Analysis of technical and economical data

- What is the subject all about?

  ↳ what kind of consequences can be drawn from measured or observed data?

  ↳ how can we „compress" information?

  (measurement with 1000 data ↦ how can we compress it and get valuable information?)

- The way of teaching the subject

  ↳ usually: basic definitions, statements, methods, applications

  ↳ in this subject:

  - simple methods of data processing

    ↳ these require no new mathematical knowledge

    ↳ • just a few new definitions: descriptive statistics

    ↳ we will pose some questions, they will be answered later ↓

the general form of these answers will lead to the axioms of probability theory and the basics of statistics

/

Guideline of the subject:

- descriptive statistics (2 h)
- probability theory and basics of statistics (2 h)
- applications in measurement techniques (1 h)
  (error definitions, direct and indirect measurement)
- confidence interval (1 h)
- quality control and reliability (1 h)
- statistical tests (parametric; non-parametric) (2 h)
- analysis of variance (1 ÷ 2 h)
- relationship between variables
  - ↳ correlation coefficient
  - ↳ regression analysis
  - ↳ Wald method

  (2 ÷ 3 h)
- Conclusion (1 h)

14 h

Elementary methods of data processing

- let's do some data collection: Mathematics A2 ~~in Matlab~~ marks

mark	ν_i	$\frac{\nu_i}{n}$
$\xi_2' = 2$	250	0,5
$\xi_3' = 3$	150	0,3
$\xi_4' = 4$	70	0,14
$\xi_5' = 5$	30	0,06
$\sum$:	500 = n	1

ν_i: frequency

$\frac{\nu_i}{n}$: relative frequency

$$\sum_{i=1}^{n} \nu_i = n = 500$$

$$\sum_{i=1}^{n} \frac{\nu_i}{n} = 1 \quad \square$$

- the average can be calculated:

$$\bar{z} = \frac{1}{n} \cdot \sum_{j=1}^{500} z_j \quad ; \text{ where:}$$

z_j: 2, 2, 5, 3, 2, 4, 2, 5, 4, ...

j: 1, 2, 3, ...

- let's create four sub-intervals, they should be disjoint (non-overlap)

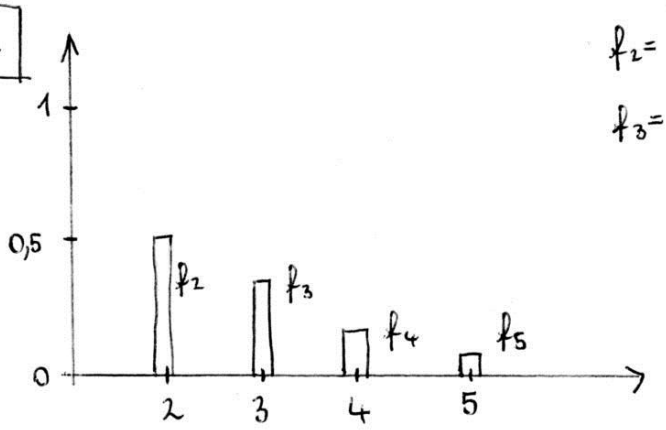
↳ the sum of the frequencies should be equal to the sum of the individuals

$$\bar{z} = \frac{1}{n} \cdot \sum_{j=1}^{500} z_j = \frac{1}{n} \cdot \sum_{i=2}^{5} v_i \cdot z_i = \sum_{i=2}^{5} \frac{v_i}{n} \cdot z_i = 0{,}5 \cdot 2 + 0{,}3 \cdot 3 + 0{,}14 \cdot 4 + 0{,}06 \cdot 5 = 2{,}76$$

changing the index!

- the average carries not enough information, let's show the distribution of the marks:

histogram



$f_2 = \frac{v_2}{n}$ $\quad f_4 = \frac{v_4}{n}$

$f_3 = \frac{v_3}{n}$ $\quad f_5 = \frac{v_5}{n}$

- the mark is a discrete variable, here we get a natural order

- if the variable is non-numerical but a quality or categorical variable (car types, countries, etc.)

↳ how can we construct a histogram? In this case the variables don't define an exact order

↓

PARETO DIAGRAM (Vilfredo Pareto)

- in case of the Pareto diagram we begin with the highest value!

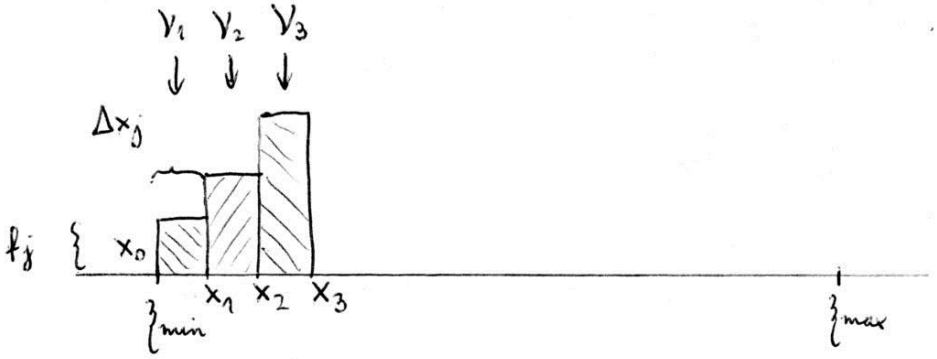
Continous variables

- for creating a histogram the data has to be assigned to groups

data: $\xi_1 < \xi_2 < \dots < \xi_n$ (ordered data)

$\xi_{min} = \xi_{x_0} = \min\{\xi_1, \dots, \xi_n\}$

$\xi_{max} = \xi_{x_{IS}} = \max\{\xi_1, \dots, \xi_n\}$



$I = \xi_{max} - \xi_{min}$ → data range

- creating discrete intervals: $[x_0, x_1)$; $[x_1, x_2)$; $[x_2, x_3)$; ... ; $[x_{IS-1}, x_{IS}]$

IS: the number of intervals

- let's calculate the frequencies: $\nu_1, \nu_2, \dots, \nu_{IS}$

relative frequencies: $\frac{\nu_1}{n}, \frac{\nu_2}{n}, \dots, \frac{\nu_{IS}}{n}$

$f_j \cdot \Delta x_j = \frac{\nu_j}{n}$ ⟼ area of the rectangle

- how can we select IS?

↳ if IS is too high, the histogram gets disturbed

↳ if IS is too low, the nature of the data is spoiled

therefore:

$n \leq 100$ $\quad IS \cong \sqrt{n}$

$n \geq 100$ $\quad IS \cong \log_2 n + 1$

↳ „logarithm (of n) to base two"

n	$\sqrt{n}$	$\log_2 n + 1$
100	10	8
200	14	9
500	22	10
1000	32	11

How should we determine Δx_j?

① uniform subdivision

↳ basic level (secretary, Excel)

② non-uniform subdivision

↳ the goal is, that every Δx_j should contain the same amount of data ($= \nu_j$) (nearly const.)

Example:

body heights ⟼ the students should guess the height of the teacher!

Queues: 170; 170; 173; 173; 175; 175; 175; 176; 177; 178; 178; 181; 183; 185

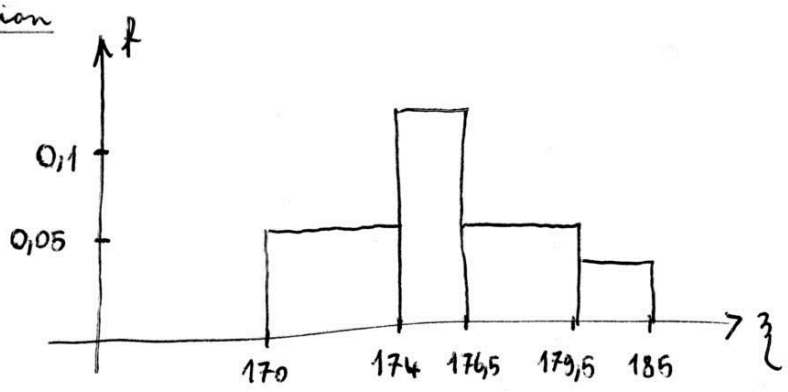
$n = 14$ $\quad x_{min} = 170$

$IS = \sqrt{n} \cong 4$ $\quad x_{max} = 185$

/

x_{Li}	x_{ui}	ν_i	ν_i / n	Δx_i	f_i
[170	174)	4	0,286	4	0,071
[174	176,5)	4	0,286	2,5	0,114
[176,5	179,5)	3	0,214	3	0,071
[179,5	185]	3	0,214	5,5	0,039
	Σ:	14	1		

Empirical density function



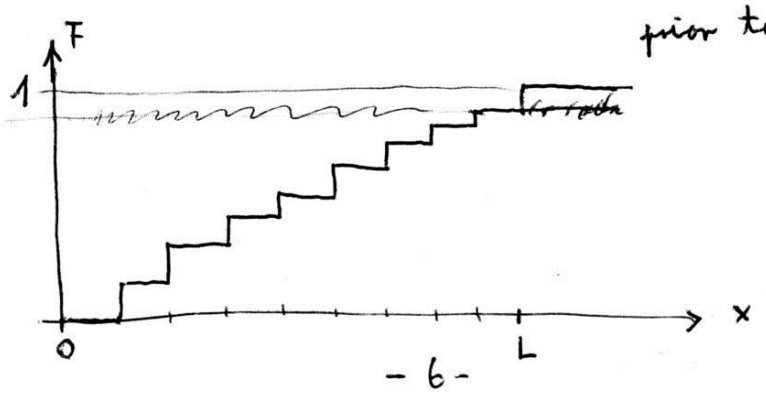
Questions:

- how does the relative frequency behave if $n \to \infty$
- how does the histogram behave if $n \to \infty$ (answers will follow later)

2012.08.06.

Cumulative distribution function (CDF) → empirical!

- what percent of a data-series is below a certain value?
- pulling a homogenous beam:
  - several measurements are carried out → in what percent of the measurements does the beam break for example prior to $\frac{4}{5} \cdot L$?

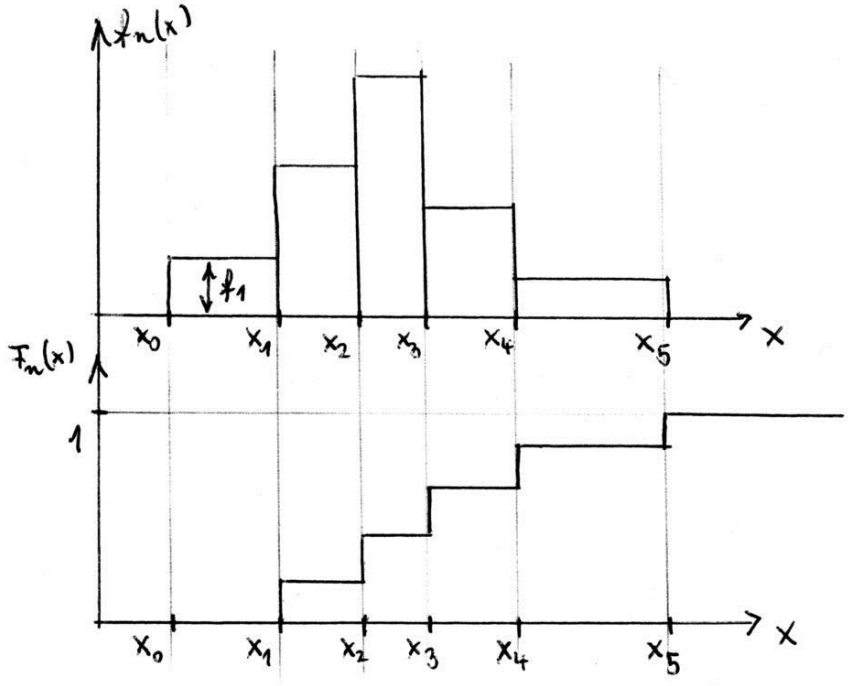


$F_n(x) = \frac{k}{n}$; with $\xi_k \leq x \leq \xi_{k+1}$

$\xi_1 ; \xi_2 ; \dots ; \xi_n$ (sample units)

- the CDF shows the relative frequency of the sample units left from x.

- What is the relationship between the empirical density function and the empirical cumulative distribution ~~fun~~ function?



$x_0 - x_1$	ν_1
$x_1 - x_2$	ν_2
$x_2 - x_3$	ν_3
$x_3 - x_4$	ν_4
$x_4 - x_5$	ν_5

$\Downarrow$

$f_i = \frac{\nu_i}{n \cdot \Delta x_i}$

where:

$$F_n(x_1) = \frac{\nu_1}{n}$$

$$F_n(x_2) = \frac{\nu_1}{n} + \frac{\nu_2}{n}$$

$$\vdots$$

$$F_n(x_5) = \frac{1}{n} \cdot (\nu_1 + \nu_2 + \dots + \nu_5)$$

general formula:

$$F_n(x_k) = \sum_{x_i \leq x_k} f_n(x_i) \cdot \Delta x_i$$

(!) → the sum of the areas left from the actual point

Question:

- what happens with $\left.\begin{matrix} f_n(x) \\ F_n(x) \end{matrix}\right\}$ if $n \to \infty$?

- what kind of relationship is between $F_n(x)$ and $f_n(x)$ if $n \to \infty$?

↓

the answer comes later on!

Median

- the median is the „middle of the sample". Half of the sample is above, half below it.

• let's have a measurement: $\xi_1, \xi_2, \dots, \xi_n$

• arrange the data: $\xi_1^* \le \xi_2^* \le \dots \le \xi_n^*$ „ordered sample"

① if n is odd

$m = \frac{n+1}{2}$; m must be even now

↳ $\tilde{\xi} = \xi_m^*$, the median

~ : „tilde"

② if n is even

$m_1 = \frac{n}{2}$

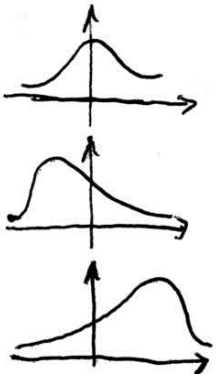
↳ $\tilde{\xi} = \frac{1}{2}\left(\xi_{m1}^* + \xi_{m1+1}^*\right)$

- relationship between median and average:

• if the histogram is about symmetrical ⇒ $\bar{\xi} \approx \tilde{\xi}$

• if —11— leans to the left → $\tilde{\xi} < \bar{\xi}$

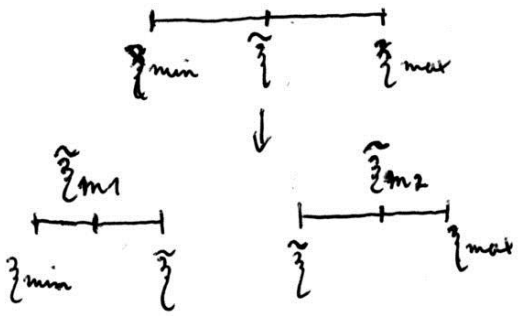
—11— leans to the right → $\tilde{\xi} > \bar{\xi}$



Lower and upper quantiles

- the median splits the sample into two parts (upper and lower)

  ↳ let's determine the medians of the upper and lower parts!



example: $n = 19$

$x_1^* \leq x_2^* \leq \dots \leq x_{19}^*$ ⟼ the median is $\tilde{x} = x_{10}^*$

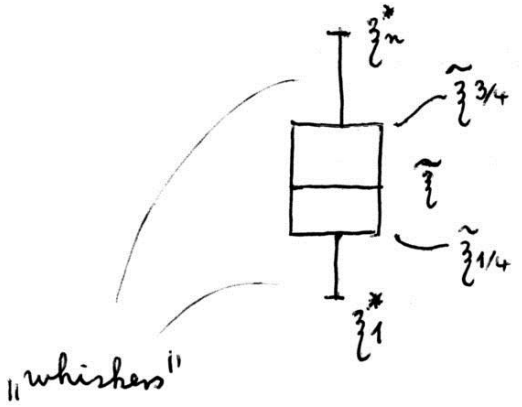
lower quartile: $\tilde{x}_{1/4} = x_5^*$

upper quartile: $\tilde{x}_{3/4} = x_{15}^*$

How can we display the median and quartiles?

Boxplot

- global representation of the measurement sample



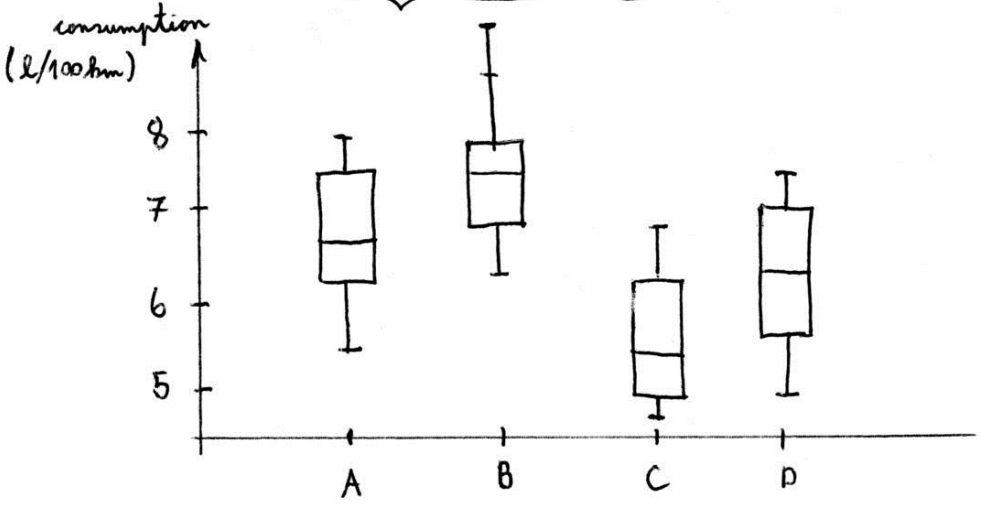
advantage of using box plots:

- different samples can be compared easily

example:

- comparison of fuel consumption of different car types

	$\tilde{x}$	$\tilde{x}_{1/4}$	$\tilde{x}_{3/4}$	x_{max}	x_{min}	
A						
B						
C						
D						



observations from the ~~box~~ box-plot:

- fuel consumption of C is significantly lower, than the others
- ~~type~~ type B can consume very much fuel under certain circumstances

↓

how to deal with the outliers?

Outliers

- definition: „Inter-quartile range" = IQR

$$IQR = \tilde{\xi}_{3/4} - \tilde{\xi}_{1/4}$$

- the data is an outlier if it is further away from the box than 1,5 · IQR

- extreme outlier, if further than 3 · IQR

calculation:

- upper outliers: $\xi > \tilde{\xi}_{3/4} + 1{,}5 \cdot IQR$
- upper extreme outliers: $\xi > \tilde{\xi}_{3/4} + 3 \cdot IQR$
- lower outliers: $\xi < \tilde{\xi}_{1/4} - 1{,}5 \cdot IQR$
- lower extreme outliers: $\xi < \tilde{\xi}_{1/4} - 3 \cdot IQR$

example → we will see it in the practice !!

Standard deviation of the sample

- how can we describe the fluctuation of the data?

variation: $\xi_i - \bar{\xi}$, for every „i" point

let's calculate the average variation:

$$\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \bar{\xi}) = \frac{1}{n}\cdot\sum_{i=1}^{n}\xi_i - \frac{1}{n}\cdot\sum_{i=1}^{n}\bar{\xi} = \bar{\xi} - \frac{1}{n}\cdot n\,\bar{\xi} = 0 \quad \square$$

- something is wrong

↳ the variation $(\xi_i - \bar{\xi})$ might be +/- , therefore the average is zero

- what can we use instead?

$$\sum |\xi_i - \bar{\xi}| \quad \text{or} \quad \boxed{\sum (\xi_i - \bar{\xi})^2}$$

↳ the absolute value is not ideal, because it is difficult to ~~dis~~ differentiate!

<u>therefore:</u>

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2 \quad ; \text{„standard deviation of the sample"}$$

<u>but</u> the formula contains $(n-1)$ independent terms instead of n

/ because $\sum(\xi_i - \bar{\xi}) = \emptyset$ /

⇓

<u>sample standard deviation:</u>

$$s^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2 \quad \text{„bracket"}$$

↳ this is corrected, adjusted

$$s^* = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2}$$

• the unit is the same as the original variable!

2012.09.13.

- a different formula might be used also:

$$s^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2 = \frac{1}{n-1} \sum (\xi_i^2 + \bar{\xi}^2 - 2 \cdot \xi_i \bar{\xi}) \overset{\nabla}{=}$$

$$= \frac{1}{n-1} \cdot \left[\sum \xi_i^2 + n \cdot \bar{\xi}^2 - 2 \cdot \underbrace{\sum \xi_i \bar{\xi}}_{\| \; 2 \cdot n \cdot \bar{\xi}^2} \right] = \frac{1}{n-1} \left(\sum \xi_i^2 - n \cdot \bar{\xi}^2 \right) =$$

$$= \frac{n}{n-1} \left[\frac{1}{n} \cdot \sum \xi_i^2 - \bar{\xi}^2 \right] \geq 0$$

"the avg. of the squares is always higher than the square of the avgs!"

(Schwartz - Cauchy - Bunyakovski)

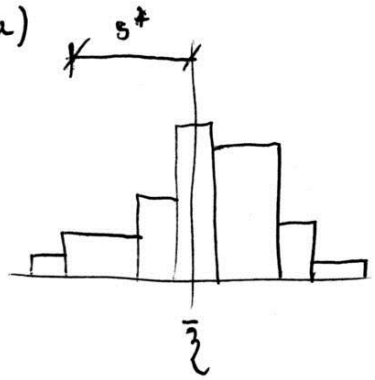
- comparing s^* and $\tilde{\xi}_{1/4}$ i $\tilde{\xi}_{3/4}$:

- few outliers increase s^* significantly, but $\tilde{\xi}_{1/4}$ and $\tilde{\xi}_{3/4}$ are not influenced
- if outliers are present, the box plot gives better information

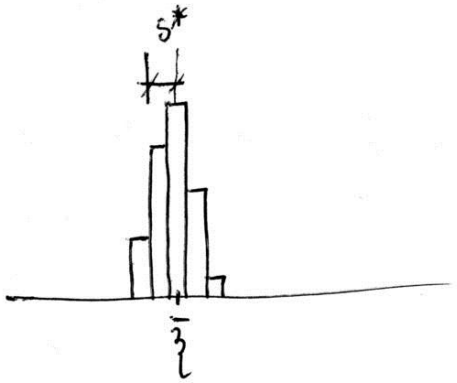
practical use:

$\bar{\xi}$ and s^* describe the measurement data

a)



b)



- let x be the expected value of a production process

(D=20 mm diameter, V=0,5 l beer, 1 kg bread, 500 g detergent)

- let's make measurements; draw a histogram, calculate $\bar{\xi}$ and s^*

if $|\bar{\xi} - x| = H > 0$ ⟶ adjustment problems

if s^* is high ⟶ inaccurate production

Data acquisition

- measurement, observation
- data sampling

(1.) measurement, observation

- globally at similar conditions performed or occured processes, their parameters are recorded.

(2.) data sampling

- the goal is to form a statement on the multitude / population
  - ↳ but we do not want to investigate every single item
    - it would be expensive, time consuming
    - there are cases, where the item is destroyed during investigation (stress test, tensile test)

„roncsolásos vizsgálat"

- a sample should be taken from the population
  - ↳ the statements for the sample should be true for the whole population

- in order to do this, the following points have to be ensured:
  - the population should be homogenous regarding the investigated property (every item should have the same properties)
  - every item should get into the sample with the same probability

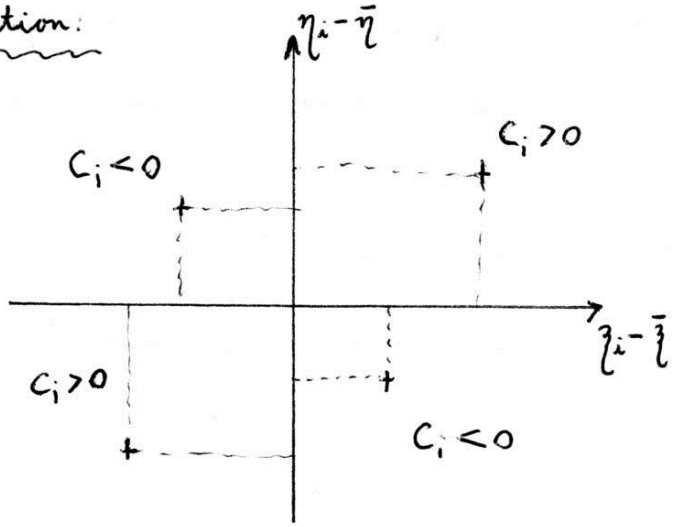
- later on we will state that the sample is a collection of independent items with the same distribution
  - ↳ same distribution is ensured by the correct definition of the population (same production line, same rice)
  - ↳ the identical probability of selection is ensured with random sampling. (either with or without putting the item back)

Empirical correlation coefficient

- to this point we only dealt with one variable
  - ↳ let's assume two variables. Is there a relationship between them?

$\xi_1 ; \xi_2 ; \dots ; \xi_n$
$\eta_1 ; \eta_2 ; \dots ; \eta_n$
} pairing the same points:
$(\xi_1, \eta_1) ; (\xi_2, \eta_2) ; \dots ; (\xi_n, \eta_n)$

graphical representation:



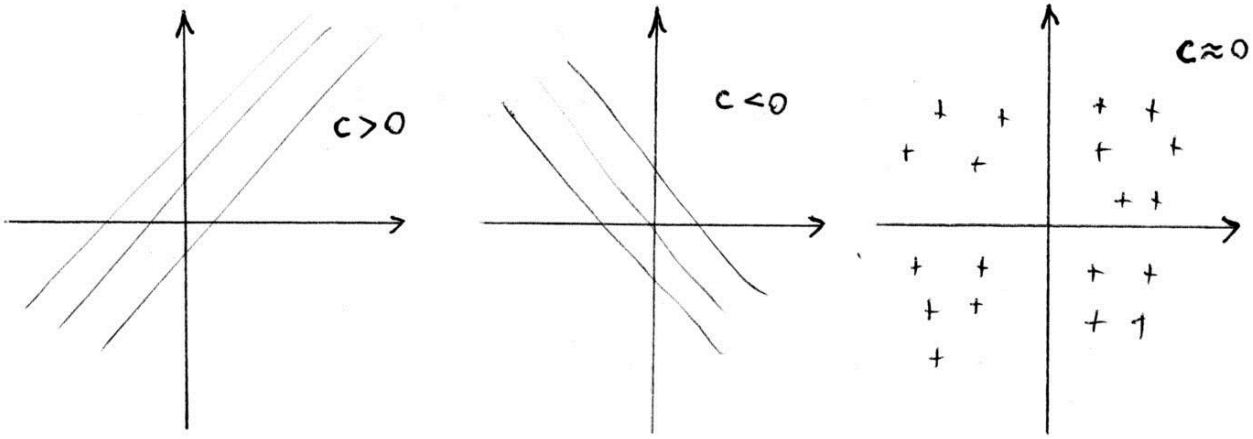
- let's define c_i :

$$c_i = (\xi_i - \bar{\xi})(\eta_i - \bar{\eta})$$

- create the sum:

$$c = \sum_{i=1}^{N} c_i = \sum_{i=1}^{N} (\xi_i - \bar{\xi})(\eta_i - \bar{\eta})$$

- what will we get for c ?



c depends on:

- amount of points : let's take the average, $\frac{1}{N} \cdot c$
- standard deviation of the variables: let's normalize, $\frac{1}{N} \cdot c \cdot \frac{1}{s_\xi} \cdot \frac{1}{s_\eta}$

$$\Downarrow$$

$$\varrho(\xi, \eta) = \frac{\frac{1}{N} \cdot \sum (\xi_i - \bar{\xi})(\eta_i - \bar{\eta})}{s_\xi \cdot s_\eta}$$

$S(\xi,\eta)$ · empirical correlation coefficient

- we can determine that:

  · if $S(\xi,\eta)>0$; the relationship is progressive

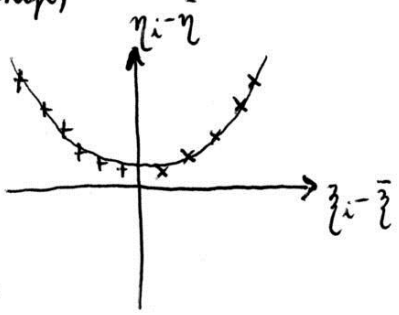
  · if $S(\xi,\eta)<0$; ——— " ——— is degressive

  · if $S(\xi,\eta)\approx 0$; no relationship

questions:

- what happens to $S(\xi,\eta)$ if $n \to \infty$?
- what values can $S(\xi,\eta)$ have? $(|S|\leq 1)$
- does it really represent a general relationship, or less than that?

  ↳ counterexample: $\eta = \xi^2$



$S\approx 0$, but the relationship is obvious!

Rank correlation

- can we simplify the formula for correlation, if the observed variable is discrete?

Example:

- a jury of two judge something (e.g. wines, ice skating, sausage competition)

the contenders: A, B, C, D

	A	B	C	D
ξ: J1	1	4	2	3
η: J2	3	4	1	2

→ ranking of the jury

- question: are the jury in concordance?

jury J1: $\xi_1 ; \xi_2 ; \xi_3 ; \xi_4$

J2: $\eta_1 ; \eta_2 ; \eta_3 ; \eta_4$

} calculate $\bar{\xi} ; \bar{\eta} ; s_\xi ; s_\eta$

$$r(\xi, \eta) = \frac{\frac{1}{N} \cdot \sum (\xi_i - \bar{\xi})(\eta_i - \bar{\eta})}{s_\xi \cdot s_\eta}$$

↑ rank correlation coeff.

if $r \approx 1 \longmapsto$ concordance

if $r \approx -1 \longmapsto$ opposite opinion

- correlation between the ~~valu~~ ranks = rank correlation (Spearman)

- the formula can be simplified:

$$d_1 = \xi_1 - \eta_1$$

$$d_2 = \xi_2 - \eta_2$$

$$\vdots$$

$$\boxed{r = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}}$$

in the example:

$d_1 = 1-3 = -2$

$d_2 = 4-4 = 0$

$d_3 = 2-1 = 1$

$d_4 = 3-2 = 1$

$\sum d_i^2 = 4+0+1+1 = 6$

$r = 1 - \frac{6 \cdot 6}{4(16-1)} = 0{,}4$

- if the jury has more than two members:

2012.08.20.

Kendall's W or Kendall's concordance

$J1, J2, \dots, J_{NJ}$ NJ: number of jurys

$C1, C2, \dots, C_{NC}$ NC: number of contenders

	→ j					
	J1	J2	J3		J_{NJ}	Σ
C1	r_{11}	r_{21}	r_{31}		r_{NJ1}	R_1
C2	r_{12}	r_{22}				R_2
↓ i	r_{13}					
C_{NC}	r_{1NC}				r_{NJNC}	R_{NC}

R_i: the sum of r_{ij} for the „i^{th}" contender ; this is the resultant rank number

$$R_1 = \sum_{j=1}^{NJ} r_{j1} \quad ; \quad R_2 = \sum_{j=1}^{NJ} r_{j2}$$

- let's arrange the resultant rank numbers:

$$R_1^* \leq R_2^* \leq \dots \leq R_{NC}^* \longmapsto \text{final order}$$

$$\bar{R} = \frac{1}{NC} \sum_{i=1}^{NC} R_i \quad ; \text{ let's calculate the squared deviation}$$

$$S^* = \sum_{i=1}^{NC} (R_i - \bar{R})^2$$

- what will be the value of S?

↳ if the members of the jury are in concordance:

• the order is the same:

$$\left.\begin{aligned} R_1^* &= 1 \cdot NJ \\ R_2^* &= 2 \cdot NJ \end{aligned}\right\} \Delta R = NJ !$$

⋮

↳ if there is no concordance:

ΔR is getting smaller! (maybe Ø)

the maximum of S: (if $\Delta R = NJ$)

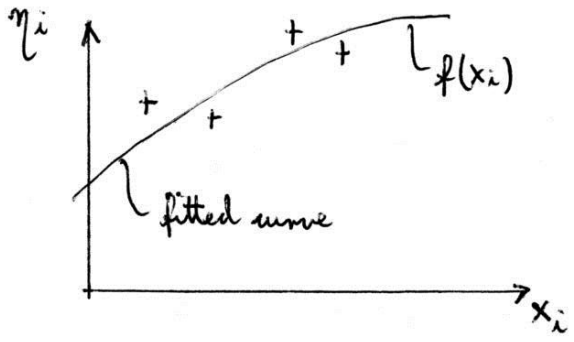
$$S_{max} = \frac{NJ^2 \cdot NC(NC^2-1)}{12}$$

- the actual S should be normalized with S_{max}:

$$W = \frac{12 \cdot \sum_{i=1}^{NC} (R_i - \bar{R})^2}{NJ^2 \cdot (NC^3 - NC)}$$

$0 \leq W \leq 1$; if there is concordance: $W = 1$ (absolute)

Approximation of the relationship



$[x_i, \eta_i]_{i=1}^{n}$ measurement

- the idea comes from Gauss (1777 - 1855) ; around 1795 LNM

(Daniel Kehlmann: Measuring the World)

$$\min \left\{ \sum_{i=1}^{n} [\eta_i - f(x_i)]^2 \right\} \rightarrow \text{that is what we are looking for!}$$

- $f(x)$ should possess some free parameters

for example:

$$f(x) = a_0 + a_1 x + a_2 x^2 \rightarrow \text{non-linear function}$$

let's determine:

$$\min\left\{\sum_{i=1}^{n}\left(\eta_i - a_0 - a_1 x_i - a_2 x_i^2\right)^2\right\} = D(a_0, a_1, a_2)$$

$$\left\{\eta_i ; x_i\right\}_{i=1}^{n} \text{ is given}$$

- let's assume that there exists a local minimum, the derivative there is zero.

$$\frac{\partial D}{\partial a_0} = 0 \quad ; \quad \frac{\partial D}{\partial a_1} = 0 \; ; \quad \frac{\partial D}{\partial a_2} = 0$$

↳ 3 unknowns, 3 equations

$$\frac{\partial D}{\partial a_0} = \sum_{i=1}^{n} 2\cdot\left(\eta_i - a_0 - a_1 x_i - a_2 x_i^2\right)(-1) = 0$$

$$\frac{\partial D}{\partial a_1} = \sum_{i=1}^{n} 2\cdot\left(\eta_i - a_0 - a_1 x_i - a_2 x_i^2\right)(-x_i) = 0$$

$$\frac{\partial D}{\partial a_2} = \sum_{i=1}^{n} 2\cdot\left(\eta_i - a_0 - a_1 x_i - a_2 x_i^2\right)(-x_i^2) = 0$$

Equation:

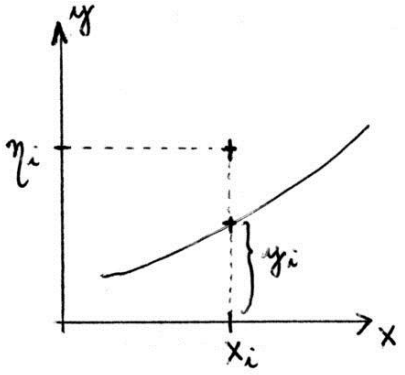
$$\sum \eta_i - n\cdot a_0 - a_1 \cdot \sum x_i - a_2 \sum x_i^2 = 0$$

$$\sum x_i \eta_i - a_0 \cdot \sum x_i - a_1 \cdot \sum x_i^2 - a_2 \sum x_i^3 = 0$$

$$\sum x_i^2 \eta_i - a_0 \sum x_i^2 - a_1 \cdot \sum x_i^3 - a_2 \sum x_i^4 = 0$$

a_0, a_1, a_2 can be determined (Excel uses the same method)

Coefficient of determination



- total standard deviation of η

$$\frac{1}{n}\sum(\eta_i - \bar{\eta})^2 = s_{tot}^2$$

- residual standard deviation

$$\frac{1}{n}\sum(\eta_i - y_i)^2 = s_{err}^2$$

coeff. of determination:

$$R^2 = \frac{s_{tot}^2 - s_{err}^2}{s_{tot}^2}$$

$R^2 = 1$, if $s_{err} = 0$

$R^2 = 0$, if $s_{tot} = s_{err}$

- if $\eta_i = y_i$, then the curve passes through the measurement points

↳ in this case: $s_{err} = 0 \Rightarrow R^2 = 1$ ∎

- if $y_i = \bar{\eta}$, then the curve is selected to be the average of the points

↳ in this case $s_{err}^2 = s_{tot}^2 \Rightarrow R^2 = 0$ ∎

⇓

the coeff. of determination shows the „goodness" of the fitting

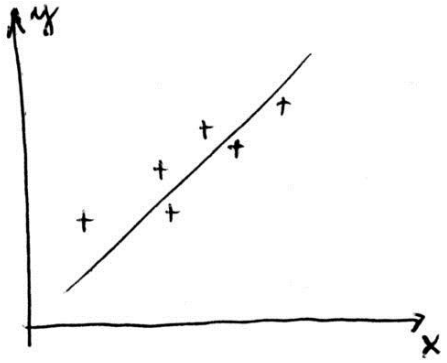
- the nearer is R^2 to 1 the nearer goes the curve to the points

Caution in the interpretation

- if $R^2 = 1$, thus $\eta_i = y_i$, the fitted curve passes through the points, this is not regression but interpolation

↳ this will not describe the process properly!

it can be shown:



$$r(xy,x) = \frac{\frac{1}{N}\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y}$$

$$R^2 = r^2(y,x)$$

Definitions to this point: (with the posed questions)

- relative frequency $\frac{V}{n}$ what if $n \to \infty$?
- average $\frac{1}{n}\sum \xi_i$ —11—
- empirical density function —11—
- empirical cumulative distribution function (CDF) —11—
- relationship between $f_n(x)$ and $F_n(x)$, if $n \to \infty$
- sample standard deviation s^* —11—
- empirical correlation coefficient ϱ —11—

/.

Probability theory

- history: • the Greek did not deal with the probability theory
  - 17th century: Pascal and Fermat (basics)
  - beginning of the 20th century: Kolmogorov

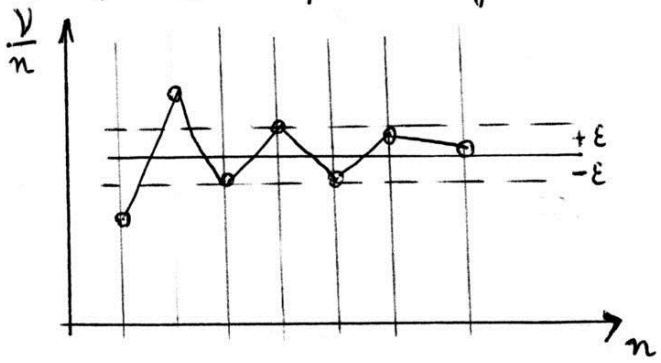
random event

(clearly)

- the result of an event (experiment, etc.) is not unequivocally determined by the circumstances taken into account.

  ↳ this is incidence

- relative frequency → probability



("successful")

ν: frequency (number of events)

n: number of all events

↳ relative frequency variates randomly

↳ even as $n \to \infty$, the rel. frequ. „stabilizes". The probability, that it jumps out from a $+\varepsilon/-\varepsilon$ interval, decreases

probability of an event

A: event (e.g. to throw a dice and get a five)

$P(A) = p \qquad 0 \leq p \leq 1$

$P(A) = 1$: certain event

$P(A) = 0$: impossible event

deterministic variable

example: $V = \frac{s}{t}$, it does not depend on coincidence

random variable

- the circumstances do not determine the variable unequivocally.

  taken into account

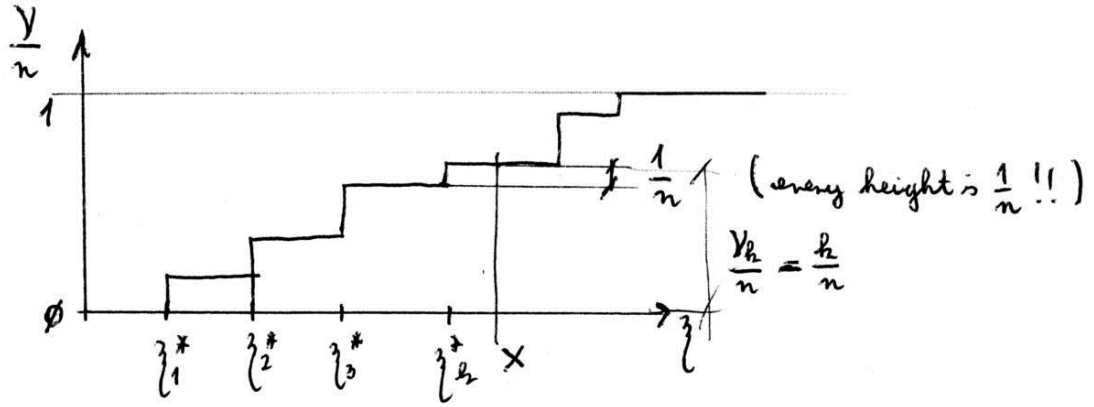
- it can be discrete or continous

examples: dice , roulette , result of a measurement , body height

the empirical cumulative distribution function ⟶ cumulative distribution function

ordered observations:

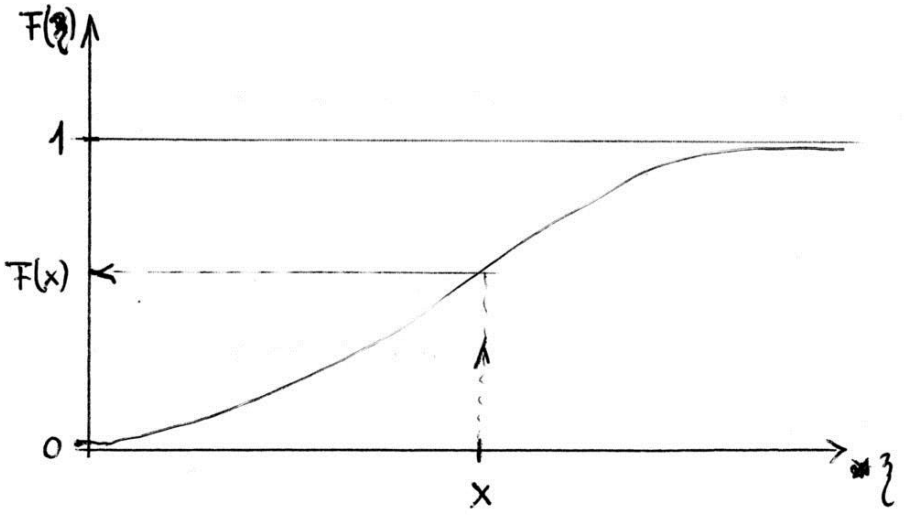
$\xi_1^* \,/\, \xi_2^* \,/ \cdots / \,\xi_n^*$ (discrete variable!)



- for an arbitrary x there belongs „k" observations, wich are less than „x":

$$F_n(x) = \frac{k}{n}$$

- if $n \longrightarrow \infty$, then $F_n(x) \longrightarrow F(x)$



event „A":

$A: \xi < x$

$P(A) = P(\xi < x) = F(x)$

Properties:

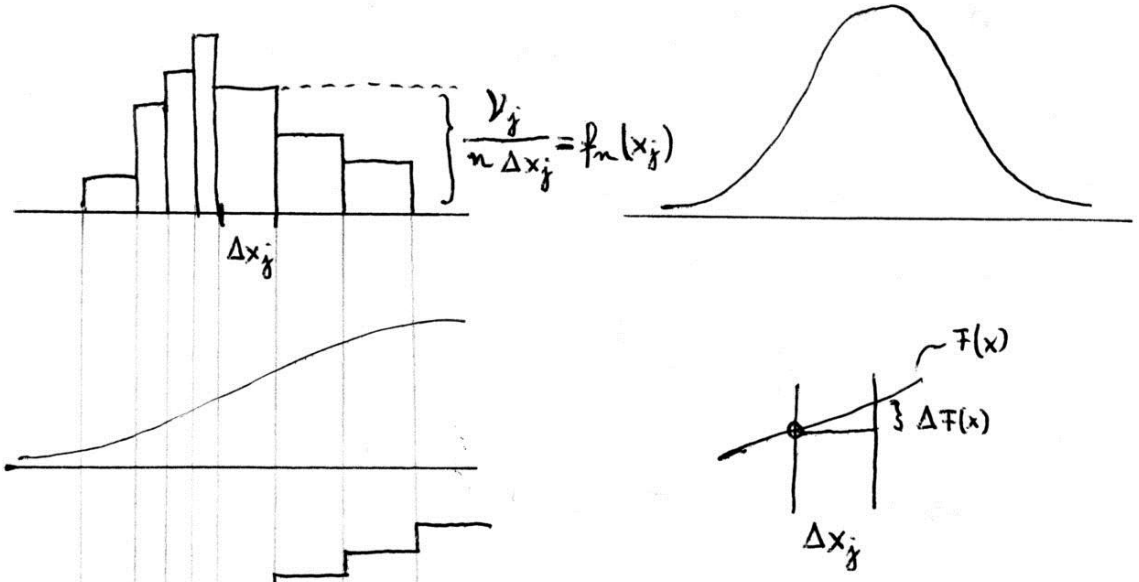
$F(x) \to 0$; if $x \to -\infty$

$F(x) \to 1$; if $x \to +\infty$

probability of belonging to an interval:

$$\left.\begin{aligned} P(\xi < a) &= F(a) \\ P(\xi < b) &= F(b) \end{aligned}\right\} \quad P(a < \xi < b) = F(b) - F(a)$$

empirical density function ⟶ density function



$$\frac{\Delta F(x)}{\Delta x} \cong \frac{F(x+\Delta x) - F(x)}{\Delta x} = \frac{\frac{\nu_j}{n}}{\Delta x}$$

$$F_n(x_j) = \sum_{x<x_j} \frac{\nu_j}{n} = \sum \underbrace{\frac{\nu_j}{n\,\Delta x_j}}_{f_n(x_j)} \cdot \Delta x_j = \sum f_n(x_j)\,\Delta x_j$$

if $n \to \infty$

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

properties of the density function

- $f(x) \longrightarrow \emptyset$; if $x \longrightarrow \pm\infty$
- $f(x) \geq \emptyset$
- $P(a \leq \xi \leq b) = \int_a^b f(x)\,dx$
- $f(x) = \frac{dF}{dx}$

examples: - normal
- uniform

average $\longrightarrow$ mean value

$\xi_1, \xi_2, \dots, \xi_n$ ← number of intervals!!

$$\bar{\xi} = \frac{1}{n} \cdot \sum_{i=1}^{n} \xi_i = \frac{1}{n} \cdot \sum_{j=1}^{K} x_j \nu_j = \frac{1}{n} \cdot \sum_{j=1}^{K} x_j \frac{\nu_j}{\Delta x_j} \cdot \Delta x_j = \sum x_j \underbrace{\frac{\nu_j}{n \cdot \Delta x_j}}_{f_n(x_j)} \Delta x_j =$$

$$= \sum x_j\, f_n(x_j)\,\Delta x_j \xrightarrow[n\to\infty]{} \int x\, f(x)\,dx$$

("first-order moment")

definition of the mean value:

$$M(\xi) = \int x\, f(x)\,dx$$

properties

$$M(\xi+\eta) = M(\xi) + M(\eta)$$

$\eta = a\xi + b$ $\qquad M(a\xi+b) = a \cdot M(\xi) + b$ $\qquad M(a) = a$

sample standard deviation ⟶ standard deviation (and variance)

$$s^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2$$

↓

$D^2(\xi) = \sigma^2 = M[(\xi - M(\xi))^2]$, variance (mean squared difference)

$\sigma = +\sqrt{D^2(\xi)}$, standard deviation

properties

- ξ ; $D^2(\xi) = \sigma_\xi^2$

$\eta = a\xi + b \longrightarrow D^2(\eta) = D^2(a\xi + b) = a^2 D^2(\xi) + D^2(b) = a^2 \cdot \sigma_\xi^2$ //. ∅ *

- ξ, η

$D^2(\xi+\eta) = D^2(\xi) + D^2(\eta)$ if ξ and η are independent

but:

$D(\xi+\eta) \neq D(\xi) + D(\eta)$

Important:

in case of a random variable ξ, $D(\xi)$ and $M(\xi)$ are constant values (not random variables anymore)!
(deterministic)

Variance of the average

$$\bar{\xi} = \frac{1}{n} \Sigma \xi_i$$

$$D^2(\bar{\xi}) = D^2\left(\frac{1}{n} \Sigma \xi_i\right) = \frac{1}{n^2} D^2\left(\Sigma \xi_i\right) = \frac{1}{n^2} \Sigma D^2(\xi_i) = \frac{n \cdot D^2(\xi)}{n^2} = \frac{\sigma^2}{n}$$

/. 2012.10.04.

Standardication

$$\xi \;;\; m = M(\xi) \;;\; D^2(\xi) = \sigma_\xi^2$$

- standardized variable:

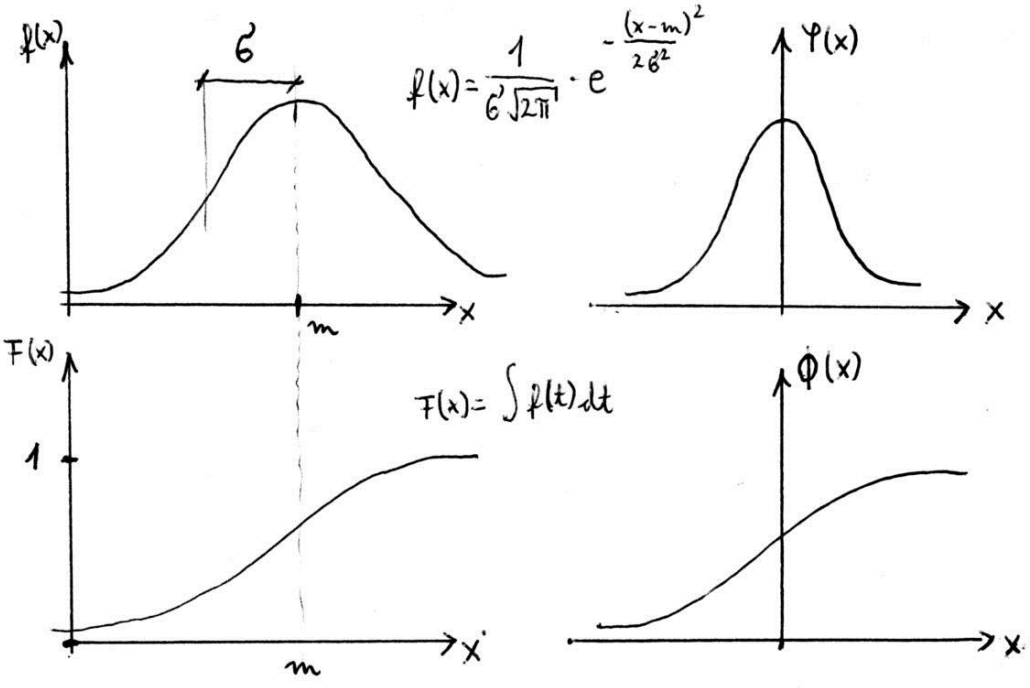
$$\eta = \frac{\xi - m}{\sigma} \qquad (\text{standard score})$$

$$M(\eta) = M\left(\frac{\xi - m}{\sigma}\right) = \frac{1}{\sigma} \cdot M(\xi - m) = \frac{1}{\sigma} \cdot (M(\xi) - M(\xi)) = 0$$

$$D^2(\eta) = D^2\left(\frac{\xi - m}{\sigma}\right) = \frac{1}{\sigma^2} \cdot D^2(\xi - m) = \frac{D^2(\xi)}{\sigma^2} = 1$$

$$\Downarrow$$

the standardized variable has a mean value of zero and a variance of one. $\left(\begin{matrix} M(\eta) = 0 \\ D(\eta) = 1 \end{matrix}\right)$



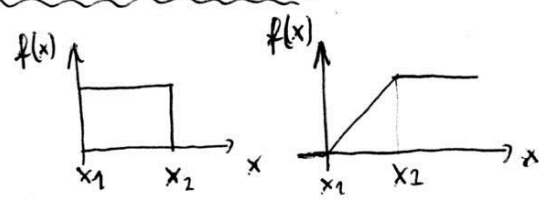
normal distribution

$N(m, \sigma)$

standard normal distribution

$N(0,1)$

uniform distribution:

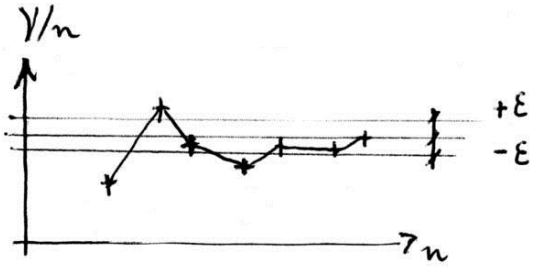


Law of large numbers

- in case of an experiment which was carried out several times:

relative frequency ⟶ probability

$$\lim_{n\to\infty} P\left(\left|\frac{\nu}{n} - p\right| \geq \varepsilon\right) = \emptyset$$



- the convergence is stochastic

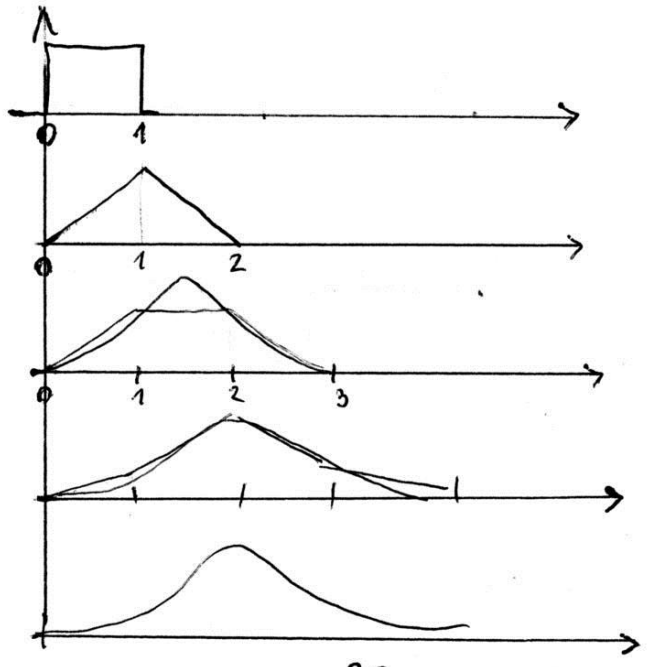
$$\bar{\xi} \xrightarrow[st.]{} M(\xi)$$

Central limit theorem (CLT)

- let's take "n" independent random variables, the distribution of their sum will be a normal distribution

$$\lim_{n\to\infty} P\left(\frac{\xi_1+\xi_2+\ldots+\xi_n - n\cdot m}{\sigma\cdot\sqrt{n}}\right) = \phi(x)$$

example: ξ_i , $i=1,2,\ldots,n$; the distribution of each variable is uniform



Reminder:

- data sampling

  ↳ the data sample should be a collection of independent items with the same distribution

  ↳ the data sample should be "sufficiently large"

  ↳ a data sample can be described:

  - average
  - sample standard distribution
  - histogram

Estimation

- the parameters (mean, variance, etc.) are calculated from the sample

estimator:

- the value which estimates the particular parameter

2012.10.11

Properties

(1) unbiased estimation

actual parameter: a

estimator: $\hat{a}$

- if $M(\hat{a}) = a$, then the estimation is unbiased
- the average is is the unbiased estimator of the mean:

$$M(\bar{\xi}) = M\left(\frac{1}{n} \cdot \sum \xi_i\right) = \frac{1}{n} \cdot \sum M(\xi_i) = \frac{1}{n} \cdot n \cdot M(\xi) = M(\xi)$$

- standard deviation, sample standard deviation

$$M(s^2) \neq D^2(\xi) \quad \text{but} \quad M(s^{*2}) = D^2(\xi)$$

② consistent estimation

d_i: estimator calculated from a sample of „i" elements (units)

$d_1, d_2, \ldots, d_n$

- the estimation is consistent if „d_i" tends to „a" with increasing n

③ efficient estimation

d_1 estimation
d_2 estimation
} d_1 is the more efficient estimator if:

$$D^2(d_1) < D^2(d_2)$$

Measurement errors

- measurement errors and error propagation
- the results of a measurement: ξ_i random variable
- every measurement has an actual value: X

$$\xi = X + \varepsilon$$

ε: the error of the measurement

↳ this might be an error in the measuring system, or the noise of the observed value

(example: 1 liter bottles)

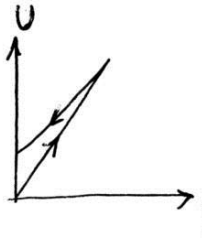
$$M(\xi) = M(X + \varepsilon) = M(X) + M(\varepsilon) = X + M(\varepsilon)$$

$$M(\varepsilon) = \Delta x \begin{cases} = \emptyset \rightarrow \varepsilon \text{ is a random error} \\ \neq \emptyset \rightarrow \Delta x \text{ is a systematic error} \end{cases}$$

$\sigma = \varepsilon - \Delta x$ is the random error

Systematic error

- the systematic error of the measuring device or the measuring method
- examples:
  - zero error (incorrect zeroing, displaced scale, etc.)
  - hysteresis of the device



  - dynamic behaviour of the device (for example an unsteady process is measured with a static device)

Random error

- several effects which cannot be taken into account influence the measurement (temperature of the air, voltage of power supply, material properties, etc.)
- since $M(\varepsilon) = \emptyset \Rightarrow M(\xi) = X$
- the evaluation of the measurement is performed with statistical methods (confidence intervals, hypothesis testing, regression)

Error propagation

- classification of measurements:

① direct measurements

- the measured quantity is directly compared to the etalon (standard)
  - ↳ a device representing the unit of the quantity

(2.) indirect measurements

- the required quantity is calculated from measured quantities

  ↳ torque: $M = k \cdot F$ (k = "lever arm")

  power: $P = U \cdot I$

- every measured quantity has an error (systematic, random)

  ↓

  these errors „propagate" through the calc. process

2012.10.18.

1.) Propagation of systematic errors

- different quantities: x_1 ; x_2 ; ... ; x_n

  sys errors: Δx_1 ; Δx_2 ; ... ; Δx_n

- the value of y has to be calculated:

$$y = f(x_1, x_2, \dots, x_n)$$

- how can we determine the systematic error of y? ✱

$$\Delta y = ?$$

$$\Delta y = f(x_1 + \Delta x_1, x_2 + \Delta x_2, \dots, x_n + \Delta x_n) - f(x_1, x_2, \dots, x_n)$$

↳ if every Δx_i is known, this can be calculated!

↳ but: how do each Δx_i influence the sys. error of y?

↓

if we want to decrease Δy, which Δx_i should be reduced?

(i.e.: what parameter should be measured more accurately)

How does the value of Δx_i influence Δy?

(first order Taylor series)

- Let us create the Taylor series for $y + \Delta y$:

$$y + \Delta y \cong f(x_1, x_2, \dots, x_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i + \dots$$

only the linear part is kept!!

$$H_y = \Delta y \cong \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i$$

H_y: the resultant systematic error (absolute error)

$\frac{\partial f}{\partial x_i}$: sensitivity parameter

↳ this shows in what extent each Δx_i influences the resultant systematic error

the maximum of the res. sys. error: (absolute error)

$$\Delta y_{max} = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \cdot |\Delta x_i|$$

relative systematic error:

$$H_{rel} = \frac{\Delta y}{y} \cong \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot \frac{\Delta x_i}{y}$$

specific case:

$$y = C \cdot x_1^{k_1} \cdot x_2^{k_2} \cdot \dots \cdot x_n^{k_n}$$

$$\frac{\partial f}{\partial x_i} = C \cdot x_1^{k_1} \cdot x_2^{k_2} \cdot \dots \cdot k_i \cdot x_i^{k_i - 1} \cdot \dots \cdot x_n^{k_n} = k_i \cdot \frac{y}{x_i}$$

therefore:

$$\frac{\Delta y}{y} = \sum_{i=1}^{n} k_i \cdot \frac{\Delta x_i}{x_i}$$

short example:

the volume of a cylinder is calculated as follows:

$$V = \frac{D^2 \pi h}{4} = \frac{\pi}{4} \cdot D^2 \cdot h$$

①

$$H_{rel} = \frac{\Delta V}{V} = 2 \cdot \frac{\Delta D}{D} + 1 \cdot \frac{\Delta h}{h} \quad //$$

②

$$\frac{\partial V}{\partial D} = \frac{\pi}{2} \cdot D \cdot h \qquad \text{i} \qquad \frac{\partial V}{\partial h} = \frac{\pi}{4} \cdot D^2$$

$$H_{abs} = \Delta V = \frac{\pi}{2} \cdot D \cdot h \cdot \Delta D + \frac{\pi}{4} \cdot D^2 \cdot \Delta h \qquad / : V = \frac{D^2 \pi}{4} \cdot h$$

$$\frac{\Delta V}{V} = 2 \cdot \frac{\Delta D}{D} + \frac{\Delta h}{h} \quad //$$

---

2. Propagation of random errors

$x_1, x_2, \dots, x_n$

random variables $\Rightarrow$ y is also a random variable

$$\varepsilon_i = X - x_i$$

- let us use the previous formula, instead of Δx_i we have ε_i

$$\boxed{\varepsilon_y \cong \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \varepsilon_i}$$

calculating the mean:

$$M(\varepsilon_y) = M\left(\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot \varepsilon_i\right) = \sum_{i=1}^{n} M\left(\frac{\partial f}{\partial x_i} \cdot \varepsilon_i\right) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} M(\varepsilon_i) = \emptyset$$

$\Downarrow$

therefore ε_y is also random type error

- the variance of ε_y:

$$D^2(\underbrace{y}_{\text{actual value}} + \varepsilon_y) = D^2(\varepsilon_y) = D^2\left(\sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\varepsilon_i\right) = \sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2 \cdot D^2(\varepsilon_i)$$

$$\sigma_y^2 = \sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2$$

$\hookrightarrow \sigma_y = \sqrt{\sigma_y^2}$; standard deviation

- relative variance:

$$\left(\frac{\sigma_y}{y}\right)^2 = \sum_{i=1}^{n}\left(\frac{\partial f}{\partial x_i}\right)^2 \cdot \left(\frac{\sigma_i}{y}\right)^2$$

* special case: $y = C \cdot x_1^{k_1} \cdot x_2^{k_2} \cdot \ldots \cdot x_n^{k_n}$

$$\left(\frac{\sigma_y}{y}\right)^2 = \sum_{i=1}^{n}\left(k_i \frac{\sigma_i}{x_i}\right)^2$$

- in case of a measurement the variance / st. dev. is unknown

$\hookrightarrow$ the sample standard deviation (s^*) is calculated

$\hookrightarrow$ the formulas above are used with „s^*"

example:

- volume of a cylinder; $V = \frac{D^2\pi}{4} \cdot h$

- measuring diameter $(D_1, D_2, \ldots, D_N)$ and height $(h_1, h_2, \ldots, h_N)$

$$\bar{D} = \frac{1}{N}\sum_{j=1}^{N} D_j \longmapsto s_D^* \qquad s_{\bar{D}}^* = \frac{s_D^*}{\sqrt{N}}$$

$$\bar{h} = \frac{1}{N}\sum_{j=1}^{N} h_j \longmapsto s_h^* \qquad s_{\bar{h}}^* = \frac{s_h^*}{\sqrt{N}}$$

- but it is better to calculate each $V_i = \frac{D_i^2 \pi}{4} h_i$

(if we don't want to investigate the effect of the components)

Confidence interval

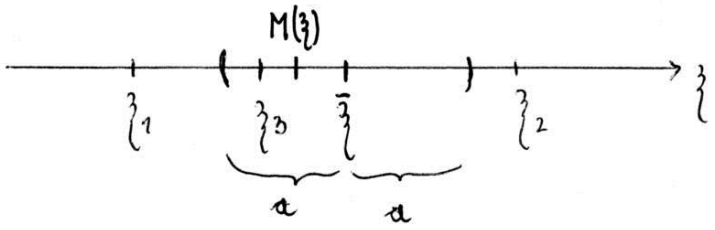
- let us assume that we deal with random errors

- a measurement is carried out: $\xi_1, \xi_2, \dots, \xi_n$

  • the average $\bar{\xi}$ can be calculated

question:

how precisely do we estimate the mean $M(\xi)$ using the average?



remember: $\bar{\xi} \to M(\xi)$ $\quad n \to \infty$

- let us create an interval around $\bar{\xi}$! The average is in the middle, the radius is „a"

- the interval should contain $M(\xi)$.

$M(\xi) \in [\bar{\xi} - a\,;\, \bar{\xi} + a]$ with a probability of „p"

- how is „a" determined?

$$P\left(\bar{\xi} - a \le M(\xi) \le \bar{\xi} + a\right) = p$$

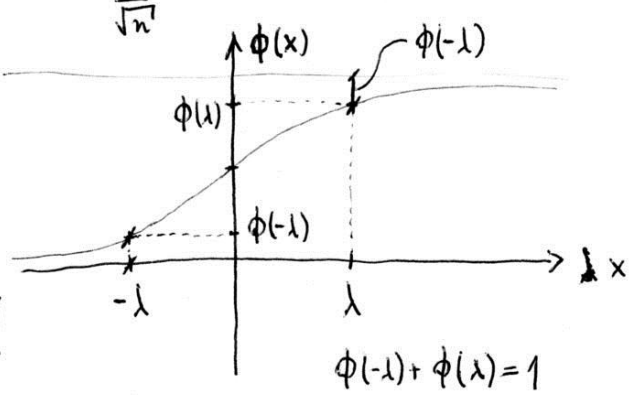
$$\eta \in N(0,1)$$

$$P\left(-a \le \underbrace{\bar{\xi} - M(\xi)}_{} \le +a\right) = P\left(-\underbrace{a \cdot \frac{\sqrt{n}}{\sigma}}_{\lambda} \le \frac{\bar{\xi} - M(\xi)}{\frac{\sigma}{\sqrt{n}}} \le +a \cdot \frac{\sqrt{n}}{\sigma}\right)$$

let us create the standard score of the avg.!

$$M(\bar{\xi}) = M(\xi)$$

$$D(\bar{\xi}) = \frac{D(\xi)}{\sqrt{n}}$$



$$P(-\lambda \le \eta \le \lambda) = p = \Phi(\lambda) - \Phi(-\lambda) = 2\Phi(\lambda) - 1$$

$$\Phi(\lambda) = \frac{p+1}{2} \quad ; \quad \lambda = \Phi^{-1}\left(\frac{p+1}{2}\right)$$

example: if the significance level is $p = 95\%$, then $\lambda = 1{,}96$

$$\lambda = a \cdot \frac{\sqrt{n}}{\sigma} \Rightarrow \boxed{a = \lambda \cdot \frac{\sigma}{\sqrt{n}}}$$, the radius of the confidence interval

$\bar{\xi} \pm a$; $\bar{\xi} - a$: lower confidence limit

$\bar{\xi} + a$: upper confidence limit

- what if the st. deviation is unknown? (This is usually the case)

• let us calculate the sample st. dev.:

$$s^{*2} = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{\xi} - \xi_i)^2$$

st in this case we have $\eta^* = \dfrac{\bar{\xi} - M(\xi)}{\frac{s^*}{\sqrt{n}}} \notin N(0,1)$

instead:

Student distribution (William Gosset, worker of the Guiness brewery)

- Student's „t" distribution

⇓

λ_{St} instead of λ, therefore:

$$a = \lambda_{St} \cdot \frac{s^*}{\sqrt{n}}$$

STATISTICAL HYPOTHESIS TESTING