Probability and Statistics

teacher:	Martin Schindler
	KAP, tel. 48 535 2836, building G
consult. hours:	by agreement
e-mail:	martin.schindler@tul.cz

last update: 14. září 2016

Course outline:

1. Descriptive statistics: Basic descriptive statistics. Types of variables, frequency distribution, graphical data processing. Basic characteristics of location and variability, ordered data.
2. The calculations of basic characteristics of the ordered data. Boxplot. Multidimensional data - correlation coefficient.
3. Probability theory: event, the definition of probability, probability properties.
4. The independence of events, conditional probability. Bayes theorem.
5. Random variable. Probability distribution. Distribution function, density, quantile function. Characteristics of random variables.
6. Discete distribution: alternative, binomial, geometric, hypergeometric, Poisson.
7. Normal distribution, Central limit theorem - Moivreova - Laplace theorem. Continuous distribution: uniform, exponential, Student and F distributions.

8. Multivariate random variable (vector). Dependence - covariance and correlation coefficient.
9. Introduction to Mathematical Statistics. Point estimates, interval estimates for parameters of normal and binomial distribution.
10. Basic concepts of statistical hypothesis testing. Tests of hypotheses on the parameters of normal distribution.
11. Non-parametric tests. Tests of hypotheses about the parameters of the binomial distribution
12. Goodness of fit tests and their application.
13. Correlation and regression. . Spearman's coefficient of serial correlation.
14. Linear regression, method of least squares.

Statistics

- **Statistics** is a discipline that deals with the collection, organization, analysis, interpretation and presentation of data. Only events appearing at a large set of cases, not only at individual cases, are of interest.
- **Data set** is a set of **statistical units** (inhabitants, towns, companies,...), on which we measure values of **variable**(age, number of inhab., turn-over,...)
- Measurements are recorded in an appropriate **scale (levels of measurement)**.
- On one unit we can measure several characteristics - that allows to study correlation (Is there a relationship between height and weight in the studied population?).

We can treat the data set two different ways:

1. **Descriptive statistics** - we make conclusions only for the studied data set from the observed data (we measured all the units in the population we want to describe)
2. **Mathematical (inferential) statistics** - studied data set is treated as a **sample data** – set of units randomly and independently selected from **target population** that is large (cannot be explored completely for time, financial or organizational reasons). We want to make conclusions about the whole population only from the sample values (second half of semester).

Types of scales

- **zero-one** (male/female, smoker/nonsmoker)
- **nominal** (marital status, eye color) - disjoint categories that cannot be ordered
- **ordinal** (education level, satisfaction level) - nominal scale with ordered categories
- **interval** (temperature in Celsia degree, year of birth) - values are numeric, distance between the neighboring values is constant, an arbitrarily-defined zero point
- **ratio** (weight, hight, number of inhabitants) - values are given in a multiple of a unit quantity, zero means nonexistence of the measured characteristic.

  - **Qualitative**: zero-one, nominal, ordinal
  - **Quantitative (continuous)**: interval, ratio

Example - one-dimensional

- one-dimesional data

- we study IQ scores of 62 pupils from 8-th grade in a certain primary school
- how to describe and evaluate what have the data in common or how much they differ from each other?
- from the data set (values of the variable) we calculate characteristics (characteristics of location, variability, shape of the distribution, for multi-dimensional dat also characteristics of correlation)
- a characteristic (a statistic) expresses (evaluate) given property by one number

Example - data set

measured values denote by $x_1, x_2 \ldots, x_n$, now $n = 62$.

107	141	105	111	112	96	103	140	136	92
92	72	123	140	112	127	120	106	117	92
107	108	117	141	109	109	106	113	112	119
138	109	80	111	86	111	120	96	103	112
104	103	125	101	132	113	108	106	97	121
134	84	108	84	129	116	107	112	128	133
96	94								

ordered data set denote by $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

Frequency distribution

- If the values are often repeated we can produce so called **frequency table**.
- If the variable is continuous and n (number of observations) is large, it is advisable to divide the range of values into M intervals with endpoints
  $a = a_0 < a_1 < a_2 < ... < a_{M-1} < a_M = b$.
- all the observations from a interval can be represented by one value (usually the center of the interval) x_i^*, $i = 1, \dots, k$.
- let n_i denotes number of observations that falls to interval $\langle a_{i-1}, a_i), i = 1, \dots, M$ – so called **absolute frequency** (Intervals are called **classes**).
- **cumulative frequency** N_i gives the number of observations in the (i-th) and all the preceding classes
- numbers n_i/n gives **relative frequency**.

Example - frequency distribution

Interval	x_i^*	absol. n_i	n_i/n	cumul. N_i	N_i/n
< 80	75	1	0.016	1	0.016
$\langle 80, 90)$	85	4	0.065	5	0.081
$\langle 90, 100)$	95	8	0.129	13	0.210
$\langle 100, 110)$	105	18	0.290	31	0.500
$\langle 110, 120)$	115	14	0.226	45	0.726
$\langle 120, 130)$	125	8	0.129	53	0.855
$\langle 130, 140)$	135	5	0.081	58	0.935
≥ 140	145	4	0.065	62	1.000

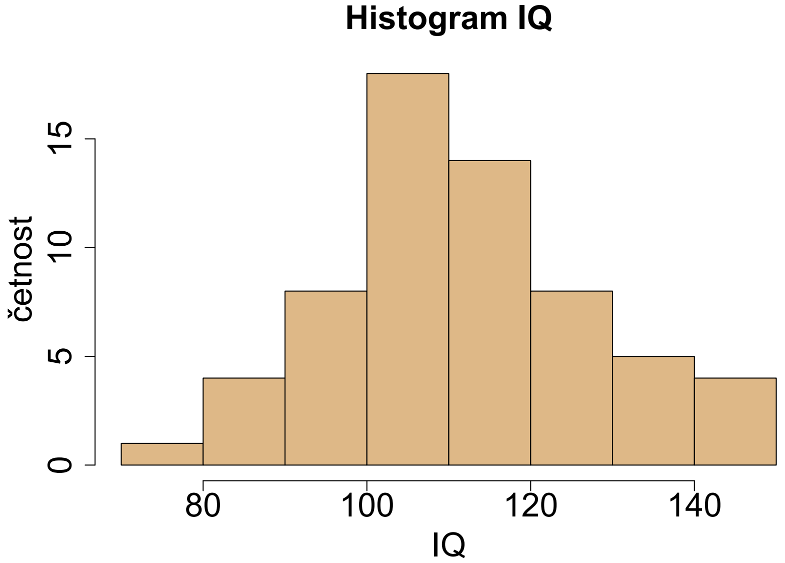
Histogram

- graphic display of frequency distribution
- we assign to each interval a box, such that its area is proportional to the frequency of the interval
- most often the intervals have equal length (often appropriately rounded), then the hight of the boxes corresponds with the frequencies.
- problem: choice of the number of intervals M
  we can use e.g. Sturges rule:

$$M \approx 1 + 3.3 \log_{10}(n) \doteq 1 + \log_2(n)$$

- for our example: $1 + \log_2(62) = 6.95$

Example - histogram



Characteristic of location

- allows to characterize the level of a variable by one number - evaluation, how the observations are small or large.
- it should hold for a characteristic m of a data set x , that it naturally changes with the change of the scale, i.e. for arbitrary constants a, b:

$$m(a \cdot x + b) = a \cdot m(x) + b$$

- if we add a constant b to all observations, then the characteristic gets larger by b
- if we multiple each observation by a, then the resulting characteristic gets bigger a-times

Aritmetic mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

- for our example: $\overline{x} = \frac{1}{62}(107 + 141 + \ldots + 94) = 111.0645$
- sensitive to outliers. Only for quantitative scales.
- can be computed from the frequency table as a weighted average

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{M} n_i x_i^* = \frac{\sum_{i=1}^{M} n_i x_i^*}{\sum_{i=1}^{M} n_i} = \frac{1 \cdot 75 + 4 \cdot 85 + \ldots + 4 \cdot 145}{62} = 111.7742$$

- for zero-one variable: $\frac{\text{number of ones}}{\text{number of zeros andones}}$ = relative frequency (percent) of ones (observations with the given property).
- for our example $y_i = 0$ (i-th pupil is a man) , $y_i = 1$ (i-th pupil is a female): $\overline{y} = \frac{32}{62} = 0.516$

Mode

- $\hat{x}$ - most frequent value
- can be used even for nominal and ordinal scales
- not necessarily unique
- for our example:

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

$$\hat{x} = 112$$

Median

- $\tilde{x}$ - number that divides the ordered sample into two equal halves, is located in the middle of the ordered sample

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} \qquad \text{for } n \text{ odd}$$

$$\tilde{x} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) \qquad \text{for } n \text{ even}$$

- robust - not influenced by large changes of a few values. often also for ordinal scale. For our example:

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

$$\tilde{x} = \frac{1}{2}\left(x_{(31)} + x_{(32)}\right) = 110$$

Quantiles: percentiles, deciles, quartiles

α-quantile x_α ($\alpha \in (0,1)$) - Dividing ordered data into two part, such that α-ratio of the smallest values is smaller than x_α

- $x_\alpha = x_{(\lceil \alpha n \rceil)}$,
  where $\lceil a \rceil$ denotes a, if it is a integer, otherwise the nearest larger integer.
- special quantiles:

  percentiles: $\alpha = 0.01, 0.02, \ldots, 0.99$
  deciles: $\alpha = 0.1, 0.2, \ldots, 0.9$
  quartiles: $\alpha = 0.25, 0.5, 0.75$

  1-st (lower) quartile is denoted by $Q_1 = x_{0.25}$
  3-rd (upper) quartile is denoted by $Q_3 = x_{0.75}$
- median is the 50% quantile, 50-th percentile, 5-th decile a 2-nd quartile

Example - quantiles

72	80	84	84	86	92	92	92	94	96
96	96	97	101	103	103	103	104	105	106
106	106	107	107	107	108	108	108	109	109
109	111	111	111	112	112	112	112	112	113
113	116	117	117	119	120	120	121	123	125
127	128	129	132	133	134	136	138	140	140
141	141								

- 1-st quartile $Q_1 = x_{0.25} = x_{(\lceil 0.25\cdot 62\rceil)} = x_{(\lceil 15.5\rceil)} = x_{(16)} = 103$
- 3-rd quartile $Q_3 = x_{0.75} = x_{(\lceil 0.75\cdot 62\rceil)} = x_{(\lceil 46.5\rceil)} = x_{(47)} = 120$
- 1-st decile (10% quantile)

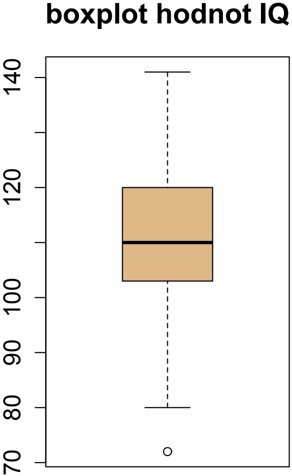
$$x_{0.1} = x_{(\lceil 0.1\cdot 62\rceil)} = x_{(\lceil 6.2\rceil)} = x_{(7)} = 92$$

- 9-th decile (90% quantile)

$$x_{0.9} = x_{(\lceil 0.9\cdot 62\rceil)} = x_{(\lceil 55.8\rceil)} = x_{(56)} = 134$$

Boxplot

- depicts quartiles, median, minimum, maximum, eventually outliers (observations further from the nearest quartile than $1.5 \cdot (Q_3 - Q_1)$)
- for our example: $Q_1 = 103, \tilde{x} = 110$, $Q_3 = 120$, 72 is an outlier



Characteristics of variability

- measures of scatter, inequality, variability of sample set.
- it should hold for a characteristic of variability s of a data set x that for arbitrary constant b and for arbitrary positive constant $a > 0$:

$$s(a \cdot x + b) = a \cdot s(x)$$

- if we add a constant b to all observations, then the characteristic does not change
- if we multiple each observation by a, then the resulting characteristic gets bigger a-times

Variance

(population) **variance** $s_x^2 = var(x)$ - mean square deviation from the mean

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2$$

- for our example:

$$s_x^2 = \frac{1}{62}\left[(107 - 111.0645)^2 + \ldots + (94 - 111.0645)^2\right] = 246.4797$$

- from our frequency table:
  $s_x^2 = \frac{1}{n}\sum_{i=1}^{M} n_i(x_i^* - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{M} n_i x_i^{*2}\right) - \overline{x}^2$
  $= (1 \cdot 75^2 + \ldots + 4 \cdot 145^2) - 111.7742^2 = 257.3361$
- it holds for variance that $s_{a \cdot x+b}^2 = a^2 s_x^2$

Standard deviation, variation coefficient

(non-sample) **standard deviation**: square root of variance

$$s_x = \sqrt{s_x^2}$$

- expressed in the same units as the data

coefficient of variation:

$$v = \frac{s_x}{\overline{x}}$$

- defined only for positive values $x_1, \ldots, x_n > 0$
- does not depend on the choice of the scale, can be used for comparison of different samples

for our data: $s_x = \sqrt{246.4797} = 15.70$

$$v = \frac{15.70}{111.0645} = 0.1414$$

range: difference of maximum and minimum of the sample

$$R = x_{(n)} - x_{(1)}$$

interquartile range: difference of the third and first quartile

$$R_M = Q_3 - Q_1 = x_{0.75} - x_{0.25}$$

mean deviation: mean absolute deviations from median (or

mean)

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}|$$

for our example: $R = 141 - 72 = 69$ $R_M = 120 - 103 = 17$

$$d = \tfrac{1}{62}(|107 - 110| + \ldots + |94 - 110|) = 12.03$$

Characteristics of shape

- measures shape of the distribution of a data set.
- it should hold for a characteristic of shape γ of a data set x that for arbitrary constant b and for arbitrary positive constant $a > 0$:

$$\gamma(a \cdot x + b) = \gamma(x)$$

- if we add a constant b to all observations or if we multiple each observation by a, then the characteristic does not change
- so for calculation we use the standardized values

$$\frac{x_i - \overline{x}}{s_x}.$$

Skewness: mean third power of standardized values

$$g_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^3$$

- measures how much the data ”leans”to one side of the mean. (symetric ≈ 0, right tail > 0, left tail < 0)

Kurtosis: mean fourth power of standardized values

$$g_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)^4 - 3$$

- measure of the ”peakedness”of the distribution (concentrated around peak and tails > 0, “flat” distribution < 0)

can be used for comparison with (verification of) normal distribution, for which $g_1 \doteq g_2 \doteq 0$.
for our data: $g_1 = 0.0159$ $g_2 = -0.241$

Example - multidimensional

- multidimensional data (more then one variable of interest)

- we find IQ score, gender, average grade in 7th class and 8th class for 62 pupils
- how to evaluate the relationship (dependence) between individual variables?
- calculate appropriate statistics (numbers) or by a plot

Example - obtained multidimensional data

Girl	1	0	0	1	0	1	0	0	1	1
Gr7	1	1	3.15	1.62	2.69	1.92	2.38	1	1.4	1.46
Gr8	1	1	3	1.73	2.09	2.09	2.55	1	1.9	1.45
IQ	107	141	105	111	112	96	103	140	136	92

Girl	1	0	0	0	1	0	1	1	1	0
Gr7	1.85	3.15	1.15	1	1.69	1.6	1.62	1.38	1.7	3.23
Gr8	1.45	3.18	1.18	1	1.91	1.72	1.63	1.36	1.9	3.36
IQ	92	72	123	140	112	127	120	106	117	92

Girl	0	0	1	1	1	1	0	1	0	1
Gr7	2.07	1.84	1.2	1.31	1.4	1.53	1.84	1	1.3	1.4
Gr8	2.45	1.9	1.36	1.45	1.73	1.6	1.54	1	1.45	1.82
IQ	107	108	117	141	109	109	106	113	112	119

Girl	0	0	1	1	0	1	0	1	0	0
Gr7	1	2.92	2.23	1.69	2.61	1.07	1.46	2.15	1.69	1.38
Gr8	1	2.82	2.45	1.54	2.54	1	1.36	1.9	1.82	1.18
IQ	138	109	80	111	86	111	120	96	103	112

vícerozměrná data - pokračování

Girl	1	1	1	0	0	1	0	1	1	0
Gr7	1.46	1.6	1.07	1.3	2.08	2	1.69	1.4	2.23	1.6
Gr8	1.54	1.63	1	1.27	1.54	2.09	1.91	1.45	2	1.81
IQ	104	103	125	101	132	113	108	106	97	121

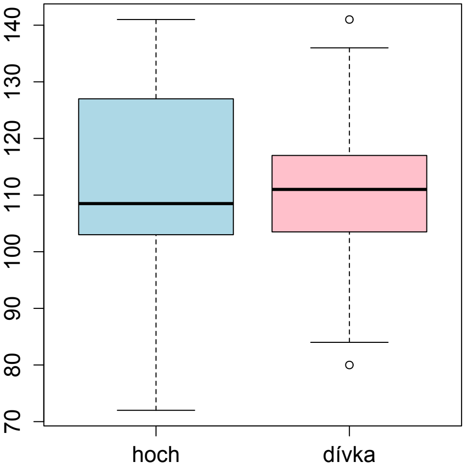
Girl	1	0	1	1	0	1	0	1	1	0
Gr7	1.07	3.13	1.84	1.8	1	1.92	2.2	1.53	1.3	1
Gr8	1.27	3.27	1.82	1.63	1	1.9	2.25	1.54	1.45	1.18
IQ	134	84	108	84	129	116	107	112	128	133

Girl	0	0
Gr7	2.85	2.61
Gr8	2.91	2.81
IQ	96	94

Graphic display of correlation

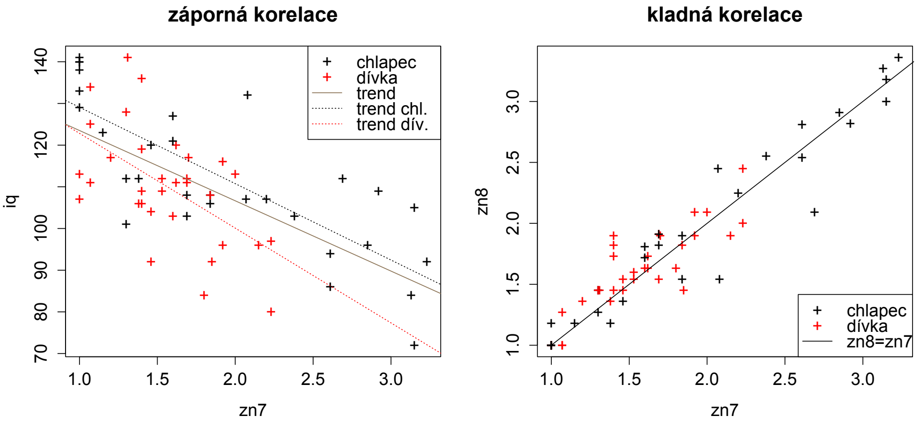
- Depends on type of the scale
- for dependence of quantitative on qualitative variable we can plot boxplot/histogram for every category of qualit. variable
- display dependence of IQ score on gender
- $\overline{x}_{boy} = 112.0$ $\overline{x}_{girl} = 110.2$

boxplot IQ zvlášť pro obě pohlaví



Graphic display of correlation - 2

Scatter plot: dependence of two quantitative variables



Correlation characteristics

two variables on every unit, i.e. we have $(x_1, y_1), \ldots, (x_n, y_n)$
covariance: measures the direction of dependence, is influenced by change of scale

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\,\overline{y},$$

- It holds: $s_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s_x^2, \quad s_{yy} = s_y^2$

(Pearson) correlation coefficient: normalized covariance, measures direction and magnitude of dependence

$$r_{x,y} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\cdot\left(\frac{y_i - \overline{y}}{s_y}\right)$$

- for var. IQ a gr7: $r_{IQ,zn7} = \frac{-6.2876}{15.6997 \cdot 0.6106} = -0.6559$

Correlation coefficient

- measures direction and the extend of linear dependence
- its value falls always into interval $\langle -1, 1\rangle$
- $r_{x,y} \approx 0$ (variables x and y mutually independent)
- $r_{x,y}$ close to 1 (positive dependence: increasing linear relationship of x and y)
- $r_{x,y}$ close to -1 (negative dependence: decreasing linear relationship of x and y)

For our data set we can calculate the correlation for every pair of variables girl, iq, gr7, gr8: so called **correlation matrix**

	girl	iq	gr7	gr8
girl	1.0000	-0.0597	-0.3054	-0.2661
iq	-0.0597	1.0000	-0.6559	-0.6236
gr7	-0.3054	-0.6559	1.0000	0.9481
gr8	-0.2661	-0.6236	0.9481	1.0000

Probability theory

- deals with **experiment**, whose possible results are called outcomes.

- a set of all possible outcomes is a sample space Ω
- elements of Ω are denoted by ω_i and are called **elementary events**
- **Event** (denote A, B, etc.) - is a subset of Ω (a set of outcomes of an experiment) , can be represented by a statement about the result of the experiment

Probability of an event A (denoted $P(A)$): expresses measure of expectation, that event A occurs.

- for large number of repetitions of the experiment the relative frequency of event A goes to $P(A)$.

Classical probability

- the set of all possible events Ω consists of finite number (n) of elementary events $\omega_1, \ldots, \omega_n$
- if elementary events are assigned equal probabilities
- $m(A)$ denotes the number of elementary events, that form the event (are favorable to the event) A

Then

$$P(A) = \frac{m(A)}{n} = \frac{\text{number of favorable elem. events}}{\text{number of all elem. events}}$$

Example: throwing dice

- we throw an honest die with numbers $1, 2, \ldots, 6$
- event A - die falls on six
- event B - die falls on some odd number
- all 6 events that can occur are equally probable
- we count $m(A) = 1$ a $m(B) = 3$

Thus

$$P(A) = \frac{m(A)}{n} = \frac{1}{6}$$

and

$$P(B) = \frac{m(B)}{n} = \frac{3}{6} = \frac{1}{2}$$

Example (permutation)

What is the probability that if we randomly rearrange letters P, A, V, E, L we get the word PAVEL?

- **factorial**: $n! = 1 \cdot 2 \cdot \ldots \cdot n$ number of ways to arrange n different items in a row - number of **permutations**
- number of ways to rearrange the letters is $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$, each is equally probable
- only one of them is favorable
- thus $P = \frac{1}{5!} = \frac{1}{120}$

What is the probability for MISSISSIPPI?

- we speak about **multiset permutation** (some elements appear multiple times), number of rearangements is $\frac{11!}{4! \cdot 4! \cdot 2!}$, out of which only one is favorable
- thus $P = \frac{1}{\frac{11!}{4! \cdot 4! \cdot 2!}} = \frac{4! \cdot 4! \cdot 2!}{11!} = \frac{24 \cdot 24 \cdot 2}{39916800} = 0.000029$

Example (combination)

There are 12 boys and 16 girls in the classroom. I choose randomly three pupils. What is the probability that I choose one boy and two girls?

- **binomial coefficient**: $\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \frac{n \cdot (n-1) \cdots (n-k+1)}{1 \cdot 2 \cdots k}$ is number of ways how to choose k elements out of n different elements (the order does not matter) - **combination** of k elements from a set of n elements.
- number of all possible ways, i.e. triples, which can be chosen, is $\binom{28}{3} = \frac{28!}{3! \cdot 25!} = \frac{28 \cdot 27 \cdot 26}{1 \cdot 2 \cdot 3} = 3276$, all of which are equally probable.
- number of favorable ways, i.e. triples with just one boy: $\binom{12}{1} \cdot \binom{16}{2} = 12 \cdot 120 = 1440$: every way how to choose 1 boy out of 12 can be combined with every way how to choose 2 girls out of 16.
- thus $P = \frac{\binom{12}{1} \cdot \binom{16}{2}}{\binom{28}{3}} = \frac{40}{91} \doteq 0.44$

Example (k-permutations of n)

From the digits $1, 2, 3, 4, 5$ we randomly form a three digit number. Every digit can be used only once. What is the probability that such a number is smaller than 200?

- number of three digit numbers $5 \cdot 4 \cdot 3 = 60$ - number of permutations of 5 elements taken 3 at a time (the order does matter), each is equally probable.
- number of favorable cases, i.e. numbers starting with 1 is: $1 \cdot 4 \cdot 3 = 12$
- thus $P = \frac{1 \cdot 4 \cdot 3}{5 \cdot 4 \cdot 3} = \frac{1}{5}$

What if every digit could be used multiple times?

- number of all three digit numbers $5 \cdot 5 \cdot 5 = 5^3 = 125$ - number of permutations with repetition of 5 elements taken 3 at a time (the order does matter).
- number of favorable cases, i.e. numbers starting with 1 is: $1 \cdot 5 \cdot 5 = 25$
- thus $P = \frac{1 \cdot 5 \cdot 5}{5 \cdot 5 \cdot 5} = \frac{1}{5}$

Notions and rules 1

- for every random event $A \subset \Omega$ holds: $0 \leq P(A) \leq 1$
- $\emptyset$ - **impossible event**, never occurs: $P(\emptyset) = 0$
- Ω - **certain event**, occurs at each realization of the experiment: $P(\Omega) = 1$
- $\overline{A}$ - **complement** of an event A, is the event that A does not occur. It holds $P(\overline{A}) = 1 - P(A)$
- $A \subset B$ - **event** A **is a subset of event** B, i.e. whenever A occurs, then B occurs as well: $P(A) \leq P(B)$ and $P(B - A) = P(B) - P(A)$.
- $A \cup B$ - **union of events** A **and** B, i.e. event that occurs if and only if at least one of A and B occurs.
- $A \cap B$ - **intersection of events** A **and** B, i.e. event that occurs if and only if both A and B occur at the same time. It holds: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Notions and rules 2

- Events A and B are called **disjoint**, if $A \cap B = \emptyset$, i.e. events A and B cannot occur both at the same time.
- $A_1, \dots A_n \subset \Omega$; $A_i \cap A_j = \emptyset$ for every $i \neq j$, then $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$
- For any random events $A_1, A_2, \cdots, A_n$ it holds (inclusion–exclusion principle)

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(A_i \cap A_j) +$$

$$+ \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} P(A_i \cap A_j \cap A_k)$$

$$- \cdots (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right).$$

- We say that $A_1, \dots, A_n$ form a **partition of sample space** Ω, if the events $A_1, \dots, A_n$ are disjoint (i.e. $A_i \cap A_j = \emptyset$, $i, j = 1, \dots, n$, $i \neq j$) and $\bigcup_{i=1}^{n} A_i = \Omega$.

Example 1

(prob. of complementary event): From digits $1, 2, 3, 4, 5$ we randomly form a three digit number. What is the prob., that any digit is repeated? (event A)?

$$P(A) = P(\text{“any digit is repeated”}) \quad \text{not easy to find}$$

- we can calculate

$$P(\overline{A}) = P(\text{“no digit is repeated”}) = \frac{\text{number of favorable}}{\text{number of all}} = \frac{5 \cdot 4 \cdot 3}{5^3}$$

- thus

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{12}{25} = \frac{13}{25}$$

Example 2

(union of two not disjoint events): We choose randomly one number from 1 to 100. What is the prob., that it is divisible by two (event A) or by three (event B)?

thus $P(A \cup B) = ?$

$$P(A) = \frac{m(A)}{n} = \frac{50}{100} = 0.5$$

$$P(B) = \frac{m(B)}{n} = \frac{33}{100} = 0.33$$

- events A and B are not disjoint: $P(A \cap B) = \frac{16}{100} = 0.16$
- thus

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.33 - 0.16 = 0.67$$

Example 3

(union of more not disjoint events, inclusion–exclusion principle): Absent-minded secretary puts the letters into the envelopes at random. What is the probability that at least one letter is put into its correct envelop?

Let A_i be the event that letter i is placed in the correct envelope. Thus $P(A_1 \cup A_2 \cup A_3) = ?$

$$P(A_1) = P(A_2) = P(A_3) = \frac{2!}{3!} = \frac{1}{3}$$

$$P(A_i \cap A_j) = \frac{1}{3!} = \frac{1}{6} \quad \forall i, j \quad i \neq j \qquad \text{and} \quad P(A_1 \cap A_2 \cap A_3) = \frac{1}{6}$$

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - \\ &- P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3) = \\ &= \frac{1}{3} + \frac{1}{3} + \frac{1}{3} - \frac{1}{6} - \frac{1}{6} - \frac{1}{6} + \frac{1}{6} = \frac{2}{3} \end{aligned}$$

Geometric probability

- generalization of classical probability for Ω uncountable

$$P(A) = \frac{n - \text{dimensional volume}(A)}{n - \text{dimensional volume}(\Omega)}$$

Ex. (meeting probability): Two people (A and B) are to arrive at a certain location at some randomly chosen time between 1:00 PM and 2:00 PM, and both A and B will wait 10 min. before leaving. Assume independent and random arrival times. What is the prob., that they meet each other (event A)?

- Ω can be pictured as a part of a plan 60×60 (in minutes)
- from picture

$$P(A) = \frac{\text{area corresponding to meeting}}{60 \cdot 60} = \frac{3600 - 2500}{3600} = \frac{11}{36}$$

Conditional probability

Let A, B are events such that $P(B) > 0$. **Conditional probability** of event A given that the event B has occurred is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- We restrict Ω only to B. We compute it as the proportion of B that is also part of A.

Ex.: We throw a die. What is the probability that the die falls on three (event A), given that an odd number was obtained (event B)?

- from the definition and because $A \subset B$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}$$

Independent events

Independent events: occurrence of one event does not change the probability of the other event , or

$$\frac{P(A \cap B)}{P(B)} = P(A|B) = P(A)$$

and similarly for $P(B|A)$.
Thus we say that, events A and B are **independent**, if

$$P(A \cap B) = P(A) \cdot P(B)$$

Ex.: Two dice are rolled. What is the probability that the first die falls on six (event A) and at the same time the second one falls on six (event B)? Are the events A and B independent?

- from the classical probability (number of all elementary events is 36):

$$\frac{1}{36} = P(A \cap B) \overset{?}{=} \; = P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{6}$$

so the events are independent.

Example

(independence): two dice are rolled.
Event A means that at least one die falls on two.
Event B means that the sum of the obtained values is eight.
Are the events A and B independent?

- from the classical probability (number of all elementary events is 36):

$$0.0556 = \frac{2}{36} = P(A \cap B) \overset{?}{=} \neq P(A) \cdot P(B) = \frac{11}{36} \cdot \frac{5}{36} = 0.0424$$

so the events are not independent.

Law of total probability

Let $D_1, D_2, \ldots, D_n$ form a partition of the sample space Ω, then for any event A

$$P(A) = \sum_{i=1}^{n} P(A|D_i) \cdot P(D_i)$$

Proof:

$P(A) = P(A \cap \Omega) = P(A \cap \bigcup\limits_{i=1}^{n} D_i) = \sum_{i=1}^{n} P(A \cap D_i) = \sum_{i=1}^{n} \frac{P(A \cap D_i)}{P(D_i)} \cdot P(D_i) = \sum_{i=1}^{n} P(A|D_i) \cdot P(D_i)$

Example

(Law of total probability): There are three bags with bonbons. In the first bag there are 10 bonbons out of which 4 are chocolate, in the second bag 1 out of 8 is chocolate and in the third one 2 out of 6 are chocolate. From one bag (randomly chosen) we draw one bonbon. What is the probability that the bonbon will be chocolate (event A)?

- denote D_i event that we draw from the ith bag.
- $P(A) = P(A|D_1)\cdot P(D_1) + P(A|D_2)\cdot P(D_2) + P(A|D_3)\cdot P(D_3) = \frac{4}{10}\cdot\frac{1}{3} + \frac{1}{8}\cdot\frac{1}{3} + \frac{2}{6}\cdot\frac{1}{3} = \frac{103}{360} = 0.286$

Bayes' theorem

Let $D_1, D_2, \ldots, D_n$ form a partition of the sample space Ω, then for any event A such that $P(A) > 0$, it holds

$$P(D_i|A) = \frac{P(A|D_i) \cdot P(D_i)}{\sum_{j=1}^{n} P(A|D_j) \cdot P(D_j)}$$

Proof:

$$\begin{aligned} P(D_i|A) &= \frac{P(A \cap D_i) \cdot P(D_i)}{P(A) \cdot P(D_i)} = \\ &= \frac{P(A|D_i) \cdot P(D_i)}{P(A)} \overset{\text{LTP}}{=} \frac{P(A|D_i) \cdot P(D_i)}{\sum_{j=1}^{n} P(A|D_j) \cdot P(D_j)} \end{aligned}$$

example

(Bayes' theorem): Suppose that only 1% of population suffers from a certain disease. There is a medical test to detect the disease with the following reliability: If a person has the disease, there is a probability of 0.8 that the test will give a positive response; whereas, if a person does not have the disease, there is a probability of 0.9 that the test will give a negative response. If a person have a positive response to the test, what is the probability that the person have the disease?

- denote *DIS* event that the person is diseased
- *HEA*: event that the person is healthy
- *POS*: event that the response to the test is positive
- *NEG*: event that the response to the test is negative

$$P(DIS|POS) \overset{\text{BT}}{=} \frac{P(POS|DIS)\cdot P(DIS)}{P(POS|DIS)\cdot P(DIS) + P(POS|HEA)\cdot P(HEA)} =$$

$$= \frac{0.8\cdot 0.01}{0.8\cdot 0.01 + 0.1\cdot 0.99} \doteq 0.075$$

Random variable

- use of events is not always sufficient
- often the result of an experiment is a number
- e.g. number of sixes in ten tosses with a die, lifetime of a light bulb

Random variable: numerical expression of the result of an experiment (real-valued function on sample space Ω)
distribution of random variable: determines probabilities associated with the possible values of random variable (a set function: assigns a probability to every subset of R)

- distribution is uniquely determined e.g. by (cumul.) distr. f.
- **(Cumulative) distribution function** $F_X(x)$ of a random variable X determines for every x probability, that the rand. var. X is smaller than x:

  $$F_X(x) = P(X < x) \qquad x \in R$$

  cumulative probability (theoretic counterpart of the cumul. relative frequency calculated for every point of R)

Types of distribution

properties of c.d.f. $F_X(x)$:

- nondecreasing, continuous from the left
- $\lim_{x\to-\infty} F_X(x) = 0, \qquad \lim_{x\to\infty} F_X(x) = 1$

Discrete distribution ($F_X(x)$ “step-function”): X is a discrete r.v., if X can take only a sequence of different values $x_1, x_2, \ldots$ with probabilities $P(X = x_1), P(X = x_2), \ldots$ (probability (mass) function) satisfying $\sum_i P(X = x_i) = 1$.

Continuous distribution ($F_X(x)$ continuous): X is a continuous r.v., if there exists a probability **density** function $f_X(x)$, for which

$$F_X(x) = P(X < x) = \int_{-\infty}^{x} f_X(t)\,dt$$

- $f_X(x) = F'_X(x)$ at every continuity point of $f_X(x)$
- $f_X(x) \geq 0\ \forall x, \quad P(a < X < b) = \int_a^b f_X(x)\,dx, \quad \int_{-\infty}^{\infty} f_X(x)\,dx = 1$
- $P(X = a) = 0$ for every $a \in R$ (theoretic counterpart of the boundary of a histogram when lengths of intervals goes to zero)

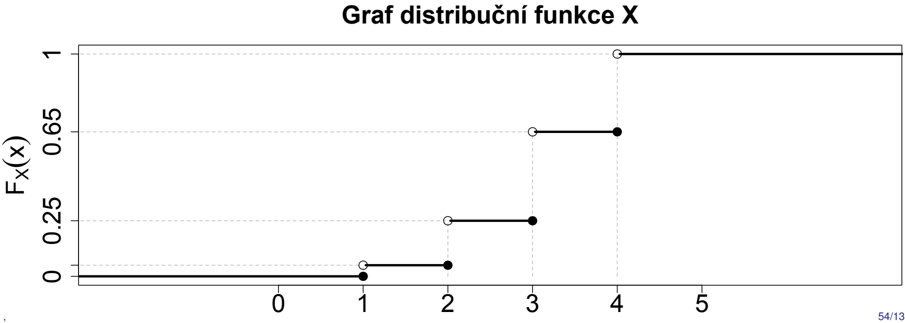
Example 1

(discrete distribution): It is known that the distribution of grades from a certain course for a random student (X) is the following:

x_i	1	2	3	4
$P(X = x_i)$	0,05	0,2	0,4	0,35

Find $P(X < 3)$ and cum. distribution function of r.v. X.

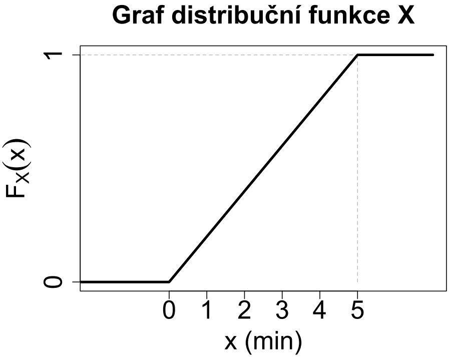
- $F_X(3) = P(X < 3) = P(X = 1) + P(X = 2) = 0{,}05 + 0{,}2 = 0{,}25$
- need to find $F_X(x) = P(X < x)$ for every $x \in R$

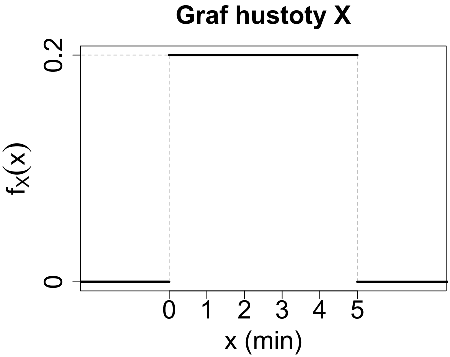


Example 2

(continuous distribution): Tram leaves regularly in five minute intervals. Assume that we come to the tram stop at a random time. What is the distribution of the r.v. X denoting our waiting time? ▸ to uniform distribution

- it is enough to find the cum. distribution f. $F_X(x)$ or the density f. $f_X(x)$ for every $x \in R$
- clearly for $x \in (0,5)$ it holds $F_X(x) = P(X < x) = \frac{x}{5}$, so $f_X(x) = \frac{1}{5}$





Determining probabilities 1

for ▸ Ex. 1 (discrete distribution): Find the probability, that the student's grade is

- less than 4 but not less than 2:

$$P(2 \leq X < 4) \overset{\text{distr. f.}}{=\!=\!=} P(X < 4) - P(X < 2) = F_X(4) - F_X(2) = 0{,}65 - 0{,}05 = 0{,}6$$

$$\overset{\text{from prob. mass f.}}{=\!=\!=\!=\!=} P(X = 3) + P(X = 2) = 0{,}4 + 0{,}2 = 0{,}6$$

- not less than 3:

$$P(X \geq 3) \overset{\text{from distr. f.}}{=\!=\!=\!=} 1 - P(X < 3) = 1 - F_X(3) = 1 - 0{,}25 = 0{,}75$$

$$\overset{\text{from prob. mass f.}}{=\!=\!=\!=\!=} P(X = 3) + P(X = 4) = 0{,}4 + 0{,}35 = 0{,}75$$

- equal to 4:

$$P(X = 4) \overset{\text{from prob. mass f.}}{=\!=\!=\!=\!=} 0{,}35 \qquad \text{height of the step of distr. f. at 4}$$

Determining probabilities 2

for ▸ Ex. 2 (continuous distribution): Find the probability, that we will wait

- less than 4 but more than 2 minutes:

$$P(2 < X < 4) \overset{P(X=2)=0}{=\!=\!=} P(X < 4) - P(X < 2) \overset{\text{distr. f.}}{=\!=\!=} F_X(4) - F_X(2) = \frac{4}{5} - \frac{2}{5} = \frac{2}{5}$$

$$\overset{\text{from density}}{=\!=\!=} \int_2^4 f_X(x)\,dx = \int_2^4 \frac{1}{5}\,dx = \frac{2}{5}$$

- longer than 4 minutes:

$$P(X > 4) \overset{\text{from distr. f.}}{=\!=\!=} 1 - P(X < 4) = 1 - F_X(4) = 1 - \frac{4}{5} = \frac{1}{5}$$

$$\overset{\text{from density}}{=\!=\!=} \int_4^\infty f_X(x)\,dx = \int_4^5 \frac{1}{5}\,dx + \int_5^\infty 0\,dx = \frac{1}{5}$$

- exactly 4 minutes:

$$P(X = 4) = \int_4^4 \frac{1}{5}\,dx = 0 \qquad \text{height of the step of distr. f. at 4 is equal 0}$$

Expectation

Expectation (expected value) of random variable X - value, around which the possible values of X cumulate

- for discrete distr.: weighted mean of possible values, weights are the probabilities

$$EX = \sum_i x_i \cdot P(X = x_i) = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \dots$$

u ▸ Ex. 1: $EX = 1 \cdot 0{,}05 + 2 \cdot 0{,}2 + 3 \cdot 0{,}4 + 4 \cdot 0{,}35 = 3{,}05$ (mean, expected grade)

- for continuous distr.: integral over possible values x, weighting function is the density

$$EX = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

u ▸ Ex. 2: $EX = \int_{-\infty}^{0} x \cdot 0\, dx + \int_0^5 x \cdot \frac{1}{5}\, dx + \int_5^{\infty} x \cdot 0\, dx = \frac{5}{2}$ (mean, expected waiting time)

Expectation of a function $Y = g(X)$ of a random variabe X - value, around which values of r.v. $g(X)$ cumulate

- for discrete distr.: weighted sum of the function values

$$Eg(X) = \sum_i g(x_i) \cdot P(X = x_i) = g(x_1) \cdot P(X = x_1) + g(x_2) \cdot P(X = x_2) + \dots$$

- for continuous distr.: integral over possible values $g(x)$, weighting function is the density

$$Eg(X) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx$$

for ▸ Ex. 1: suppose, we are not interested in expected grade, but expected tuition fee, that is derived from the grade by a relation $g(x) = 1000 \cdot x^2$ Kč

$Eg(X) = 1000 \cdot 1^2 \cdot 0{,}05 + 1000 \cdot 2^2 \cdot 0{,}2 + 1000 \cdot 3^2 \cdot 0{,}4 + 1000 \cdot 4^2 \cdot 0{,}35 = 10\,050$ Kč

Variance

Variance of rand. var. X: $var\,X = E(X - EX)^2$ - gives variability of the distribution of X around its expectation, it is the expected value of the squared deviation from the mean

- for discrete distr.:

$$var\,X = E(X - EX)^2 = \sum_i (x_i - EX)^2 \cdot P(X = x_i) =$$

$$= (x_1 - EX)^2 \cdot P(X = x_1) + (x_2 - EX)^2 \cdot P(X = x_2) + \dots$$

for ▸ Ex. 1:
$var\,X = 2{,}05^2 \cdot 0{,}05 + 1{,}05^2 \cdot 0{,}2 + 0{,}05^2 \cdot 0{,}4 + 0{,}95^2 \cdot 0{,}35 = 0{,}7475$

- for continuous distr.:

$$var\,X = E(X - EX)^2 = \int_{-\infty}^{\infty} (x - EX)^2 \cdot f_X(x)\,dx$$

for ▸ Ex. 2:
$var\,X = \int_{-\infty}^{0} (x - \frac{5}{2})^2 \cdot 0\,dx + \int_0^5 (x - \frac{5}{2})^2 \cdot \frac{1}{5}\,dx + \int_5^{\infty} (x - \frac{5}{2})^2 \cdot 0\,dx \doteq 2{,}083$

$\sqrt{var\,X}$ is called **standard deviation** of rand. var X

Independent random variables

Alike for events we can speak about independence of random variables. Independence means that knowing value of one r. v. does not effect the probability distribution of the second r. v.

We say that rand. variables X and Y are **independent** if for every $x, y \in R$

$$P(X < x, Y < y) = P(X < x) \cdot P(Y < y)$$

specially for discrete distr. it can be replaced by the condition that for every i, j

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$$

Properties of expectation and variance

Let $a, b \in R$ and X is a random var., then

1) $E(a+b\cdot X)=a+b\cdot EX$
2) $var\,(a+b\cdot X)=b^2\cdot var\,X$
3) $var\,X\geq 0$
4) $var\,X=EX^2-(EX)^2$
5) $E(X+Y)=EX+EY$
6) for independent X, Y: $var\,(X+Y)=var\,X+var\,Y$

proof: 1), 2), 4) and 5) follows from linearity of sum or integral:

ad 1) e.g. for continuous distribution:

$$E(a+b\cdot X)=\int_{-\infty}^{\infty}(a+b\cdot x)\cdot f_X(x)\,dx \overset{\text{lin. int.}}{=\!=\!=}$$
$$=a\cdot\int_{-\infty}^{\infty}f_X(x)\,dx+b\cdot\int_{-\infty}^{\infty}x\cdot f_X(x)\,dx=a+b\cdot EX$$

ad 2):

$$var(a+b\cdot X)=E[a+b\cdot X-E(a+b\cdot X)]^2\overset{1)}{=}E[a+b\cdot X-(a+b\cdot EX)]^2=$$
$$=E[b\cdot(X-EX)]^2=b^2\cdot var\,X$$

ad 3): foll. from fact that $var\,X$ is integral (sum) of nonneg. funct. (values)
ad 4): similar as 1) and 2) (homework). ad 5) and 6): w/o proof

Quantile function

Let F_X is the distribution function of random variable X. Then the function F_X^{-1} given by the relation

$$F_X^{-1}(\alpha) = \inf\{x \in R \,:\, F_X(x) \geq \alpha\} \qquad 0 < \alpha < 1,$$

is called **quantile function**

Infimum of a set A, inf *A: is the maximum from those elements, that are smaller or equal to all the elements of A.*

- Value of the function $F_X^{-1}(\alpha)$ is called α**-quantile** (or $100 \cdot \alpha\,\%$ quantile)
- for continuous distr. it is the inverse of F_X. It holds

  $$P(X < F_X^{-1}(\alpha)) = \alpha$$

  α-quantile is such value that the rand. var. is smaller that this value with probability α
- specially $F_X^{-1}(0{,}5)$ is called **median** of a distribution.

u ▸ Ex. 1: $F_X^{-1}(0{,}5) = \inf\{x \,:\, F_X(x) \geq 0{,}5\} \xlongequal{\text{from graph } F_X} 3$

u ▸ Ex. 2: $F_X^{-1}(0{,}5) = \inf\{x \,:\, F_X(x) \geq 0{,}5\} \xlongequal{\text{from inv. function of } F_X} 5 \cdot 0{,}5 = 2{,}5$

with probab. 50 % I will wait less then 2,5 minutes

Bernoulli distribution

Example: only one out of the four answers a), b), c), d) to a question is correct. What is the probability of the correct answer for random guessing?

Let $X = 1$ (or 0), if we answer correctly (or incorrectly)

$$P(X = 1) = 1/4, \quad P(X = 0) = 3/4$$

X is a r.v. with Bernoulli distribution with parameter $p = 1/4$

Generally:

X has **Bernoulli distribution** with param. p if

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p, \qquad 0 < p < 1$$

- expectation $\quad EX = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = p$
- variance
$$\begin{aligned} var\,X &= EX^2 - (EX)^2 = 1^2 \cdot P(X = 1) + 0^2 \cdot P(X = 0) - p^2 = \\ &= p - p^2 = p \cdot (1 - p) \end{aligned}$$

in Ex.: $\quad EX = \dfrac{1}{4} \qquad var\,X = \dfrac{1}{4} \cdot \dfrac{3}{4} = \dfrac{3}{16}$

Example

(binomial distribution): In a test there are 5 questions, with only one correct answer out of a), b), c), d). What is the probability of getting exactly 3 correct answers for random guessing?

- set X number of the correct answers
- for each the probab. of correct answer is $p = 1/4$
- answers to the questions are independent
- i.e. probability that we succeed in three (e.g. the first three) questions and fail in the remaining (denote by 11100), is $p^3 \cdot (1-p)^2$
- we can succeed in different three answers: number of ways to choose three questions from five $\binom{5}{3} = 10$

$$10\times\left\{\begin{array}{l}11100\\11010\\10110\\01110\\11001\\10101\\01101\\10011\\01011\\00111\end{array}\right.$$

Probability of answering exactly three question correctly is

$$P(X=3) = \binom{5}{3}\cdot p^3\cdot(1-p)^2 = 10\cdot(1/4)^3\cdot(3/4)^2 = 0{,}088$$

Binomial distribution

We conduct n independent trials. We are interested in X number of occurrences of a certain event in these n trials. Probability of occurrence of that event is equal for every trial, equal to p. X can only take values $0, 1, \ldots, n$ with probability mass function

$$P(X = i) = \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i}, \quad i = 0, 1, \ldots, n; \quad \text{where } 0 < p < 1$$

- we say that X has **binomial distribution** with parameters n and p
- for short $X \sim Bi(n, p)$
- can be understood as a sum of n independent Bernoulli trials
- expectation $EX = \sum_{i=0}^{n} i \cdot \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} = n \cdot p$
- variance

$$var\, X = EX^2 - (EX)^2 = \sum_{i=0}^{n} i^2 \cdot \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} - (n \cdot p)^2 = n \cdot p \cdot (1-p)$$

in Ex.: $X \sim Bi(5, 1/4)$ $EX = \frac{5}{4}$ $var\, X = 5 \cdot \frac{1}{4} \cdot \frac{3}{4} = \frac{15}{16}$

Example

(geometric distribution): only one out of the four answers a), b), c), d) to every question is correct. Consecutively we answer the questions by random guessing until the first correct answer. What is the probability that the first correctly answered question will be the third one.

- set X number of incorrectly answered question before the first success
- every question is correctly answer with prob. $p = 1/4$
- answers to different questions are independent
- we must fail in the first and second and succeed in the third question

So the probability is

$$(1-p)^2 \cdot p = (3/4)^2 \cdot (1/4) \doteq 0{,}14$$

Geometric distribution

We conduct independent trials until a certain event occurs. We are interested in X number of trials before the first occurrence of that event. Probability of occurrence of that event is equal for every trial, equal to p. X can only take values $0, 1, \ldots$ with probability mass function

$$P(X = i) = (1 - p)^i \cdot p, \quad i = 0, 1, \ldots$$

where $0 < p < 1$

- we say that X has **geometric distribution** with parameter p
- we write $X \sim Ge(p)$
- expectation $EX = \sum_{i=0}^{\infty} i \cdot (1 - p)^i \cdot p = \frac{1-p}{p}$
- variance

$$var\,X = EX^2 - (EX)^2 = \sum_{i=0}^{\infty} i^2 \cdot (1 - p)^i \cdot p - \left(\frac{1 - p}{p}\right)^2 = \frac{1 - p}{p^2}$$

in Ex.: $X \sim Ge(1/4)$ $EX = 3$ $var\,X = \frac{3}{4}/(\frac{1}{4})^2 = 12$

Example

(hypergeometric distribution): In a pot there are 30 sweet dumplings, out of which 10 are with strawberry and 20 with plum inside. We draw 6 dumplings. What is the prob. that less that two of them are straberry?

- set X number of strawberry dumplings among the six
- we draw “without replacement”, i.e. the draws are not independent
- want to find $P(X < 2) = P(X = 0) + P(X = 1)$
- $P(X = 0)$ or $P(X = 1)$ follows from classical definition

$$P(X = 0) = \frac{\binom{10}{0} \cdot \binom{20}{6}}{\binom{30}{6}} \doteq 0{,}065 \quad \text{resp. } P(X = 1) = \frac{\binom{10}{1} \cdot \binom{20}{5}}{\binom{30}{6}} \doteq 0{,}261$$

the result is

$$P(X < 2) \doteq 0{,}065 + 0{,}261 = 0{,}326$$

Hypergeometric distribution

We have a set of N objects, out of which M have a certain property. We draw n objects. Let X denote number of drawn objects with the property. X can only take integer values with probabilities

$$P(X=i)=\frac{\binom{M}{i}\cdot\binom{N-M}{n-i}}{\binom{N}{n}},\quad \text{pro} \quad \max(0,M+n-N)\leq i\leq\min(M,n)$$

- we say that X has **hypergeometric distribution** with parameters N, M and n
- we write $X\sim Hg(N,M,n)$
- expectation $EX=\sum_i i\cdot\frac{\binom{M}{i}\cdot\binom{N-M}{n-i}}{\binom{N}{n}}=\frac{n\cdot M}{N}$
- variance $var\,X=\frac{n\cdot M\cdot(N-M)}{N^2}\cdot\left(1-\frac{n-1}{N-1}\right)$

in Ex.:

$X\sim Hg(N=30,M=10,n=6)$ $\quad EX=\frac{6\cdot 10}{30}=2$ $\quad var\,X\doteq 1{,}103$

Poisson distribution

Let X be a random variable that can take only values $i = 0, 1, 2, \ldots$ with probabilities

$$P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \qquad i = 0, 1, 2, \ldots$$

where $\lambda > 0$ is a given number.

- we say that X has **Poisson distribution** with parameter λ
- we write $X \sim Po(\lambda)$
- expectation and variance $\qquad EX = var\, X = \lambda$

Let $Y_n \sim Bi(n, p)$, where n is large and p is small such that $n \cdot p = \lambda$.
Then $\lim_{n\to\infty} P(Y_n = i) = P(X = i)$.
i.e. for n large and p small the distribution $Bi(n, p)$ can be approximated by distribution $Po(n \cdot p)$
e.g. for $\quad Y \sim Bi(20, 0{,}1) \quad$ and $\quad X \sim Po(20 \cdot 0{,}1) = Po(2)$
is $\qquad P(Y = 3) \doteq 0.19$ and $\quad P(X = 3) \doteq 0.18$

► Most often The Poisson distribution is used to model the number of events occurring within a given time interval if the events are arriving independently with an intensity λ (number of telephone calls, car accidents, customers arriving at a counter etc.)

Example

(Poisson distribution): On average there are 30 calls to a call center during one hour. What is the probability that more that one call arrives during one minute?

- let X be the number of incoming calls during 1 min.
- X is a number of events occurring within a given time interval, so $X \sim Po(\lambda)$
- λ is not known, but $EX = \lambda$
- mean number of calls per 1 minute $EX = \lambda$ can be estimated by $\frac{30}{60} = 0{,}5$
- so we set $\lambda = 0{,}5$ and calculate

$$P(X > 1) = 1 - [P(X = 0) + P(X = 1)] =$$

$$= 1 - \left[\frac{0{,}5^0}{0!}e^{-0{,}5} + \frac{0{,}5^1}{1!}e^{-0{,}5}\right] \doteq 1 - 0{,}606 - 0{,}303 \doteq 0{,}09$$

Uniform distribution

In ▸ example we dealt with uniform distr. on interval $(0, 5)$

Let X is a random variable with continuous distr. with density

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{for } x \leq a \text{ or } x \geq b. \end{cases}$$

and distribution function

$$F_X(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b. \end{cases}$$

- we say that X has **uniform distribution** on interval (a, b)
- we $X \sim U(a, b)$
- expectation and variance (homework)

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

Example: rounding error when rounding number to the nearest integer

Exponential distribution

Let X is a random variable with contin. distr. and density

$$f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & x \geq 0 \\ 0 & otherwise, \end{cases}$$

and distribution function

$$F_X(x) = \int_{-\infty}^{x} f(t)\, dt = \begin{cases} 1 - e^{-\lambda \cdot x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

where $\lambda > 0$ is a given number

- we say that X has **exponential distribution** with parameter λ
- we write $X \sim Exp(\lambda)$
- expectation $EX = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx = \int_0^{\infty} x \cdot \lambda \cdot e^{-\lambda \cdot x}\, dx \overset{\text{p. p.}}{=\!=} \frac{1}{\lambda}$
- variance $var\, X = EX^2 - (EX)^2 = \int_0^{\infty} x^2 \cdot \lambda \cdot e^{-\lambda \cdot x}\, dx - \left(\frac{1}{\lambda}\right)^2 \overset{2\times\text{ p. p.}}{=\!=\!=} \frac{1}{\lambda^2}$

► is a continuous analog of geometric distribution. It is used to describe the waiting time or time between events if they occur continuously and independently at a constant average rate (time before the next telephone call, customer arrival, time to failure etc.)

Example

(exponential distribution): Average lifetime of a certain component is 14 years and can be modeled as an exponential distribution r. v. Find

a) probab. that it breaks down in the first year after the two-year warranty

b) what maximal warranty period can the seller provide, so that not more then 20% of the sold components breaks down during the period

- set X the lifetime of the component, $X \sim Exp(\lambda)$
- λ is not known, but $EX = 1/\lambda$
- expected lifetime $EX = 1/\lambda$ can be estimated by 14
- so we set $\lambda = \frac{1}{14}$ and calculate

a)
$$P(X \in (2,3)) = \int_2^3 f_X(x)\,dx = \int_2^3 \frac{1}{14} \cdot e^{-\frac{x}{14}}\,dx = e^{-\frac{2}{14}} - e^{-\frac{3}{14}}$$

or $= P(X < 3) - P(X < 2) = F_X(3) - F_X(2) = 1 - e^{-\frac{3}{14}} - \left(1 - e^{-\frac{2}{14}}\right) \doteq 0{,}06$

b) want to find the period p such that $P(X < p) = 0{,}2$

- so $p = F_X^{-1}(0{,}2)$ (20% quantile of the distribution $Exp(\lambda)$)
- $F_X^{-1}(u)$ is the inverse function to $F_X(x)$: $F_X^{-1}(u) = -\frac{1}{\lambda} \cdot \ln(1-u)$

the warranty period $p = -14 \cdot \ln(0{,}8) \doteq 3{,}12 \doteq 3$ years and 1,5 month

Normal (Gaussian) distribution

Let X is continuous random variable with prob. density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad \text{pro}\, x \in R.$$

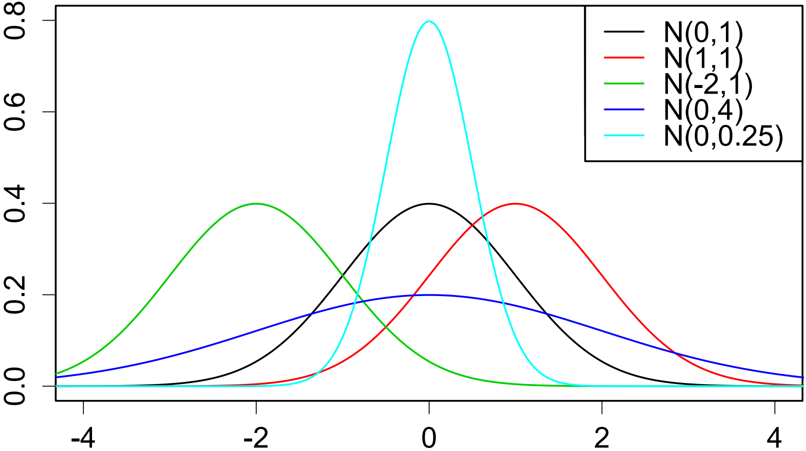
where $\mu = EX$ and $\sigma^2 = var\, X$ are parameters of the distribution.

- then X has **normal distribution** with expectation μ and variance σ^2
- shortly $X \sim N(\mu, \sigma^2)$
- distribution function $F_X(x) = \int_{-\infty}^{x} f(t)\, dt$ cannot be integrated in closed form
- for $N(0, 1)$ there are tables $F_X(x)$
- most important contin. distribution

► Origin: sum of many small independent pieces

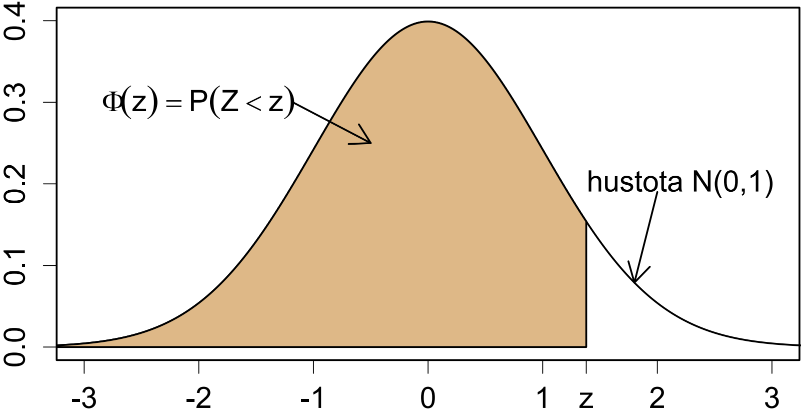
Plots of densities of normal distributrion $N(\mu, \sigma^2)$

- symmetric around expectation



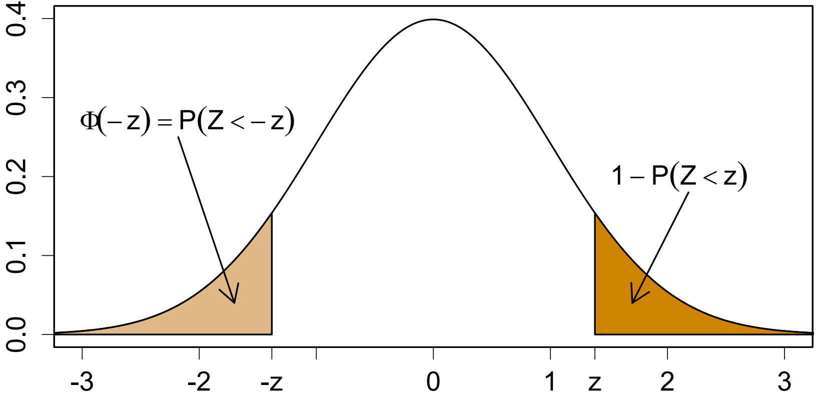
Standard normal distribution $Z \sim N(0,1)$

- distrib. function $N(0,1)$ is den. $\Phi(z) = P(Z < z)$
- e.g. $\Phi(1{,}38) = P(Z < 1{,}38) \overset{\text{from tables}}{=\!=\!=} 0{,}916$



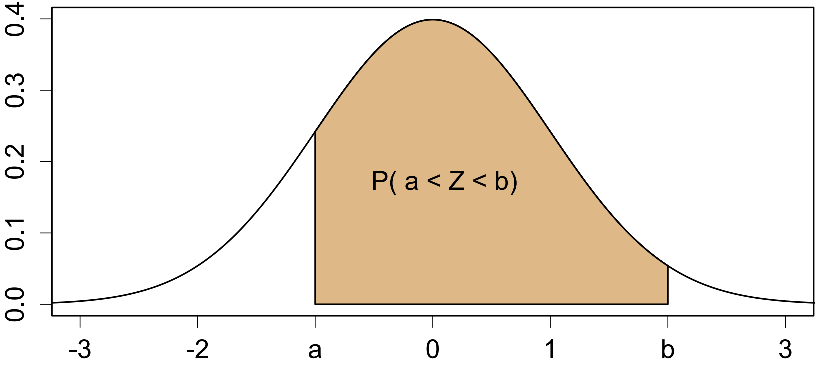
Standard normal distribution $Z \sim N(0,1)$

- from symmetry of $N(0,1)$: $\Phi(-z) = 1 - \Phi(z)$
- e.g. $P(Z < -1{,}38) = \Phi(-1{,}38) = 1 - \Phi(1{,}38) \overset{\text{from tab.}}{=\!=\!=} 1 - 0{,}916 = 0{,}084$



Standard normal distribution $Z \sim N(0,1)$

- $P(a < Z < b) = P(Z < b) - P(Z < a) = \Phi(b) - \Phi(a)$
- e.g. $P(-1 < Z < 2) = \Phi(2) - \Phi(-1) \overset{\text{from tab.}}{=\!=\!=} 0{,}977 - 0{,}158 = 0{,}819$



General normal distribution $Z \sim N(\mu, \sigma^2)$

- for $X \sim N(\mu, \sigma^2)$ it holds that

$$Z \overset{\text{den.}}{=\!=} \frac{X-\mu}{\sigma} \sim N(0,1)$$

- $P(X < x) = P(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}) = \Phi\left(\frac{x-\mu}{\sigma}\right)$
- so

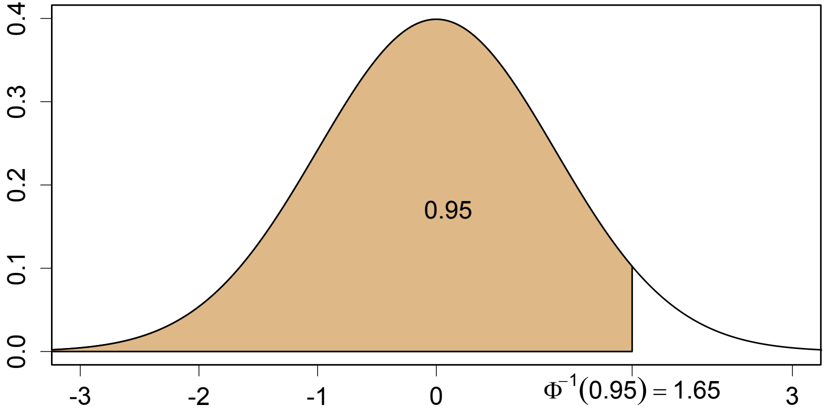
$$P(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Ex.: Hight of boys in the sixth grade $X \sim N(\mu = 143, \sigma^2 = 49)$:
find $P(130 < X < 150) = \Phi\left(\frac{150-143}{7}\right) - \Phi\left(\frac{130-143}{7}\right) \doteq 0{,}81$
so approximately 81% of boys in the sixth grade are 130 to 150 cm tall.

Ex.: What hight is exceeded by only 5% of boys in the sixth grade?
. . . 95% quantile of distribution $N(\mu = 143, \sigma^2 = 49)$

Quantiles of standard normal distribution 1

- quantile funct. of $Z \sim N(0,1)$ is denoted by $\Phi^{-1}(\alpha)$
- it holds: $P(Z < \Phi^{-1}(\alpha)) = \Phi(\Phi^{-1}(\alpha)) = \alpha$
- can be found in tables of $\Phi(x)$ opposite way
- frequently used: $\Phi^{-1}(0{,}95) = 1{,}65$ a $\Phi^{-1}(0{,}975) = 1{,}96$

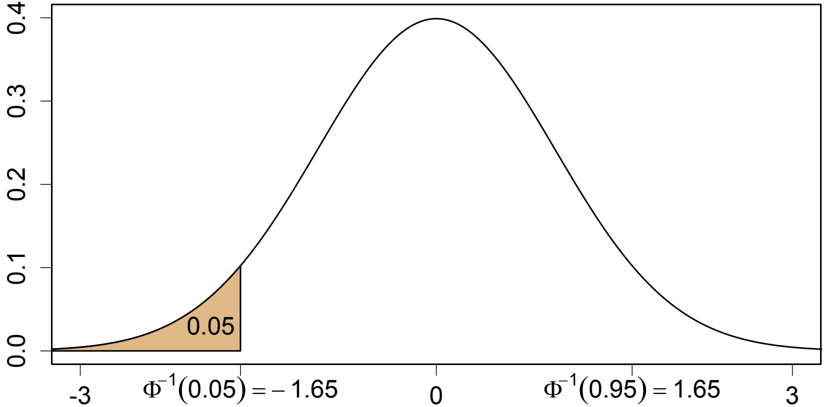


Quantiles of standard normal distribution 2

- in tables often quantiles only for $\alpha \geq 0{,}5$
- for $\alpha < 0{,}5$ we can use (follows from symmetry of distribution):

$$\Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$$

- e.g.: 5% quantile $N(0,1)$ is $\Phi^{-1}(0{,}05) = -\Phi^{-1}(0{,}95) = -1{,}65$



Quantiles of general normal distribution

- for $X \sim N(\mu, \sigma^2)$ it holds that $Z \overset{\text{den.}}{=\!=} \frac{X-\mu}{\sigma} \sim N(0,1)$
- α-quantile of r. v. X is such value h, for which

$$P(X < h) = \alpha \qquad \Phi\left(\frac{h-\mu}{\sigma}\right) = \alpha$$

$$P\left(\frac{X-\mu}{\sigma} < \frac{h-\mu}{\sigma}\right) = \alpha \qquad \frac{h-\mu}{\sigma} = \Phi^{-1}(\alpha)$$

$$P\left(Z < \frac{h-\mu}{\sigma}\right) = \alpha \qquad h = \sigma \cdot \Phi^{-1}(\alpha) + \mu$$

Ex.: Find 95% quantile of distribution $N(\mu = 143, \sigma^2 = 49)$
is equal to $\sigma \cdot \Phi^{-1}(0{,}95) + \mu = 7 \cdot 1{,}65 + 143 = 154{,}5$
so only 5% of boys in the sixth grade is taller than 154,5 cm.

Random sample

Random sample is a set $X_1, X_2, \dots, X_n$ of independent and identically distributed random variables.

▶ Ex. 1: Hight of boys in the sixth grade, large population, we choose n boys at random and measure their height X_i

▶ Ex. 2: Measuring strength of a fabric, we measure the strength at n randomly chosen samples

- number of variables **n** is called **sample size**
- parameters of distribution (expectation μ, variance σ^2, etc.) of ran. var. X_i is often not known
- these parameters can be inferred from the ran. sample
- **sample mean** $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a (point) estimate of expectation (of height, strength)
- **sample variance** $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is a (point) estimate of variance of the distribution
- $\overline{X}$ and S^2 random variables as well

Properties of sample mean

Let $X_1, X_2, \ldots, X_n$ is random sample from distribution with expectation μ and variance σ^2. Then

1) $E\overline{X} = \mu$ ($\overline{X}$ is unbiased estimate of μ)
2) $var\,(\overline{X}) = \frac{\sigma^2}{n}$

Proof: ad 1) From ▸ properties of expectation (points 1) and 5)) follows:

$$E\overline{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

ad 2) From ▸ Properties of variance (points 2) and 6)) follows:

$$var\,(\overline{X}) = var\,\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} var\,X_i = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

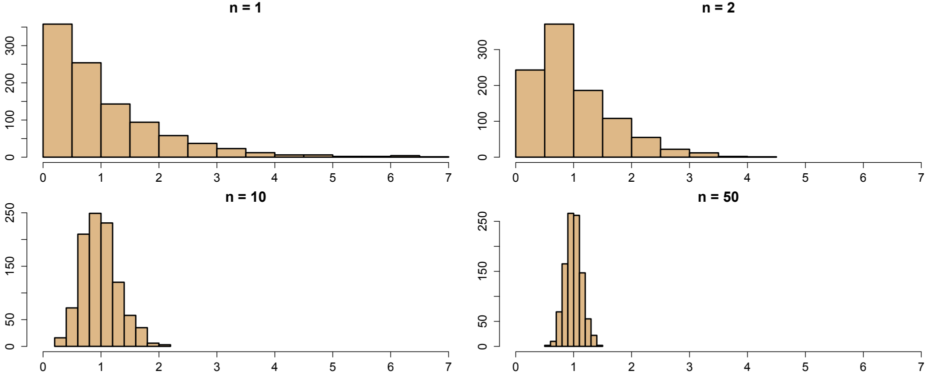
Comment:

- from the proof follows, that $E(\sum_{i=1}^{n} X_i) = n \cdot \mu$ and $var\,(\sum_{i=1}^{n} X_i) = n \cdot \sigma^2$
- unbiasedness of sample variance follows similarly, i.e. $ES^2 = \sigma^2$

Histograms of averages

Ex.: lifetime of the fluor tube is of interest, we choose randomly *n* tubes, test them and calculate their average lifetime. We determine 1 000 such averages and draw their histogram. (Data generated from *Exp*$(\lambda = 1)$)

- with increasing *n* the variability of averages decreases and normality improves (see CLT)



Central limit theorem

Let $X_1, X_2, \ldots, X_n$ is a random sample from distribution with expectation μ and finite variance σ^2. Then in the limit for $n \to \infty$

1) sample mean $\overline{X}$ is normally distributed $N(\mu, \frac{\sigma^2}{n})$
2) the sum $\sum_{i=1}^n X_i$ is normally distributed $N(n \cdot \mu, n \cdot \sigma^2)$

for sufficiently large n we can write

$$Z = \frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} = \frac{\sum_{i=1}^n X_i - n \cdot \mu}{\sqrt{n} \cdot \sigma} \stackrel{\cdot}{\sim} N(0,1)$$

and so

- $$P(\overline{X} < y) = P\left(\frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} < \frac{y - \mu}{\sigma} \cdot \sqrt{n}\right) \doteq \Phi\left(\frac{y - \mu}{\sigma} \cdot \sqrt{n}\right)$$

- $$P\left(\sum_{i=1}^n X_i < y\right) = P\left(\frac{\sum_{i=1}^n X_i - n \cdot \mu}{\sqrt{n} \cdot \sigma} < \frac{y - n \cdot \mu}{\sqrt{n} \cdot \sigma}\right) \doteq \Phi\left(\frac{y - n \cdot \mu}{\sqrt{n} \cdot \sigma}\right)$$

Ex. (CLT): 300 numbers are rounded off to one decimal place and then summed. Approximate the probability that the resultant sum differs from the exact sum by more than 1.

- Rounding error of one number is less then 0,05; the resultant sum will differ from the exact sum by less then $300 \cdot 0{,}05 = 15$.
- Roundoff errors X_i ($i = 1, \dots, 300$) can be assumed to be independent ran. var. from uniform distribution ▸ on interval $(-0{,}05; 0{,}05)$.
- So $EX_i = \frac{-0{,}05+0{,}05}{2} = 0$ and $var\, X_i = \frac{(0{,}05+0{,}05)^2}{12} = \frac{1}{1200}$
- The difference $Y = \sum_{i=1}^{n} X_i$ is thus approx. normally distr. $N(0, \frac{300}{1200} = \frac{1}{4})$ and the probability is

$$P(|Y| < 1) = P(-1 < Y < 1) = P(Y < 1) - P(Y < -1) =$$

$$= P\left(\frac{Y}{\sqrt{1/4}} < \frac{1}{\sqrt{1/4}}\right) - P\left(\frac{Y}{\sqrt{1/4}} < \frac{-1}{\sqrt{1/4}}\right) =$$

$$= \Phi(2) - \Phi(-2) = \Phi(2) - (1 - \Phi(2)) = 2 \cdot \Phi(2) - 1 =$$

$$= 2 \cdot 0{,}9772 - 1 = 0{,}9544$$

de Moivre-Laplace theorem

Let Y binomial random variable $Bi(n, p)$. Then for $n \to \infty$ Y is normally distributed

$$N(n \cdot p, n \cdot p \cdot (1 - p))$$

Proof:

- binomial random var. $Bi(n, p)$ can be seen as a sum of n independent Bernoulli random var. with par. p
- Thus (from CLT for sum) Y is as $n \to \infty$ normal r. v. with expectation $EY = n \cdot p$ and variance $var\ Y = n \cdot p \cdot (1 - p)$

Ex. (de Moivre-Laplace theorem): it is known, that 52% of population agrees with the death penalty. What is the probability that in a survey of $n = 1\,000$ people the majority will be against the death penalty?

- denote Y the number of supporters in the sample
- if randomly selected, then $Y \sim Bi(n = 1\,000, p = 0{,}52)$
- according to dM-L theorem Y is approx. normal $N(1\,000 \cdot 0{,}52 = 520, 1\,000 \cdot 0{,}52 \cdot 0{,}48 = 249{,}6)$
- majority in the survey is against if the number of supporters is less than 500, so the probability is

$$P(Y < 500) = P\left(\frac{Y - 520}{\sqrt{249{,}6}} < \frac{500 - 520}{\sqrt{249{,}6}}\right) = \Phi\left(\frac{500 - 520}{\sqrt{249{,}6}}\right) =$$

$$= \Phi(-1{,}27) = 1 - \Phi(1{,}27) = 1 - 0{,}898 = 0{,}102$$

Ex.: Consider an automatic machine which bottles cola into 2-liter (2000 ml) bottles. Consumer protection requires the average amount to be at least 2000 ml and want to check this. So there were 100 bottles randomly selected and tested for the exact amount with mean $\overline{X} = 1{,}982$ liter. Moreover we know the standard deviation of the machine is $\sigma = 0{,}05$ liter (so the variance $\sigma^2 = 0{,}0025$ liter^2) and the amount in a bottle is approx. normally distributed r. v. $N(\mu, \sigma^2 = 0{,}0025)$. Do the data confirm the hypothesis that the machine is incorrectly adjusted and consumers do not get their money's worth?

- $\overline{X} = 1{,}982$ is a point estimate of average amount in a bottle μ. For each random sample of bottles we would get different estimate (average). What now?
- Cannot we find an interval (...interval estimate), about which we could say that covers the unknown mean amount μ with large probability?
- How to verify the hypothesis (...hypothesis test), that the machine is incorrectly adjusted?

Mathematical statistics

Assume that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution (usually) with unknown parameters
We usually assume, that the distribution is given (often normal) and we try to estimate its unknown parameters or verify (test) hypotheses about the parameters (for norm. distr. expectation μ and variance σ^2)

- **point estimate** of an unknown parameter is a value calculated from the realized random sample, e.g. $\overline{X}$ is a point estimate of μ
- **interval estimate** of an unknown parameter (also **confidence interval**) is an interval (which is calculated from the observed sample), that covers the unknown parameter with given probability
- by **hypothesis testing** we try to decide between two antagonistic hypotheses about a given parameter of the distribution, e.g. the machine is adequately calibrated ($\mu = 2$ liter) or not ($\mu \neq 2$ liter)

Confidence int. for μ when σ^2 is known, for $N(\mu, \sigma^2)$

For normal random sample $X_1, X_2, \ldots, X_n$ from $N(\mu, \sigma^2)$ it holds

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

proto

$$\frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} \sim N(0, 1)$$

and so

$$P\left(-\Phi^{-1}(1 - \alpha/2) < \frac{\overline{X} - \mu}{\sigma} \cdot \sqrt{n} < \Phi^{-1}(1 - \alpha/2)\right) = 1 - \alpha$$

$100(1 - \alpha)\%$ confidence interval for μ and known σ^2 is

$$\left(\overline{X} - \Phi^{-1}(1 - \alpha/2) \cdot \frac{\sigma}{\sqrt{n}}; \ \overline{X} + \Phi^{-1}(1 - \alpha/2) \cdot \frac{\sigma}{\sqrt{n}}\right)$$

this interval (is random) covers the unknown mean μ with probability $1 - \alpha$

- only approx. $100(1 - \alpha)\%$ such intervals include the unknown μ

back to ▸ Ex.: 100 bottles of cola randomly selected, the average amount $\overline{X} = 1{,}982$ liter. The individual amounts are considered realization of a random sample from distribution $N(\mu, \sigma^2 = 0{,}0025)$. We calculate 95% confidence interval for the mean amount of coly in a bottle μ.

- $100(1-\alpha)\%$ conf. int. is
  $\left(\overline{X} - \Phi^{-1}(1-\alpha/2)\cdot\frac{\sigma}{\sqrt{n}};\ \overline{X} + \Phi^{-1}(1-\alpha/2)\cdot\frac{\sigma}{\sqrt{n}}\right)$
- for 95% conf. int. set $\alpha = 0{,}05$ and find
  $\Phi^{-1}(1-0{,}05/2) = \Phi^{-1}(0{,}975) = 1{,}96$
- plugging $\overline{X} = 1{,}982$, $\sigma = 0{,}05$ a $n = 100$:

$$\left(1{,}982 - 1{,}96\cdot\frac{0{,}05}{\sqrt{100}};\ 1{,}982 + 1{,}96\cdot\frac{0{,}05}{\sqrt{100}}\right) \doteq$$

$$\doteq (1{,}982 - 0{,}010;\ 1{,}982 + 0{,}010) =$$

$$= (1{,}972;\ 1{,}992)$$

With probability 95% this interval includes the unknown mean μ, but does not contain 2. With high certainty we can claim, that the machine is not adequately calibrated.

Ex.: 16 samples of a new alloy were tested on strength in tension with the following results (in megapascals):

13,1 16,7 14,5 10,5 15,9 16,5 20,5 17,9
18,2 19,5 8,9 16,3 15,5 15,8 25,8 23,4

Measurements will be considered a random sample from distribution $N(\mu, \sigma^2)$. We want 95% confidence interval for the mean tensile strength.

Problem: cannot use the foregoing procedure, since the standard deviation σ is not known.

Confidence interval for μ and σ^2 unknown for $N(\mu, \sigma^2)$

unknown σ replaced by estimate, by so called **sample standard deviation**

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$100(1-\alpha)\%$ confidence interval for μ and σ^2 uknown for sample from normal distr. is

$$\left(\overline{X} - t_{n-1}(1-\alpha/2)\cdot\frac{S}{\sqrt{n}};\ \overline{X} + t_{n-1}(1-\alpha/2)\cdot\frac{S}{\sqrt{n}}\right)$$

- quantile $\Phi^{-1}(1-\alpha/2)$ is replaced by the quantile $t_{n-1}(1-\alpha/2)$ (is larger $\rightarrow$ wider interval) is a penalty for replacing the unknown value of σ its estimate S.
- $t_n(\alpha)$ denotes the α-quantile of so called Student's t-distribution with n degrees of freedom; can be found in tables
- interpretation is the same as for the previous interval

back to ▸ Ex.: From 16 measurements we want to calculate the 95% confidence interval for the mean tensile strength.

- we find $\overline{X} = 16{,}8125$, $S = 4{,}2711$ and set $n = 16$
- for 95% conf. int. we set $\alpha = 0{,}05$ and find $t_{15}(1 - 0{,}05/2) = t_{15}(0{,}975) \doteq 2{,}13$

So with probability 95% the mean tensile strength is covered by interval:

$$\left(\overline{X} - t_{n-1}(1-\alpha/2)\cdot\frac{S}{\sqrt{n}};\ \overline{X} + t_{n-1}(1-\alpha/2)\cdot\frac{S}{\sqrt{n}}\right) \doteq$$

$$\doteq \left(16{,}8125 - 2{,}13\cdot\frac{4{,}2711}{\sqrt{16}};\ 16{,}8125 + 2{,}13\cdot\frac{4{,}2711}{\sqrt{16}}\right) \doteq$$

$$\doteq (16{,}8125 - 2{,}274;\ 16{,}8125 + 2{,}274) \doteq$$

$$\doteq (14{,}54;\ 19{,}09)$$

- for 99% conf. int. is $\alpha = 0{,}01$ and $t_{15}(1 - 0{,}01/2) = t_{15}(0{,}995) = 2{,}95$

so 99% conf. interval forμ is $(13{,}66;\ 19{,}96)$

How to compute a conf. interval for the variance (variability of measurements) σ^2?

Conf. interval for σ^2 for $N(\mu, \sigma^2)$

Assume that $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$. can be proven that

$$P\left(\chi^2_{n-1}(\alpha/2) < \frac{(n-1)\cdot S^2}{\sigma^2} < \chi^2_{n-1}(1-\alpha/2))\right) = 1-\alpha$$

- where $\chi^2_n(\alpha)$ denotes the α-quantile of so called χ^2-distribution with n degrees of freedom; can be found in tables

$100(1-\alpha)\%$ conf. interval for σ^2 for a smaple from normal distribution is

$$\left(\frac{(n-1)\cdot S^2}{\chi^2_{n-1}(1-\alpha/2)}; \frac{(n-1)\cdot S^2}{\chi^2_{n-1}(\alpha/2)}\right)$$

- interpretation is again the same

back to ▸ Ex.: From 16 measurements we want to calculate the 95% confidence interval the variance of tensile strength.

- we have $\overline{X} = 16{,}8125$, $S^2 = 4{,}2711^2$ and $n = 16$
- for 95% conf. int. we set $\alpha = 0{,}05$
- and find $\chi^2_{15}(1 - 0{,}05/2) = \chi^2_{15}(0{,}975) = 27{,}49$ and $\chi^2_{15}(0{,}05/2) = \chi^2_{15}(0{,}025) = 6{,}26$

So with probability 95% the variance is covered by interval:

$$\left(\frac{(n-1) \cdot S^2}{\chi^2_{n-1}(1-\alpha/2)};\ \frac{(n-1) \cdot S^2}{\chi^2_{n-1}(\alpha/2)} \right) \doteq$$

$$\doteq \left(\frac{15 \cdot 4{,}2711^2}{27{,}49};\ \frac{15 \cdot 4{,}2711^2}{6{,}26} \right) \doteq$$

$$\doteq (9{,}95;\ 43{,}71)$$

Ex.: Error rate of a machine producing certain component should not exceed 10%. Inspection of a random sample of 400 components found 42 defective components. How to find 95% amd 99% confidence interval for the error rate of the machine?

- denote p the unknown error rate
- $n = 400$ of components randomly chosen, each is defective with probability p
- so the total number of defective is $Y \sim Bi(n = 400, p)$
- in the random sample the number of defective was (absolute frequency) $y = 42$ (by realization of Y the value of y was found)

- the point estimate of p is the relative freq. $\hat{p} = \frac{y}{n} = \frac{42}{400} = 0{,}105$
- how can we obtain an interval estimate of p?
  - from CLT (deMoivre-Laplace ▸ theorem): for $Y \sim Bi(n, p)$ $Y \overset{\cdot}{\sim} N(n \cdot p, n \cdot p \cdot (1 - p))$ for sufficiently large n
  - so $\frac{Y}{n} \overset{\cdot}{\sim} N(p, \frac{p \cdot (1-p)}{n})$

Conf. interval for parameter p of binomial distr.

Let Y is a binomial $Bi(n,p)$ random varible, then $\frac{Y}{n} \dot{\sim} N(p, \frac{p\cdot(1-p)}{n})$ and because variance of Y is unknown (due to unknown p), we replace p in the variance term by its estimate $\hat{p}$. So $\frac{Y}{n} \dot{\sim} N(p, \frac{\hat{p}\cdot(1-\hat{p})}{n})$ and

$$P\left(-\Phi^{-1}(1-\alpha/2) < \frac{\frac{Y}{n}-p}{\sqrt{\hat{p}\cdot(1-\hat{p})}} \cdot \sqrt{n} < \Phi^{-1}(1-\alpha/2)\right) = 1-\alpha$$

$\frac{Y}{n}$ is then replaced by the observed relative frequency $\frac{y}{n} = \hat{p}$. So we get:
$100(1-\alpha)\%$ conf. int. for parameter p of binomial distribution:

$$\left(\hat{p} - \Phi^{-1}(1-\alpha/2) \cdot \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}};\ \hat{p} + \Phi^{-1}(1-\alpha/2) \cdot \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}\right)$$

- interpretation is similar

back to ▸ Ex.: From 400 randomly chosen components were 42 defective. We want to determine the 95% and 99% conf. interval for the error rate.

- point estimate of the error rate p is the ratio of defective in the sample $\hat{p} = \frac{y}{n} = \frac{42}{400} = 0{,}105$
- for 95% (or 99%) conf. int. we set $\alpha = 0{,}05$ (or. $\alpha = 0{,}01$)
- and find $\Phi^{-1}(1 - 0{,}05/2) = \Phi^{-1}(0{,}975) = 1{,}96$ and $\Phi^{-1}(1 - 0{,}01/2) = \Phi^{-1}(0{,}995) = 2{,}58$

So 95% conf. int. for the error rate p is:

$$\left(\hat{p} - \Phi^{-1}(1 - \alpha/2) \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}};\ \hat{p} - \Phi^{-1}(1 - \alpha/2) \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}\right) \doteq$$

$$\doteq \left(0{,}105 - 1{,}96 \cdot \sqrt{\frac{0{,}105 \cdot (1 - 0{,}105)}{400}};\ 0{,}105 + 1{,}96 \cdot \sqrt{\frac{0{,}105 \cdot (1 - 0{,}105)}{400}}\right)$$

$$\doteq (0{,}075;\ 0{,}135) = (7{,}5\%;\ 13{,}5\%)$$

or 99% conf. int. $(0{,}065;\ 0{,}145) = (6{,}5\%;\ 14{,}5\%)$

Properties of confidence intervals

- interval is wider for higher confidence level (see the last example)
- interval is narrower for larger n (sample size)
  ► e.g. for interval for μ for $N(\mu, \sigma^2)$ or for p for $Bi(n, p)$ the width is inversely proportional to the square root of n; and so for half width (more precise) interval we need 4-times more observations
- in some situations from the requirement on the width we can estimate the necessary sample size n.

How to verify hypotheses?

- how to decide whether a hypothesis about an unknown parameter of a distribution is true?
- we have calculated a confidence interval for mean amount of cola in a bottle μ: $(1{,}972;\ 1{,}992)$
- can we (and with what certainty) claim, that the machine is incorrectly adjusted?
- requirement: we would like e.g. that the probability of “false accusation” was small
- thus: we introduce standardized methods of making such decisions

Hypothesis testing

$X_1, X_2, \ldots, X_n$ is a random sample from a distr. with unknown parameter(s).

We have two hypothesis about a parameter(s) of the given distribution:

- so called **null hypothesis $\mathbf{H_0}$**: parameter is equal to certain value, parameters are equal,...
- so called **alternative hypothesis $\mathbf{H_1}$**: opposite of the null hypothesis, often what we want to prove

According to the type of H_0 and H_1 we choose the criterion (a test), is a function of the realized random sample (observed data).

Possible decisions:

- reject H_0, if data (and so the test) give evidence against it
- do not reject H_0, if data (and so the test) does not provide enough “evidence” against H_0

Method and possible errors

- **type** 1 **error**: H_0 is true and we reject it
- **type** 2 **error**: H_0 is not true and we do not reject it

significance level of a test: denoted by α (we set it, often $=$ 0,05), is the maximal acceptable level of type 1 error

decision\reality	H_0 holds	H_0 does not hold
do not reject H_0	right	type 2 error
reject H_0	type 1 error $\leq \alpha$	right

Strategy: according to what we want to find out we formulate H_0 and H_1 and set α. then we choose appropriate test (criterion): i.e. from all the tests with significance level less than α we usually choose that with the minimal probability of type 2 error back to ▸ Ex.: Randomly chosen 100 bottles of cola with average amount $\overline{X} = 1{,}982$ liter. Obtained values are considered a realization of random sample from $N(\mu, \sigma^2 = 0{,}0025)$. Can we claim that the machine is incorrectly adjusted?

We would like to test at level $\alpha = 0{,}05$ a hypothesis

- $H_0 : \mu = 2$ liter (adequately calibrated)

against an alternative

- $H_1 : \mu \neq 2$ liter (inadequately calibrated)

What test to use?

Z-test: one-sample test of mean (σ^2 known)

$X_1, X_2, \dots, X_n$ is a random sample from $N(\mu, \sigma^2)$, where σ^2 is known. From what was derived follows

$$P\left(\frac{|\overline{X} - \mu|}{\sigma} \cdot \sqrt{n} \geq \Phi^{-1}(1 - \alpha/2)\right) = \alpha$$

So to test the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ we can use statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma} \cdot \sqrt{n}$$

and at level α we reject the hypothesis H_0 (accept H_1), if $|Z| \geq \Phi^{-1}(1 - \alpha/2)$

- if $|Z| < \Phi^{-1}(1 - \alpha/2)$, then H_0 is not rejected. Conclusion: H_0 can be true
- Note: this holds for sufficiently large n also for other distributions than normal thanks to Central limit theorem

back to ▸ Ex.: 100 bottles of cola randomly chosen, $\overline{X} = 1{,}982$ liter. Assume, that data come from $N(\mu, \sigma^2 = 0{,}0025)$. Can we claim that the machine is inadequately calibrated?

We would like to test at level $\alpha = 0{,}05$ a hypothesis

- $H_0 : \mu = 2$ liter (adequately calibrated)

against

- $H_1 : \mu \neq 2$ liter (inadequately calibrated)

Criterion (test statistic) is

$$Z = \frac{\overline{X} - \mu_0}{\sigma} \cdot \sqrt{n} = \frac{1{,}982 - 2}{0{,}05} \cdot \sqrt{100} = -3{,}6$$

So

$$|Z| = 3{,}6 \geq \Phi^{-1}(1 - \alpha/2) = \Phi^{-1}(0{,}975) = 1{,}96$$

and that is why at level 0,05 we reject H_0 and accept H_1

Conclusion: automatic machine is inadequately calibrated back to ▸ Ex.: 16 samples of a new alloy were tested on strength in tension. Assume, that data come from $N(\mu, \sigma^2)$. Can we conclude, that the strength has changed compared to the previous alloy with strength 14 megapascalů? Let the level of the test be $\alpha = 0{,}01$

We would like to test on the level $\alpha = 0{,}01$ the hypothesis

- $H_0 : \mu = 14$ MPa (the strength has not changed)

against the alternative

- $H_1 : \mu \neq 14$ MPa (the strength has changed)

Problem: cannot use the foregoing method, since we do not know the standard deviation of the measurements σ.

One-sample t-test: test for the mean (σ^2 unknown)

$X_1, X_2, \ldots, X_n$ is random sample from $N(\mu, \sigma^2)$, where σ^2 is unknown. It holds that $\frac{\overline{X}-\mu}{S} \cdot \sqrt{n} \sim t_{n-1}$, thus similarly as for Z-test it follows:

$$P\left(\frac{|\overline{X}-\mu|}{S} \cdot \sqrt{n} \geq t_{n-1}(1-\alpha/2)\right) = \alpha$$

so for the test of $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ we can use the test statistic

$$T = \frac{\overline{X}-\mu_0}{S} \cdot \sqrt{n}$$

and the test of level α reject the hypothesis H_0 (H_1 is accepted), if $|T| \geq t_{n-1}(1-\alpha/2)$

- if $|T| < t_{n-1}(1-\alpha/2)$, then H_0 is not rejected. conclusion: H_0 can be true back to ▸ Ex.: We have 16 sample measurements. They come from $N(\mu, \sigma^2)$. Can we conclude, that the strength has changed compared to the previous alloy with strength 14 MPa?

We set the significance level $\alpha = 0{,}01$, and we test the hypothesis

- $H_0 : \mu = 14$ MPa (the strength has not changed)

against alternative

- $H_1 : \mu \neq 14$ MPa (the strength has changed)

Criterion (test statistic) is

$$T = \frac{\overline{X} - \mu_0}{S} \cdot \sqrt{n} = \frac{16{,}8125 - 14}{4{,}2711} \cdot \sqrt{16} = 2{,}634$$

So

$$|T| = 2{,}634 < t_{n-1}(1 - \alpha/2) = t_{15}(0{,}995) = 2{,}947$$

and so on the level 0,01 we do not reject H_0

Conclusion: Strength can be equal to the strength of the previous alloy

- Note: T-test of significance level $\alpha = 0{,}05$ would reject H_0 (H_1 would be accepted), because

$$|T| = 2{,}634 \geq t_{n-1}(1 - \alpha/2) = t_{15}(0{,}975) = 2{,}131$$

Paired t-test

Sometimes we have two sets of data (measurements) and try to compare them (their means). Denote the observed variables by $(X_1, Y_1), \dots, (X_n, Y_n)$ and assume that the random variables X and Y with the same index cannot be considered independent (often because they are measured on the same object), but rand. variables with different indices can be considered independent (measurements are unrelated, e.g. because they are made on different objects).

Ex.: Random sample of 8 people were keeping a certain type of diet. Table shows their weigth (in kg) before the diet and after.

Person	1	2	3	4	5	6	7	8
Before	81	85	92	82	86	88	79	85
After	84	68	73	79	71	80	71	72

We would like to find out whether the diet influence the weigth.

What test to use?

Paired t-test

We assume to have two-dimensional random sample (X_1, Y_1), $\ldots, (X_n, Y_n)$ such that X and Y form pairs, that can be assumed independent. denote $\mu_X = EX_i$ a $\mu_Y = EY_i$.

Then set $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$ and assume that variables Z can be considered to be a random sample from $N(\mu, \sigma^2)$, where $\mu = \mu_X - \mu_Y$.
So the test of hypothesis, that both sets of measurements come from a distributions with identical mean $H_0: \mu_X - \mu_Y = 0$ is equivalent to the hypothesis $H_0: \mu = 0$. Test of hypotheses $H_0: \mu = 0$ against $H_1: \mu \neq 0$ is a one-sample t-test problem.

So we calculate $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ a $S_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2$ and if

$$|T| = \frac{|\overline{Z} - 0|}{S_Z} \cdot \sqrt{n} \geq t_{n-1}(1 - \alpha/2)$$

then the test of level α rejects the hypothesis H_0 (we accept $H_1: \mu_X \neq \mu_Y$)

back to ▸ Ex.: 8 people keeping a diet. Does it influence the weight?

Person	1	2	3	4	5	6	7	8
X=Before	81	85	92	82	86	88	79	85
Y=After	84	68	73	79	71	80	71	72
Z=Difference	-3	17	19	3	15	8	8	13

We conduct level $\alpha = 0{,}05$ test of hypothesis

- $H_0 : \mu = \mu_X - \mu_Y = 0$ kg (diet does not influence weight)
- against $H_1 : \mu = \mu_X - \mu_Y \neq 0$ kg (diet does influence weight)

Calculate $\overline{Z} = 10$ and $S_Z = \sqrt{S_Z^2} = \sqrt{55{,}71429} = 7{,}4642$ Test statistic is

$$T = \frac{\overline{Z} - 0}{S_Z} \cdot \sqrt{n} = \frac{10 - 0}{7{,}4642} \cdot \sqrt{8} = 3{,}789$$

So

$$|T| = 3{,}789 \geq t_{n-1}(1 - \alpha/2) = t_7(0{,}975) = 2{,}365$$

and thus at the signif. level 0,05 we reject H_0.
Conclusion: diet does influence the weight.

- Note: even for $\alpha = 0{,}01$ we would reject H_0 ($t_7(0{,}995) = 3{,}499$)

Two-sample t-test

Sometimes we have two sets of data (measurements) and try to compare them (their means), but the variables in the pairs are not dependent and the two samples can be of different sample size. Denote the observable variables as $X_1, \dots, X_n$ and $Y_1 \dots, Y_m$ and we assume them to be two independent random samples (all the variables are independent).

Ex.: The following heights of students in the classroom were found out (in cm):

Boys	130	140	136	141	139	133	149	151
Girls	135	141	143	132	146	146	151	141
Boys	139	136	138	142	127	139	147	
Girls	141	131	142	141				

Test that boys and girls are on average equally tall. Set $\alpha = 0{,}05$.
What test to use?

Two-sample t-test

Assume we have random sample $X_1, \ldots, X_n \sim N(\mu_X, \sigma^2)$ and random sample $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2)$ and these two samples are independent with equal variance.
We set

$$S^{*2} = \frac{1}{n+m-2} \cdot \left((n-1) \cdot S_X^2 + (m-1) \cdot S_Y^2\right),$$

where $S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ a $S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \overline{Y})^2$.

For the test of hypothesis, that both sets of measurements come from distributions with the same mean $H_0 : \mu_X - \mu_Y = 0$ against alternative $H_1 : \mu_X - \mu_Y \neq 0$ we can use test statistic:

$$T = \frac{\overline{X} - \overline{Y} - 0}{S^*} \cdot \sqrt{\frac{n \cdot m}{n+m}}$$

and if $|T| \geq t_{n+m-2}(1 - \alpha/2)$ then at level α the hypothesis H_0 is rejected (we accept $H_1 : \mu_X \neq \mu_Y$ means are equal)

back to ▸ Ex.: Test at level $\alpha = 0{,}05$ hypothesis that boys and girls are on average equally tall.

Boys	130	140	136	141	139	133	149	151	139	136	138	142	127	139	147
Girls	135	141	143	132	146	146	151	141	141	131	142	141			

- test $H_0 : \mu_X - \mu_Y = 0$ cm (equally tall)
- against $H_1 : \mu_X - \mu_Y \neq 0$ cm (not equally tall)

We calculate $\overline{X} = 139{,}133$; $\overline{Y} = 140{,}833$; $S_X^2 = 42{,}981$; $S_Y^2 = 33{,}788$;

$$S^* = \sqrt{\frac{1}{n+m-2} \cdot \left((n-1) \cdot S_X^2 + (m-1) \cdot S_Y^2\right)} = \sqrt{\frac{1}{25}\left(14 \cdot 42{,}981 + 11 \cdot 33{,}788\right)} = 6{,}240$$

Test statistic is

$$T = \frac{\overline{X} - \overline{Y} - 0}{S^*} \cdot \sqrt{\frac{n \cdot m}{n+m}} = \frac{139{,}133 - 140{,}833 - 0}{6{,}240} \cdot \sqrt{\frac{15 \cdot 12}{15+12}} = -0{,}703$$

So $|T| = 0{,}703 < t_{n+m-2}(1 - \alpha/2) = t_{25}(0{,}975) = 2{,}060$ and so at level $0{,}05$ we do not reject H_0.

Conclusion: it is possible that boys and girls are equally tall on average.

Sign test

Sometimes we have only information how many times in a set of independent trials a variable exceeded (+) or not exceeded (-) certain value. We want to test a hypothesis, that both happens with the same probability, i.e. that median (50% quantile) of the distribution is equal to that value.

Ex.: From 46 beers, that were ordered at our table during one night, were 27 undersized and 19 oversized. Can we claim that the barman does not keep the correct size of a beer? (cheats either us or the bar owner)?
We want to verify whether the median amount of beer in a glass can be half a liter. And we know only number of beers below and above that measure. What test to choose?

Sign test - asymptotic (for large *n*)

Assume random sample $X_1, \dots, X_n$ from continuous distribution with median $\tilde{x}$. So it holds

$$P(X_i < \tilde{x}) = P(X_i > \tilde{x}) = \frac{1}{2} \qquad i = 1, \dots, n$$

We want to test $H_0 : \tilde{x} = x_0$ against $H_1 : \tilde{x} \neq x_0$, where x_0 is a given number.

We calculate the differences $X_1 - x_0, \dots, X_n - x_0$ and those equal to zero are omitted (and n is decreased adequately).

Under H_0 number of differences with a positive sign

$Y \sim Bi(n, p = 1/2)$ and so according to ▸ Moivreovy-Laplaceovy věty for large n:

Y is approx. normally distributed $N(n/2, n/4)$

Under H_0 thus

$$U = \frac{Y - n/2}{\sqrt{n/4}} = \frac{2Y - n}{\sqrt{n}} \overset{\cdot}{\sim} N(0, 1)$$

$H_0 : \tilde{x} = x_0$ at level α is rejected if $|U| \geq \Phi^{-1}(1 - \alpha/2)$

Sign test - exact

- is used if n is small

We use the fact that under H_0 the number of differences with positive sign $Y \sim Bi(n, p = 1/2)$ and so we expect the observed value Y to be close to its expectation $n/2$.
We accept $H_1 : \tilde{x} \neq x_0$ if Y too small ($\leq k_1$) or too large ($\geq k_2$).
We set level α.
Then k_1 is chosen as the largest number for which it still holds that

- $P(Y \leq k_1) \leq \alpha/2$

and k_2 is chosen as the smallest number for which it still holds that

- $P(Y \geq k_2) \leq \alpha/2$

H_0 is rejected at level α, if $Y \leq k_1$ or $Y \geq k_2$.
Note: The true signif. level of the test is often smaller than α

back to ▸ Ex.: From 46 beers 27 undersized and 19 oversized. Can we claim that the barman does not keep the correct size (is biased one or the other way)?

At level $\alpha = 0{,}05$ we test $H_0 : \tilde{x} = 500$ ml against $H_1 : \tilde{x} \neq 500$ ml.

Exact test:

We have $Y \sim Bi(n = 46, p = 1/2)$, $\alpha/2 = 0{,}025$ and find k_1 and k_2

k	14	**15**	16	...	30	**31**	32
$P(Y = k)$	0,003	0,007	0,014	...	0,014	0,007	0,003
$P(Y \leq k)$	0,006	0,013	0,027	...	0,987	0,994	0,998
$P(Y \geq k)$	0,998	0,994	0,987	...	0,027	0,013	0,006

Since $k_1 = 15 < Y = 19 < k_2 = 31$, we do not reject H_0 at level 0,05

Note: true level of the test (prob. of type 1. error) is only $2 \cdot 0{,}013 = 0{,}026$.

Asymptotic test: We calculate

$$U = \frac{2Y - n}{\sqrt{n}} = \frac{2 \cdot 19 - 46}{\sqrt{46}} = -1{,}180$$

neither this test rejects H_0, since $|U| = 1{,}180 \ngeq \Phi^{-1}(0{,}975) = 1{,}960$

Sign test - usage

- test about median for ran. sample $X_1, \ldots, X_n$ from contin. distr.
- can be used instead of one-sample (or paired) t-test
- advantage: no need for normality assumption
- disadvantage: for normal sample probab. of type 2. error is a bit larger compared to t-test
- For data from normal distribution t-test is the best choice

Tests for independence

Assume we have a random sample from two-dimensional distribution (repeated measurements of two variables) and try to find out, whether there is a dependence (correlation) between these two variables. Denote the observed values by $(X_1, Y_1), \dots, (X_n, Y_n)$.

Ex.: 9 students of statistical course were randomly selected and put through a math and language test with the following results:

Student number	1	2	3	4	5	6	7	8	9
Language test	50	23	28	34	14	54	46	52	53
Math test	38	28	14	26	18	40	23	30	27.

We want to find out, whether students' math and language scores are correlated.

Note: Not the same as decision whether the math and lang. scores are at the same level (in that case it would be a paired t-test problem)

What test to choose?

(Pearson) correlation coefficient

Assume we have a two-dimensional random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, i.e. variables with different indeces are independent. Denote S_X^2 and S_Y^2 to be sample variances of X and Y and **sample covariance** between X and Y as

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) \cdot \left(Y_i - \overline{Y}\right) = \frac{1}{n-1} \left[\sum_{i=1}^{n} (X_i \cdot Y_i) - n \cdot \overline{X} \cdot \overline{Y}\right]$$

(Pearson) sample correlation coefficient:

$$r_{XY} = r = \frac{S_{XY}}{\sqrt{S_x^2 \cdot S_y^2}} = \frac{\sum_{i=1}^{n} (X_i \cdot Y_i) - n \cdot \overline{X} \cdot \overline{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n \cdot \overline{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n \cdot \overline{Y}^2\right)}}$$

Under normality assumtion we calculate

$$T = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}$$

and hypothesis of independence of X and Y at level α is rejected if $|T| \geq t_{n-2}(1 - \alpha/2)$

At level $\alpha = 0{,}05$ we test the hypothesis of independence between math and language scores from ▸ example, where 9 students were chosen and given both tests.

Language test	50	23	28	34	14	54	46	52	53
Math test	38	28	14	26	18	40	23	30	27

We get $S_X^2 = 223{,}25$ and $S_Y^2 = 70{,}86$ and
$S_{XY} = \frac{1}{8}\left(50 \cdot 38 + \ldots + 53 \cdot 27 - 9 \cdot 39{,}33 \cdot 27{,}11\right) = 85{,}46$
correlation coef. is thus $r = \frac{S_{XY}}{\sqrt{S_x^2 \cdot S_y^2}} = \frac{85{,}46}{14{,}94 \cdot 8{,}42} = 0{,}679$
We get

$$T = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n-2} = \frac{0{,}679}{\sqrt{1 - 0{,}679^2}} \cdot \sqrt{7} = 2{,}450$$

and since $|T| = 2{,}450 \geq t_{n-2}(0{,}975) = 2{,}365$, we reject the hypothesis of independence at level 0,05. We can claim, that there is a relationship between math and language scores for students of that course

Test of independence in contingency table

Sometimes the observed data form a contingency table. For example, in case of two variables measured in nominal scale on *n* different object. The aim is to determine if two variables are dependent.

Ex.: middle-school teacher was interested in determining if there was a relationship between math anxiety and gender among students at her school. 100 students were randomly selected and given a psychological test which assessed a student's level of math anxiety (low, medium, and high). The gender of each student was also noted. The results are presented in the contingency table below:

	math anxiety			
gender	low	medium	high	sum
male	10	26	20	56
female	4	10	30	44
sum	14	36	50	100

we can perform a χ^2-test of independence: compare observed counts and expected cell counts under independence of the variables

	math anx.			
gender	low	med	hig	sum
male	10	26	20	56
female	4	10	30	44
sum	14	36	50	100

	math anx.			
gender	low	med	hig	sum
male	18%	46%	36%	100%
female	9%	23%	68%	100%
sum	14%	36%	50%	100%

- Does the level of math anxiety depend on gender?
- If independent, the percentages for both genders should be similar
- estimate of prob., that gender is female $P(\text{gend.} = \text{F}) = 44/100$
- estimate of prob., that anxiety is high $P(\text{anx.} = \text{v}) = 50/100$
- so estimate of prob. (if independent), that student is female with high anxiety
  $P(\text{gend.} = \text{F} \cap \text{anx.} = \text{H}) = (44/100) \cdot (50/100) = 0{,}22$
- so among 100 students we would expect
  $100 \cdot (44/100) \cdot (50/100) = 22$ such students
- similarly: expected counts for the remaining 5 cells.

χ^2 test of independence in contingency table

- denote n_{ij} count in i-th row and j-th column of the table (we have I rows and J columns)
- denote n_{i+} (or n_{+j}) sum of counts in i-th row (or j-th column)
- expected count in i-th row and j-th column under hypoth. of independence is
$$e_{ij} = \frac{n_{i+} \cdot n_{+j}}{n}$$

Test statistic is a goodness of fit measure between n_{ij} and o_{ij}:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

If $\chi^2 \geq \chi^2_{(I-1)\cdot(J-1)}(1-\alpha)$, we reject the hypothesis of independence of those two variables at level α.

► for the test to be valid all the expected cell counts are required to be larger than 5

At level $\alpha = 0{,}05$ we test hypothesis of independence between gender and math anxiety from ▸ example.

Observed (or expected) cell counts are:

	math anxiety			
gender	low	medium	high	sum
male	10 (7,84)	26 (20,16)	20 (28)	56
female	4 (6,16)	10 (15,84)	30 (22)	44
sum	14	36	50	100

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{(10 - 7{,}84)^2}{7{,}84} + \frac{(26 - 20{,}16)^2}{20{,}16} +$$

$$+ \frac{(20 - 28)^2}{28} + \frac{(4 - 6{,}16)^2}{6{,}16} + \frac{(10 - 15{,}84)^2}{15{,}84} + \frac{(30 - 22)^2}{22} = 10{,}39$$

We find out that $\chi^2 = 10{,}39 \geq \chi^2_{(I-1)\cdot(J-1)}(1 - \alpha) = \chi^2_2(0{,}95) = 5{,}99$

So we reject the hypothesis of independence at level 5%. Math anxiety level is related to gender.

- We can say that math anxiety is influenced by gender.

Predicting house prices

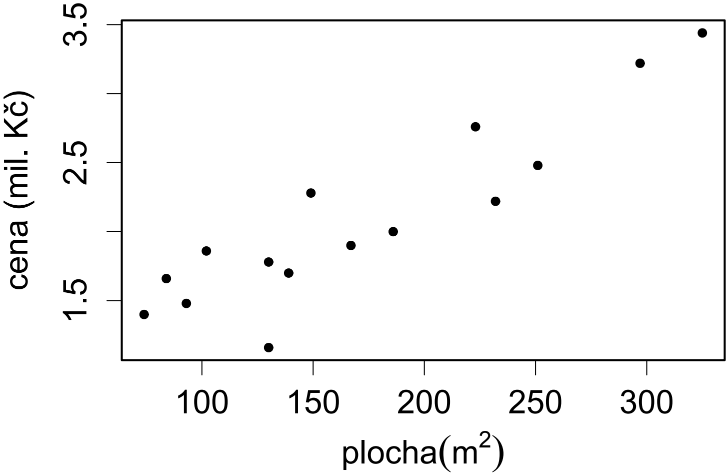
You want to sell a house of $200m^2$. How to predict its market price? We can use property sales data where we know the size and price.

- price of a house is influenced by many factors (neighborhood, size, in what condition it is etc.)
- for simplicity we use only size to predict price
- How to predict the price? Is there an exact way?
- price of house (in mil. Kč) and size of house (in m^2) were:

Size (x_i)	Price(Y_i)
74	1,40
84	1,66
93	1,48
102	1,86
130	1,78
130	1,16
139	1,70
149	2,28
167	1,90
186	2,00
223	2,76
232	2,22
251	2,48
297	3,22
325	3,44

Dependence of price on size

It is much more useful to look at the scatterplot:



- We can see that price more or less linearly changes with size

Regression line - least square method

- We have set of values $(x_i, Y_i), i = 1, \ldots, n$. We want from the set of explanatory variable x_i to estimate values of response variable Y_i (dependent variable)
- assumption: each size x_i corresponds to a average (mean) price Y_i that depends linearly on x_i:

$$EY_i = a + b \cdot x_i, \qquad i = 1, \ldots, n$$

- Moreover assume that Y_i are independent
  $Y_i \sim N(a + b \cdot x_i, \sigma^2), \quad i = 1, \ldots, n$
- Parameters a and b of regression line are estimated by means of **least square method**, i.e. we look for the values for which the expression $\sum_{i=1}^{n}(Y_i - (a + b \cdot x_i))^2$ is minimal. The solution is:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i \cdot Y_i) - n \cdot \overline{x} \cdot \overline{Y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} = \frac{S_{xY}}{S_x^2} \qquad \hat{a} = \overline{Y} - \hat{b} \cdot \overline{x}$$

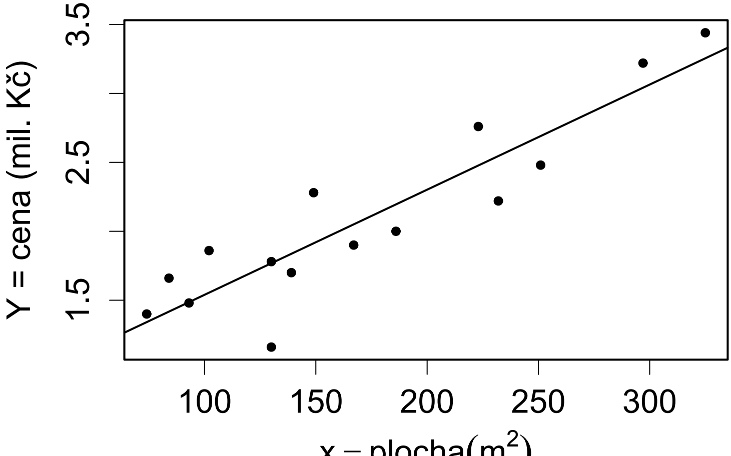
- **Residual sum of squares** (unexplained variability of Y): $S_e = \sum_{i=1}^{n}(Y_i - (\hat{a} + \hat{b} \cdot x_i))^2$ min. value of sum of squares
- **Residual variance:** $s^2 = S_e/(n-2)$
- equation of line estimating the dependence: $y = \hat{a} + \hat{b} \cdot x$
- Is the dependence significant? We test $H_0 : \; b = 0$ against $H_1 : \; b \neq 0$ using the statistic

$$T = \frac{\hat{b}}{s} \cdot \sqrt{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2}$$

  the hypothesis H_0 (that Y does not depend on x) at level α is rejected, if $|T| \geq t_{n-2}(1-\alpha/2)$
- **Coefficient of determination:** what part of the overall variability of dependent variable ($\sum_{i=1}^{n}(Y_i - \overline{Y})^2$) is explained by explanatory variable:

$$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}(= r_{xY}^2)$$

back to example. We want to estimate the least square line of how price depend on size. We get $\hat{b} = 0{,}0076(mil./m^2)$ and $\hat{a} = 0{,}777(mil.)$
equation of the line: $y = 0{,}777 + 0{,}0076 \cdot x$
interpretation of $\hat{b}$: with every m^2 the mean price of the house rises by 7 600 Kč interpretation of $\hat{a}$ (not always reasonable): price of 0 m^2 house is 777 600 Kč?



- the residual sum of squares: $S_e = 1{,}036$
- residual variance: $s^2 = S_e/(n-2) = 0.0797$
- Is this linear dependence significant? We test $H_0 : \; b = 0$ against $H_1 : \; b \neq 0$ using statistic

  $$T = \frac{\hat{b}}{s} \cdot \sqrt{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} = \frac{0{,}0076}{0{,}282} \cdot \sqrt{529780 - 15 \cdot 29629{,}88} = 7{,}9$$

  and since $|T| = 7{,}9 \geq t_{13}(0, 975) = 2{,}16$, the hypothesis $H_0 : \; b = 0$ (that price is independent of size) at level 0,05 is rejected.
- coefficient of determination:

  $$R^2 = 1 - \frac{S_e}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{1{,}036}{5.997} = 0{,}8272$$

  So 83% of variability of the price is explained by the linear dependence on size.
- estimate of the mean price of the $200 m^2$ house: $\hat{Y} = 0{,}777 + 0{,}0076 \cdot 200 = 2{,}297$