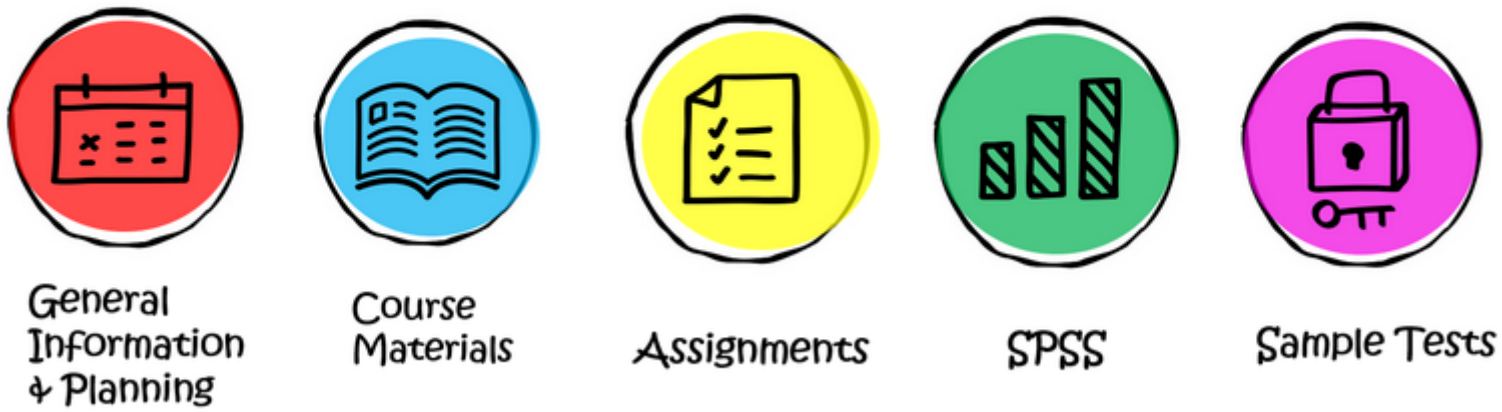
Statistical Techniques for CS/BIT

[Stat] Home Page



- **General goal**: give a probability model for situations where chance plays a role and argue correct statements for it.

- **Descriptive Statistics (Data analysis):** summarize observed data and present them graphically.

- **Probability Theory:** if the probability model is fully specified, we can exactly compute probabilities of events and expectations.

- **Statistics:** if a probability model is not completely specified, we want to give statements about characteristics of a population on the basis of a relatively small sample, drawn from this population.

- Introduce the course
- Meet your teachers (online)
- **Statistical Techniques**

  Chapter 1
- Break
- Statistical Techniques

  Chapter 1 (cont'd)

Probability Theory:

A random variable X : either discrete or continuous (or mixed).

Discrete distributions are given with the **probability function** $\boldsymbol{P(X = x)}$, where $\sum_x P(X = x) = 1$.

$$\mu = E(X) = \sum_x x \cdot P(X = x)$$

$$\sigma^2 = var(X) = E(X - \mu)^2, \sigma = \sqrt{var(X)}$$

Continuous distributions are given by the **density function** $\boldsymbol{f(x)}$, such that $\boldsymbol{P(a < x < b)}$ = **the area** $\boldsymbol{\int_a^b f(x)dx}$

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx.$$

$var(X) = E(X^2) - (EX)^2$ {while $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$}

- Introduce the course
- Meet your teachers (online)
- **Statistical Techniques**

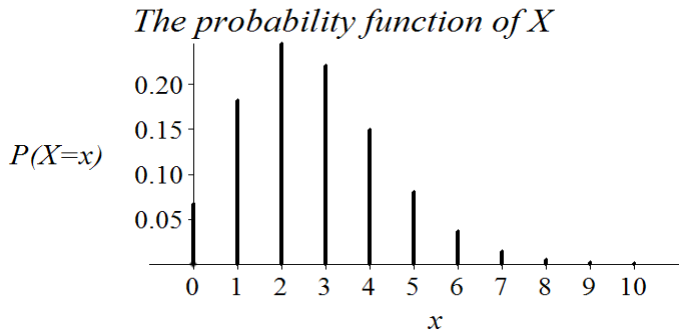
  Chapter 1
- Break
- Statistical Techniques

  Chapter 1 (cont'd)

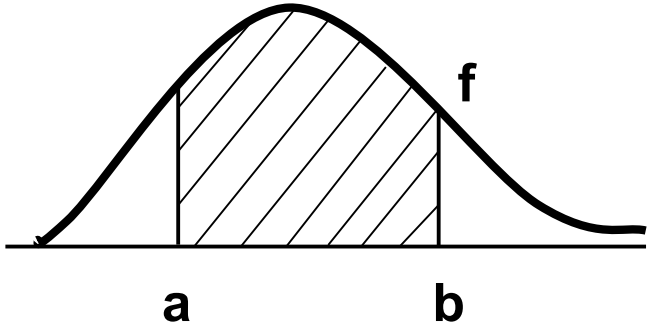
Probability Theory:

A random variable X : either discrete or continuous (or mixed).

Discrete distributions are given with the **probability function**



Continuous distributions are given by the **density function**



- Introduce the course
- Meet your teachers (online)
- **Statistical Techniques**

  Chapter 1
- Break
- Statistical Techniques

  Chapter 1 (cont'd)

Discrete distributions

The **Poisson** distribution: X counts the number of "rare events" in an area and/or period, with expectation μ.

$$P(X = x) = \frac{\mu^x}{x!} e^{-\mu},$$

$x = 0,1,2, \ldots$

and $E(X) = var(X) = \mu$

The **hypergeometric** distribution: n draws without replacement from R red and $N - R$ white balls; $X =$ "# of red balls"

$$P(X = x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{R}},$$

$x = 0, 1, \ldots, n$

The **binomial** distribution applies to situations, where we count the number of successes in n Bernoulli trials with success rate p: $X =$ "# of successes":

$X \sim B(n, p)$.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

$x = 0, 1, \ldots, n$

$E(X) = np$ and $var(X) = np(1-p)$

- Introduce the course
- Meet your teachers (online)
- **Statistical Techniques** Chapter 1
- Break
- Statistical Techniques Chapter 1 (cont'd)

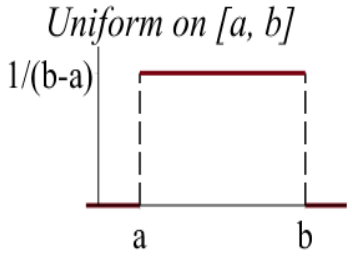
Normal approximation of the binomial distribution:

with the $\boldsymbol{N(np, np(1-p))}$-dist., if $n \geq 25, np \geq 5$ and $n(1-p) \geq 5$

Do not forget **continuity correction**,

e.g. $P(X \leq 9) \overset{c.c.}{=} P(X \leq 9.5)$

Continuous distributions



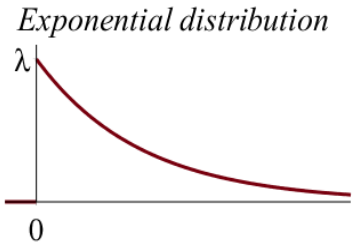
Uniform $X \sim U(a,b)$

$f(x) = \frac{1}{b-a}, a \leq x \leq b$

$E(X) = \frac{a+b}{2}$

$var(X) = \frac{(b-a)^2}{12}$

Model for random numbers, drawn from an interval, especially the interval $(0,1)$



Exponential dist.

$X \sim Exp(\lambda)$

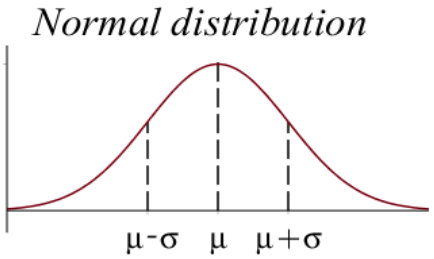
$f(x) = \lambda e^{-\lambda x}, x \geq 0$

$E(X) = \frac{1}{\lambda}$

$var(X) = \frac{1}{\lambda^2}$

$P(X > x) = e^{-\lambda x}$

Model for waiting times, interarrival times and lifetimes



Normal distr.

$X \sim N(\mu, \sigma^2)$

$E(X) = \mu$

$var(X) = \sigma^2$

$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Model for "natural quantities" variables in nature, economy, etc.

- Introduce the course
- Meet your teachers (online)
- **Statistical Techniques Chapter 1**
- Break
- Statistical Techniques Chapter 1 (cont'd)

Some applications of the normal distribution

1. If X is $N(\mu, \sigma^2)$, then $Y = aX + b$ is $N(a\mu + b, a^2\sigma^2)$

2. If $X_1, \ldots, X_n$ is a random sample, drawn from $N(\mu, \sigma^2)$, then:

$$\sum_{i=1}^{n} X_i \quad \text{is } N(n\mu, n\sigma^2),$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{is } N\left(\mu, \frac{\sigma^2}{n}\right) \text{ and}$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \quad \text{is the stand. dev. of the sample mean.}$$

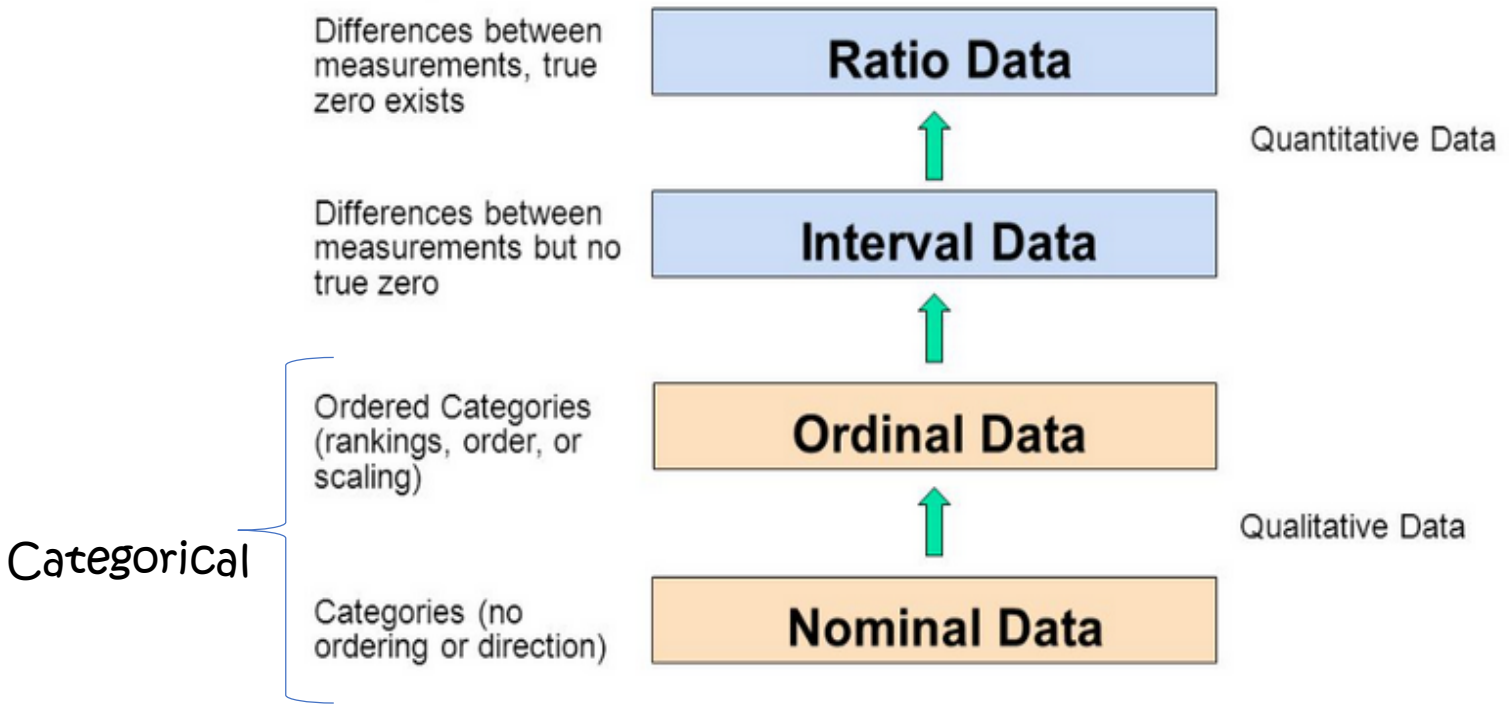
3. The Central Limit Theorem states that, if the **population is not normal,** then the distributions in 2. are **approximately** true **for sufficiently large n**. (rule of thumb: $n \geq 25$).

- Introduce the course
- Meet your teachers (online)
- **Statistical Techniques**

  Chapter 1
- Break
- Statistical Techniques

  Chapter 1 (cont'd)

Descriptive Statistics (Data analysis): summarize observed data and present them graphically.



- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

  Chapter 1
- Break
- **Statistical Techniques**

  Chapter 1 (cont'd)

Descriptive Statistics (Data analysis): summarize observed data and present them graphically.

Table 4.1 Sample Demographic Data in Its Original Form

	Age	Gender	Highest Degree	Previous Experience In Software A
Participant 1	34	Male	College	Yes
Participant 2	28	Female	Graduate	No
Participant 3	21	Female	High school	No

Table 4.2 Sample Demographic Data in Coded Form

	Age	Gender	Highest Degree	Previous Experience In Software A
Participant 1	34	1	2	1
Participant 2	28	0	3	0
Participant 3	21	0	1	0

Choose the appropriate measure!

Mean

Mode (occurs the most frequently)

Median

- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

  Chapter 1
- Break
- Statistical Techniques

  Chapter 1 (cont'd)

Dot diagram (for small samples)

A line of numbers, on which the observations are presented as "dots": equal observations are stacked.

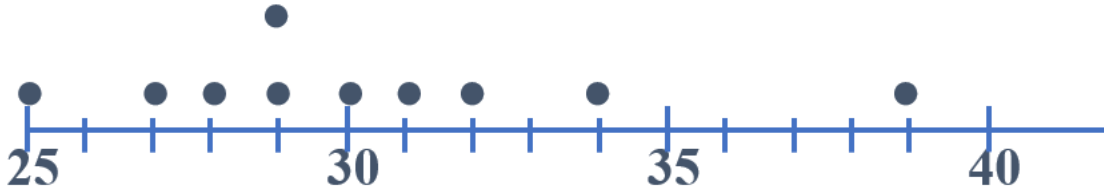
Applicable for quantitative variables if the sample is small (≤ 40 observations)

Example (n = 10 starting salaries in € per year):

x_i : 25, 28, 39, 30, 32, 29, 31, 34, 29, 27

Order from small to large: the order statistics.

$x_{(i)}$: 25, 27, 28, 29, 29, 30, 31, 32, 34, 39



- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

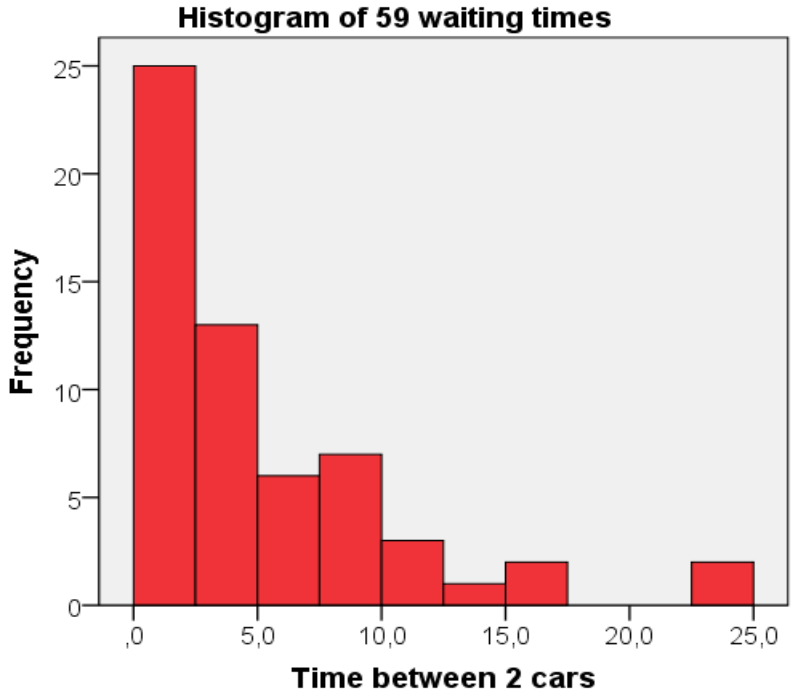
  Chapter 1
- Break
- **Statistical Techniques**

  Chapter 1 (cont'd)

Histogram

- Choose a distribution in intervals: not too many nor too few observations per interval
- Count the number of observations in each interval, **the frequency**
  or determine the **relative frequency** = $\frac{frequency}{n}$
- Build a rectangle above each interval and choose as height either the frequency or the relative frequency

A table of intervals with their (relative) frequency is called the **frequency distribution of the dataset:** the histogram is its graphical presentation



- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

  Chapter 1
- Break
- **Statistical Techniques**

  Chapter 1 (cont'd)

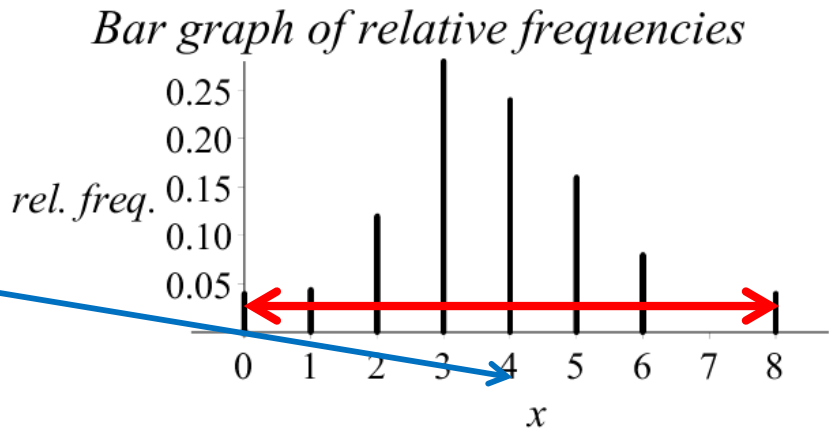
Bar graph

The variable is quantitative and discrete (integer values 0, 1, 2…, 8)

Sample of $n = 25$ observations, we record the number of times that 0, 1, .. , 8 are observed:

# accidents	0	1	2	3	4	5	6	7	8	**Total**
Frequency	1	1	3	7	6	4	2	0	1	**25**
Relative freq.	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{3}{25}$	$\frac{7}{25}$	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	0	$\frac{1}{25}$	**1**

Graphical presentation



"Center"

Spread: range

- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

  Chapter 1
- Break
- **Statistical Techniques**

  Chapter 1 (cont'd)

Measures of Centre (quantitative variables)

Mean: "arithmetic average"

$$\text{The } \textbf{Sample Mean}: \ \bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Distinguish *sample* mean $\bar{x}$ and *population* mean μ

Median (m): the middle observation

the observations are arranged from small to large (*order statistics*)

$x_1, x_2, \ldots, x_n \rightarrow x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

If n, the number of observations, is even, then compute the mean of the middle two observations.

Mode: The most frequently occurring observation

- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

  Chapter 1
- Break
- **Statistical Techniques**

  Chapter 1 (cont'd)

Percentiles and quartiles

- The median m is also called the 50th percentile: (about) 50% of the observations is smaller and 50% is greater than the median m
- The quartiles Q_1, m and Q_3 are the 25th, 50th and 75th percentiles: they split the observations in 4 roughly equal quarters.

Definition 1.2.5 The (sample) **median** is $$\boldsymbol{m} = \begin{cases} \boldsymbol{x}_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is } \textbf{odd} \\ \frac{1}{2}\left[\boldsymbol{x}_{\left(\frac{1}{2}n\right)} + \boldsymbol{x}_{\left(\frac{1}{2}n+1\right)}\right] & \text{if } n \text{ is } \textbf{even} \end{cases}$$

Without formal definition we found a univocal method to determine the $\boldsymbol{k}^{\text{th}}$ **percentile of** $\boldsymbol{n}$ **observations** $\boldsymbol{x_1, x_2, \ldots, x_n}$:

- Compute k% of n: $c = \frac{k}{100} \cdot n$.
- If $\boldsymbol{c}$ is **not integer**, round c upward to the first larger integer $[\boldsymbol{c}]$: the k^{th} percentile is $x_{([c])}$.)
- If $\boldsymbol{c}$ **is integer**, then the k^{th} percentile $= \frac{x_{(c)}+x_{(c+1)}}{2}$.

- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

  Chapter 1
- Break
- Statistical Techniques

  Chapter 1 (cont'd)

Percentiles and quartiles

A simple dot diagram (for small samples)

- A line of numbers, on which the observations are presented as "dots": equal observations are stacked.
- **Applicable** for quantitative variables if the sample is small (≤ 40 observations)

Example (n = 10 starting salaries in k€ per year):

x_i : 25, 28, 39, 30, 32, 29, 31, 34, 29, 27

Order from small to large: the **order statistics.**

$x_{(i)}$: 25, 27, 28, 29, 29, 30, 31, 32, 34, 39

- n = 10 is even, so the median is the average of $x_{(5)}$ and $x_{(6)}$: m = (29+30)/2 = 29.5
  Note that 50% of 10 equals 5, an integer.
- Q_1 is the 25th percentile: 25% of 10 is 2.5, so "2.5 observations" are smaller and 7.5 observations are greater.
  So Q_1 is 28, the 3rd observation in magnitude: $x_{(3)}$
  Note that 25% of 10 the rational number 2.5, rounded op to the next greater integer 3.
- What is the 80th percentile? 80% of 10 is 8 (integer), so the 80th perc. is the average of $x_{(8)}$ and $x_{(9)}$: (32+34)/2 = 33

- Introduce the course
- Meet your teachers (online)
- Statistical Techniques
  Chapter 1
- Break
- **Statistical Techniques**
  Chapter 1 (cont'd)

Box-plot

5-number summary of the observations

- The box plot graphs the 5-number summary of the observations: the smallest, the quartiles (Q_1, m, Q_3) and the greatest

- The "box" indicates the position of Q_1, m, Q_3 and the "whiskers" the smallest and the greatest observations.

A simple dot diagram (for small samples)

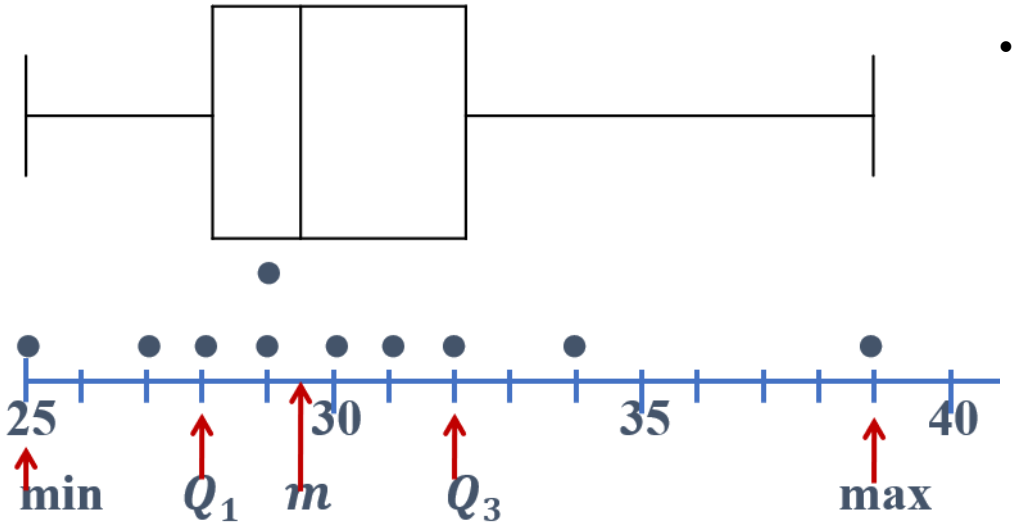
- A line of numbers, on which the observations are presented as "dots": equal observations are stacked.
- **Applicable** for quantitative variables if the sample is small (≤ 40 observations)

Example (n = 10 starting salaries in ke per year):

x_i : 25, 28, 39, 30, 32, 29, 31, 34, 29, 27

Order from small to large: the order statistics.

$x_{(i)}$: 25, 27, 28, 29, 29, 30, 31, 32, 34, 39



- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

  Chapter 1
- Break
- **Statistical Techniques**

  Chapter 1 (cont'd)

Measures of Variability

Range: the range r = largest – smallest observation

The Inter Quartile Range: $IQR = Q_3 - Q_1$

Variance: The **Sample Variance**: $\mathbf{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

Standard deviation: the **Sample Standard Deviation** $s = \sqrt{s^2}$
Distinguish *sample* variance s^2 ↔ *population* variance σ^2

Resistant measures are *not sensitive* for outliers:
e.g. - the Median and the IQR are resistant.
Non-resistant measures are $\overline{x}$, s and s^2

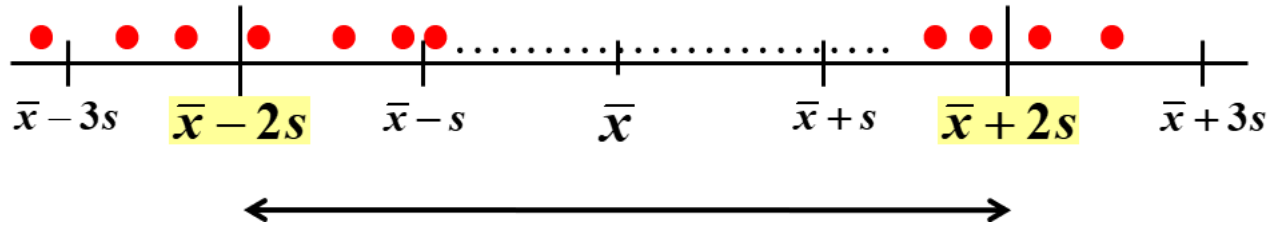
- Introduce the course
- Meet your teachers (online)
- Statistical Techniques
  Chapter 1
- Break
- **Statistical Techniques**
  Chapter 1 (cont'd)

Standard Deviation

Chebyshev`s rule: $P(|X - \mu_X| \geq c) \leq \frac{var(X)}{c^2}$ $(c > 0)$

This rule also applies to the mean $\bar{x}$ and standard deviation s:

if $c = ks$ then the interval $(\bar{x} - ks, \bar{x} + ks)$ contains at least $\left(1 - \frac{1}{k^2}\right) 100\%$ of the observations (with probability $\leq \frac{1}{k^2}$ outside the interval)



$k = 2$: $(\bar{x} - 2s, \bar{x} + 2s)$ contains at least $1 - \frac{1}{k^2} = 75\%$ of the observations

$k = 3$: $(\bar{x} - 3s, \bar{x} + 3s)$ contains at least $\frac{8}{9} \approx 89\%$ of the observations

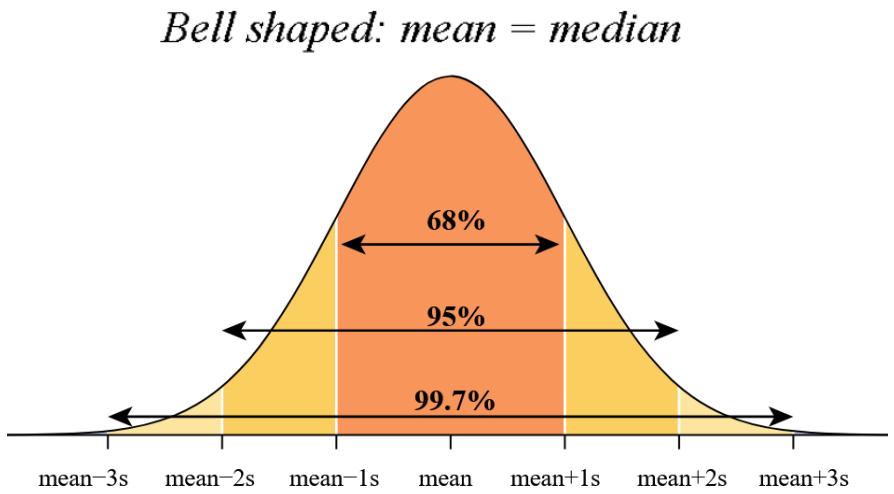
- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

  Chapter 1
- Break
- **Statistical Techniques**

  Chapter 1 (cont'd)

The Empirical Rule

This rule is **only valid for bell shaped** (mound shaped) histograms and distributions



Interval	Empirical rule Approximate percentage observations in interval	*General According to "Chebyshev"*
$(\bar{x}-s,\ \bar{x}+s)$	**68%**	*≥ 0%*
$(\bar{x}-2s,\ \bar{x}+2s)$	**95%**	*≥ 75%*
$(\bar{x}-3s,\ \bar{x}+3s)$	**99.7%**	*≥ 89%*

- Introduce the course
- Meet your teachers (online)
- Statistical Techniques

  Chapter 1
- Break
- **Statistical Techniques**

  Chapter 1 (cont'd)

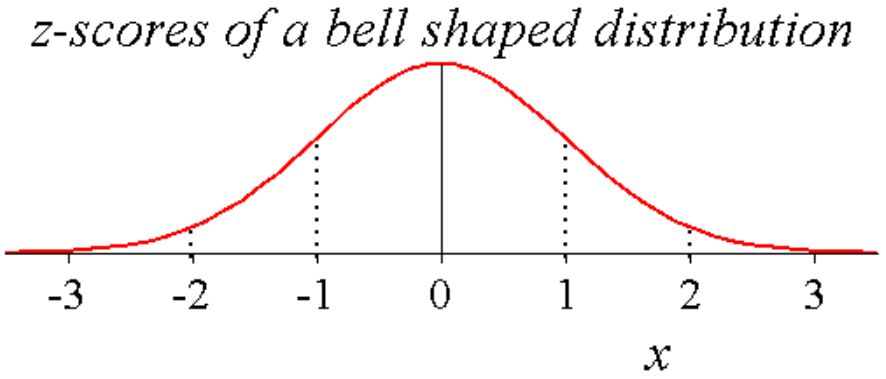
The z-scores

For samples with mean $\bar{x}$ and standard deviation s :

the $z-score$ of an observation x is $\frac{x-\bar{x}}{s}$.

Interpretation of the z-score:

- Measure for deviation from the mean: "*the number of standard deviations smaller or greater than the mean*"
- **Empirical rule for bell shaped distributions:**
  about **68%** of the observations has z-score between **-1 and 1**,
  about **95%** between **-2 and +2**
  and
  about **99.7%** between **-3 and +3**



For ***populations*** with mean μ and standard deviation σ:

the $z-score$ of an observation or value x is $\frac{x-\mu}{\sigma}$.

- Introduce the course
- Meet your teachers (online)
- Statistical Techniques
  Chapter 1
- Break
- **Statistical Techniques**
  Chapter 1 (cont'd)