?
- This is strongly arguable but there are examples of its use with numeric ordinal scales such as the Likert scale
- The Likert uses an ordered scale, e.g., integers from 1 (highest disagreement) to 5 (highest agreement)

Please circle the number that better fits your experience with the given information

I am satisfied with it
Strongly disagree 1 2 3 4 5 Strongly agree

It is simple to use
Strongly disagree 1 2 3 4 5 Strongly agree

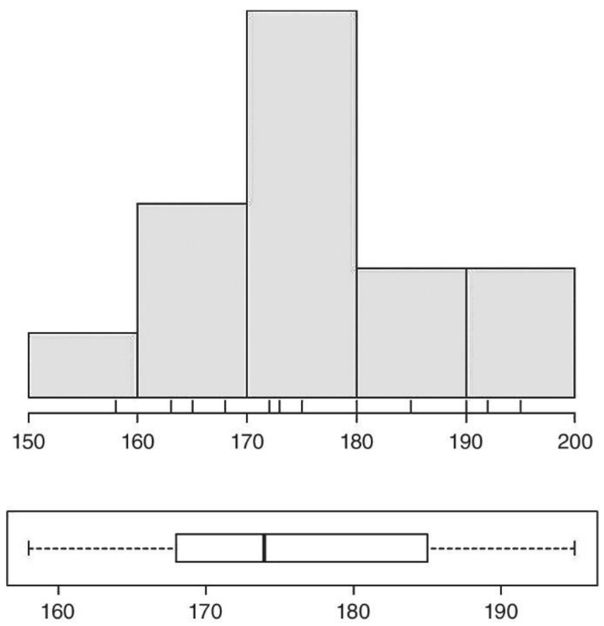
It has good graphics
Strongly disagree 1 2 3 4 5 Strongly agree

It is in accordance to my expectations
Strongly disagree 1 2 3 4 5 Strongly agree

Everything make sense
Strongly disagree 1 2 3 4 5 Strongly agree

Descriptive Univariate Analysis: statistics

- Plots can also be combined
  - *An example with the attribute length*



- There is only one value for the mean of a population
- There is only one value for the mean of a sample but can exist several samples from a single population
- The population mean and the sample mean are calculated in the same way but are differently represented:
  - μ_x is the mean population of x
  - $\bar{x}$ is a mean sample of x

Descriptive Univariate Analysis: statistics

- Dispersion statistic measures how distant the different values are
- **Dispersion statistics**:
  - **Amplitude**: is the difference between the maximum and the minimum values
  - **Interquartile range**: is the difference between the values of the 3rd and 1st quartiles
- Dispersion statistics (cont.):
  - **Mean absolute deviation** is a measure for the mean absolute distance between the observations and the mean
    - Its math formula for the population is:

$$MAD_x = \frac{\sum_{i=1}^{n} |x_i - \mu_x|}{n}$$

    - Its math formula for a sample is:

$$\overline{MAD_x} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n-1}$$

Descriptive Univariate Analysis: statistics

- Dispersion statistics (cont):
  - Standard deviation: is another measure for the typical distance between the observations and their mean
    - Its math formula for the population is: $\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu_x)^2}{n}}$
    - Its math formula for a sample is: $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$
    - The square of the standard deviation is named variance

- *Using again as example the weight attribute, dispersion statistics are as shown in the table*

Dispersion statistic	Weight (kg)
Amplitude	60.00
Interquartile range	21.75
$\overline{MAD}$	14.31
s	17.38

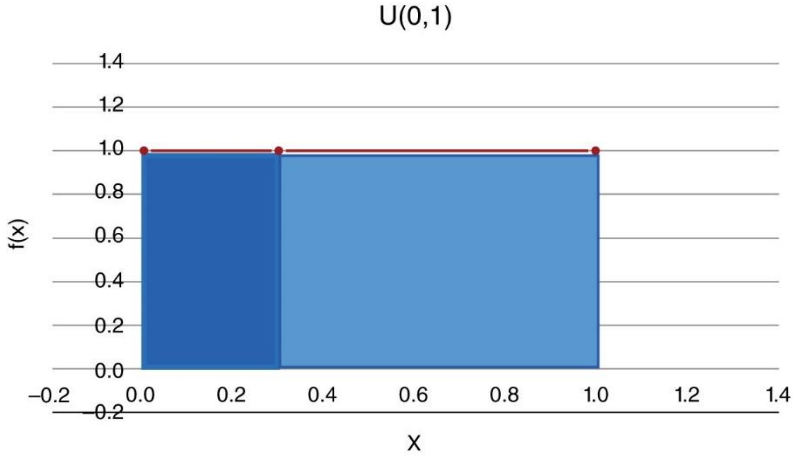
Descriptive Univariate Analysis: common univariate probability distributions

- Different events of our life follow already studied distributions
- E.g. the height of adult men, the value of a random number, or the number of cars passing in a given highway toll

- We present two of these distributions:
  - The Uniform distribution
  - The Normal distribution, also known as the Gaussian
- Both are continuous distributions and have known probability density functions

Descriptive Univariate Analysis: common univariate probability distributions

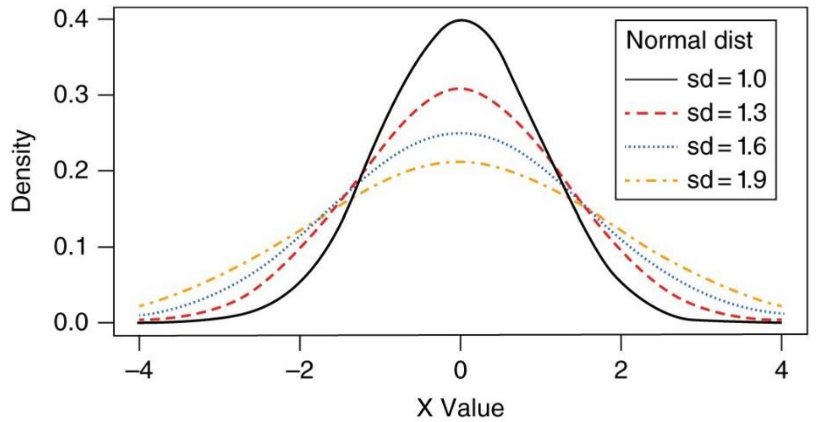
- An attribute x that follows the **uniform distribution** with parameters a and b, has equal frequency of occurrence of values in any interval of a given size



- $x \sim U(a,b)$
- $P(x < x_0) = \begin{cases} 0, if\ x_0 < a \\ \frac{x_0 - a}{b-a}, if\ a \leq x_0 \leq b \\ 1, if\ x_0 > b \end{cases}$
- $\mu_x = \frac{a+b}{2}$ $\qquad \sigma_x^2 = \frac{(b-a)^2}{12}$

Descriptive Univariate Analysis: common univariate probability distributions

- The Normal distribution
  - Physical quantities that are expected to be the sum of many independent factors (e.g., the men' height or the perimeter of 30 years old Quercus Rubra) typically have approximately Normal distributions



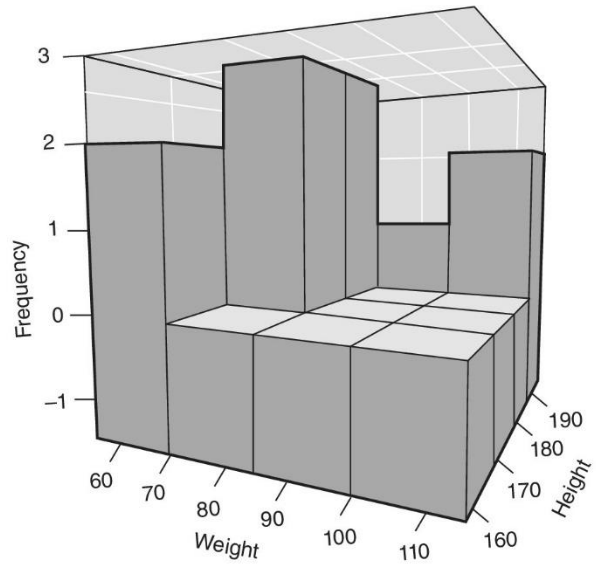
- The Normal distribution is a symmetric and continuous distribution with two parameters:
  - The mean localizes the highest point of the bell like distribution
  - The standard deviation defines how thin or larger the bell form of the distribution is
- $x \sim N(\mu_x, \sigma_x)$

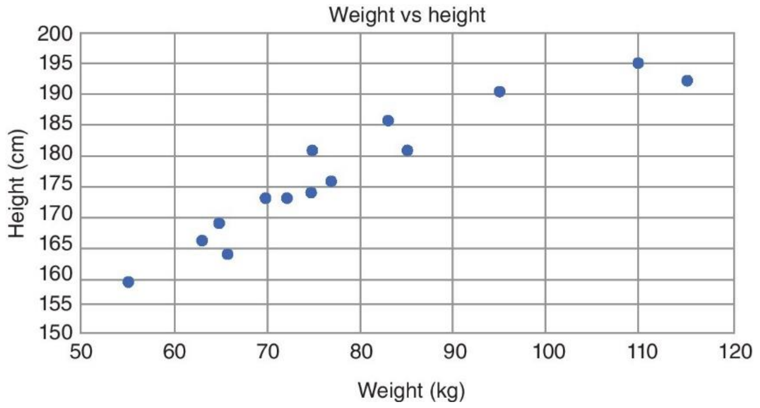
Descriptive bivariate analysis

- Talking about pairs of attributes, the relative behaviour between them
- Cases according to the scale types of the attributes:
  1. When the two attributes are quantitative
  2. When one of the attributes is qualitative and the other is quantitative
  3. When the two attributes are qualitative, at least one of them nominal
  4. When the two attributes are ordinal

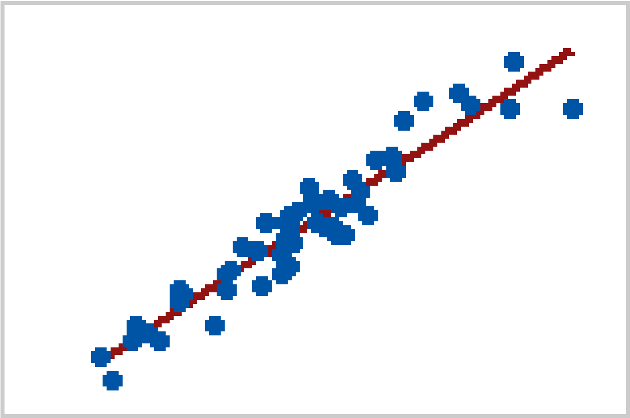
Descriptive bivariate analysis

- **When the two attributes of the pair are quantitative**
- There are several visualization techniques able to visually show the distribution of points with two quantitative attributes
  - One of these techniques is an extension of histograms, named 3-dimensional histograms
  - Another are the scatter plots

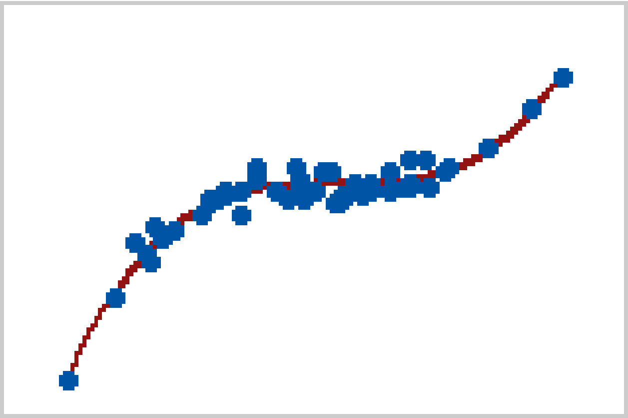




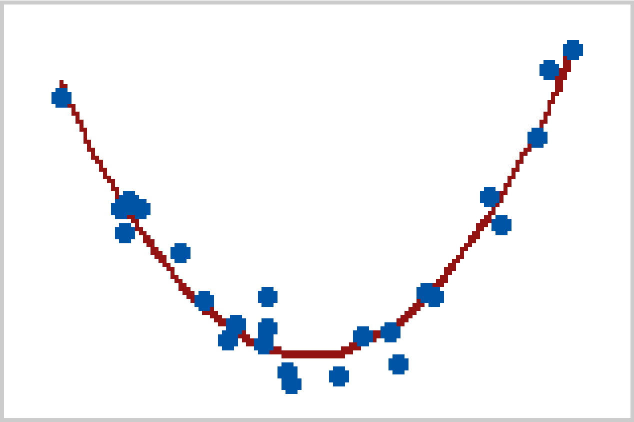
Bivariate relationship types



Linear



Monotonic



Non-Linear

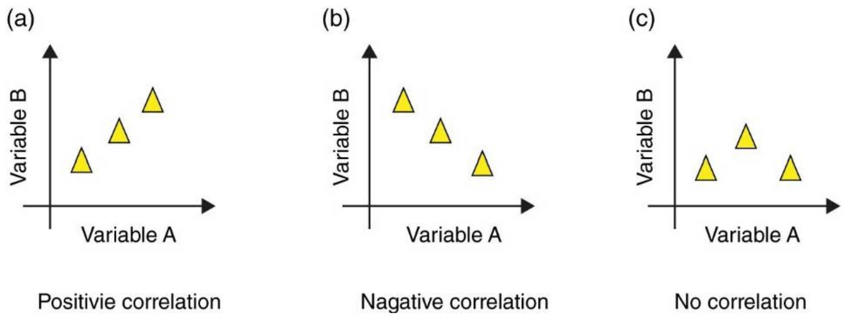
Source: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/linear-nonlinear-and-monotonic-relationships/

Descriptive bivariate analysis

- **Covariance**
  - Measures the degree of presence of linear relation between two attributes
  - Sample covariance:
    - $s_{ij} = cov(x_i, x_j) = \frac{1}{n-1} \times \sum_{k=1}^{n} (x_{ki} - \bar{x}_i) \times (x_{kj} - \bar{x}_j)$
  - The scale of the attributes influence the covariance values obtained

Descriptive bivariate analysis

- **Pearson correlation**
  - Sample Pearson correlation
    - $r_{ij} = cor\left(x_i, x_j\right) = \frac{cov\left(x_i, x_j\right)}{s_i \times s_j}$



  - Is scale independent: values always between [-1, 1]
  - If the points form:
    - an increasing line, the Pearson correlation coefficient will be 1
    - a decreasing line, its value will be -1
    - a horizontal line or a cloud without increasing or decreasing tendency, its value will be 0

Descriptive bivariate analysis

- The **Spearman's rank correlation**, as the name suggests, is based on rankings
- Compares how similar are the ranking positions of the values of the two attributes
  - $\rho_{ij} = \frac{\sum_{k=1}^{n}(rx_{ki} - \overline{rx}_i) \times (rx_{kj} - \overline{rx}_j)}{s_{rx_i} \times s_{rx_j}}$
- Ranking columns are calculated for both variable. For resolving ties we can use mean, min, or max value

Example

Friend	Weight (kg)	Height (cm)	Ranked Weight	Ranked Height
Andrew	77	175	9.0	8.0
Bernhard	110	195	13.0	14.0
Carolina	70	172	5.0	5.5
Dennis	85	180	11.0	9.5
Eve	65	168	3.0	4.0
Fred	75	173	7.5	7.0
Gwyneth	75	180	7.5	9.5
Hayden	63	165	2.0	3.0
Irene	55	158	1.0	1.0
James	66	163	4.0	2.0
Kevin	95	190	12.0	12.0
Lea	72	172	6.0	5.5
Marcus	83	185	10.0	11.0
Nigel	115	192	14.0	13.0

- *Pearson correlation*
  - $r_{weight,height} = 0.94$
- *Spearman's rank correlation*
  - $\rho_{weight,height} = 0.96$
- *Note the tie breaker ranked values. Mean is used here.*

Descriptive bivariate analysis

- **When one of the attributes is qualitative and the other is quantitative**
  - Box-plots can be used as previously discussed using one box plot for the values of the quantitative attribute per each different value of the qualitative attribute

Descriptive bivariate analysis

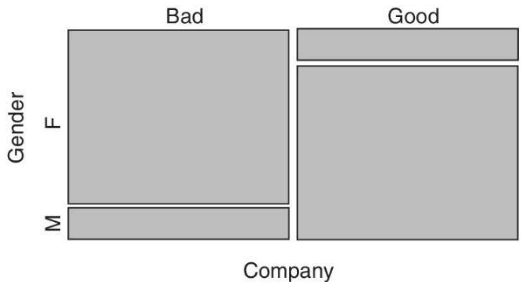
- **When one of the attributes is qualitative and the other is quantitative**
  - **Contingency tables**
    - They have a matrix like format, i.e., cells in a square with labels in the left and in the top
    - In the right most column are the totals per row while in the bottom most row are the totals per column
    - The bottom right corner has the total number of values

Descriptive bivariate analysis

- Two qualitative attributes, at least one of them nominal
  - **Mosaic plots**
    - Show the same information than contingency tables but in a more appealing visual way
    - The areas displayed are proportional to their relative frequency

Contingency

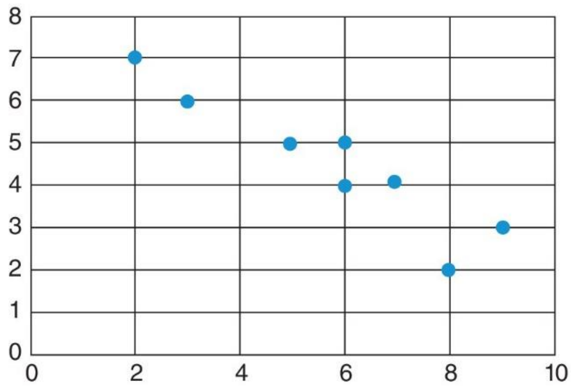
		Company		
		Good	Bad	
Gender	Male	6	2	8
	Female	1	5	6
		7	7	14



Mosaic

Descriptive bivariate analysis

- **When the two attributes are ordinal**
  - Any of the methods previously described to bivariate analysis can also be used in the presence of two ordinal attributes:



- The Spearman's rank correlation should be used instead of the Pearson correlation
- Scatter plots with ordinal attributes
  - Use the jitter effect, which add a random deviation to the values, in order to avoid that all points with the same values are represented as a unique point
- Contingency tables can be used and mosaic plots too
  - The values should be in increasing order

Questions?