VŠB - Technical University of Ostrava
Faculty of Electrical Engineering and Computer Science
Department of Applied Mathematics

STATISTICS I.

Radim Briš
Petra Škňouřilová

Ostrava 2007

INSTRUCTIONS FOR EDUCATION

STATISTICS I

The lecture notes are divided into fractions (chapters) that correspond to logical dividing of studied subject matter. Large chapters are divided into numbering subchapters. Any subchapter has got this structure.

Study time

Time that is of needed to understanding matters. The time is orientation and it can serve as guide for study layout.

Aim

Then are introduced the aims which do u have achieve after work up these chapter – concrete knowledge and acquirements.

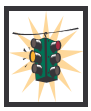
Explication

Follows personal interpretation studied matters, introduction new notions, and their explication, everything accompanied by buckthorn examples.

Summary of notions

In conclusion are rerun main notions which do you have develop yourself. If some of them you don't understand yet return towards them once more.

Study guide



Solved example

Questions

To be really sure that you completely understand discussed problems you got several theoretical questions here. Results of these tasks are mostly mentioned in brackets or they can be found at the end of textbook in KEYS TO SOLUTIONS.

Problems

In the end practical tasks for solution are presented.

1 EXPLORATORY DATA ANALYSIS

Study time: 70 minutes

Aim

- general notions of exploratory (preliminary) statistics
- data variable types
- statistical characteristics and methods of graphical presentation qualitative variables
- statistical characteristics and methods of graphical presentation quantitative variables

Explication

Original goal of statistics was to acquire data about population based on a sampling population. By population we mean a group of all existing elements which we observe during statistical research. For example:

> *If we perform a statistical research about 15 years old girls altitude by population we mean all girls currently 15.*

Considering the fact that usually a number of elements in populations is high we perform a research on so called **sample examination** where we use only part of the population instead of complete one. Examined part of population is called **the sample**. What's important is to define really representative selection.

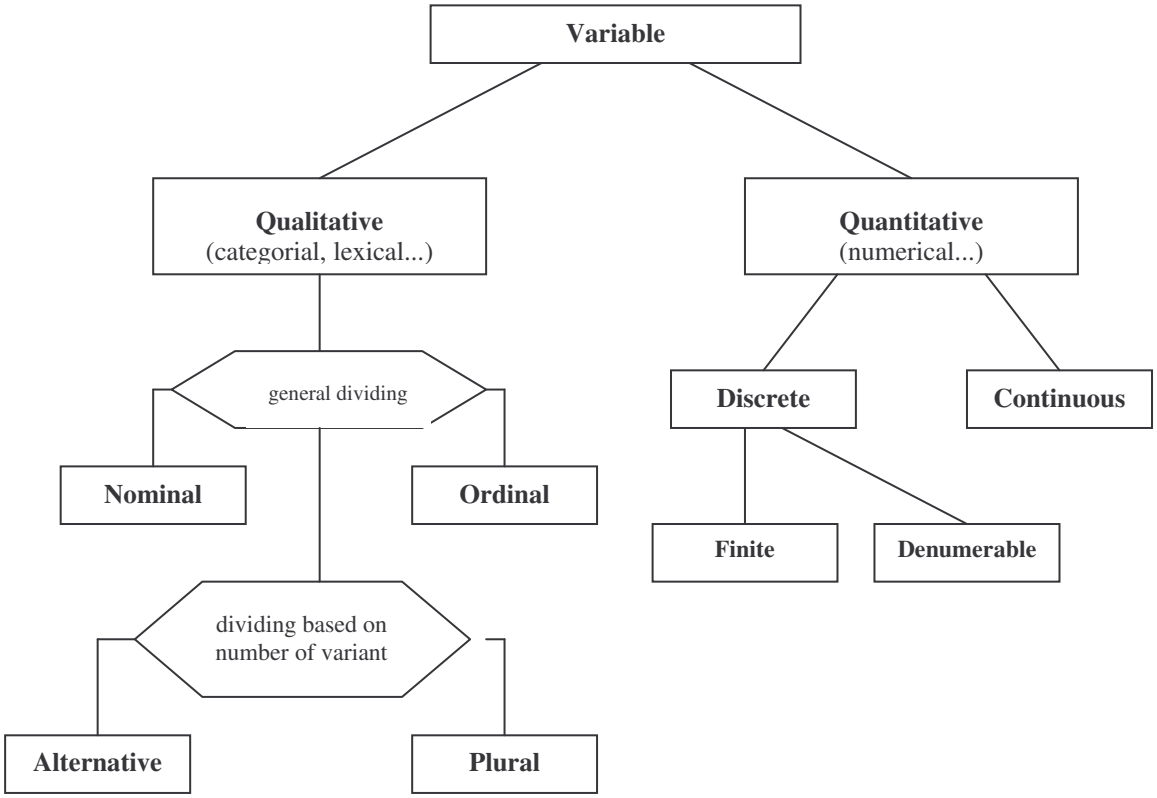
There are several ways how to make this selection. To avoid the omittance of some elements of population we choose so called **random sample** in which each element of population has a same chance to be selected.

It's obvious that sample examination can't ever be as correct as exploration of all population. Why we do prefer it?

1. Reduction of time and financial cost (especially in large population)
2. Destructive testing (some tests – cholesterol contended in blood etc.– lead to the destruction of examined elements)
3. Inaccessibility of all population

Now u know that statistics can describe all population based on knowledge gathered from population sample. Now we move on to exploratory data analysis (EDA). Data we observe will be called **the variables** and their values **variable variants**. **EDA** is often a first step in revealing information hidden in large amount variables and their variants.

Considering the fact that the way of variables processing depends most of all on their type we now make ourselves familiar with a basic dividing of variables into different categories. This dividing is presented on following image.



- **Qualitative variable** – its variants are expressed verbally and it's divided into two general subgroups according to reference between particular values:
    - **Nominal variable** – has equivalent variants: it is impossible compare them nor sort them (for example: sex, nationality ...)
    - **Ordinal variable** – it forms pass between qualitative and quantitative variables: individual variant can be sort and it is possible compare each other (for example: clothing size (S, M, L, XL))

    Second way of dividing is dividing based on number of variants:

    - **Alternative variable** – it has only two various variants (e.g. sex - male, female; ...)
    - **Plural variable** – it has more than two various variants (e.g. education, name, eye color, ...)

- **Quantitative variable –** it is expressed numerically and it's divided into:
    - **Discrete variable** – it has finite or denumerable number of variants

- **Discrete finite variable –** it has finite number of variants (e.g. mark from math - 1,2,3,4,5)

- **Discrete denumerable variable –** it has denumerable number of variants (e.g. age (year), height (cm), weight (kg), ...)

- **Continuous variable -** it has any value from $\Re$ or from some subset $\Re$ (e.g. distance between cities, ...)

Study guide

Imagine a situation when we got a large statistical group at our disposal and you face a question how to best describe it. Numbers of values with which we "replace" such a large group describe a basic attributes of this group and we shall call it ***statistical characteristics****.*

In following chapters we shall learn how to set statistical characteristics for various types of variables and how to represent the larger statistical groups.

1.1 Statistical characteristics of qualitative variables

We know that qualitative variable has two basic types - nominal and ordinal.

1.1.1 Nominal variables

Nominal variable has different but equivalent variants in one group. Number of these variants is usually low and that's why the first statistical characteristics we use to describe it will be its frequency.

- **Frequency n_i** (absolute frequency)

  - is defined as number of occurrence variant of the qualitative variable

  In case that qualitative variable has *k* different variants (we describe their frequency n_1, n_2… n_k) in the statistical group (n values large) it must hold true:

$$n_1 + n_2 + ... + n_k = \sum_{i=1}^{k} n_i = n$$

If we want express what part of the group forms variables with any variant we use relative frequency for description of variable.

- **Relative frequency p_i**

  - is defined as:

$$p_i = \frac{n_i}{n}$$

eventually:

$$p_i = \frac{n_i}{n} \cdot 100 \quad [\%]$$

(We use second formula in case if we want express of the relative frequency in percents).
It must hold true for relative frequency:

$$p_1 + p_2 + \ldots + p_k = \sum_{i=1}^{k} p_i = 1$$

When qualitative variables are processed it is good to order frequency and relative frequency into so-called **frequency table**:

FREQUENCY TABLE		
Values x_i	**Absolute frequency**	**Relative frequency**
	n_i	**p_i**
x_1	n_1	p_1
x_2	n_2	p_2
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	p_k
Total	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$

The last characteristic for nominal variable is mode.

- **Mode**

  - is defined as a variant name that have for the variable the most frequency

  The mode represented a typical element of the group. We don't determine mode in case that there is more variants with maximum frequency in the statistical group.

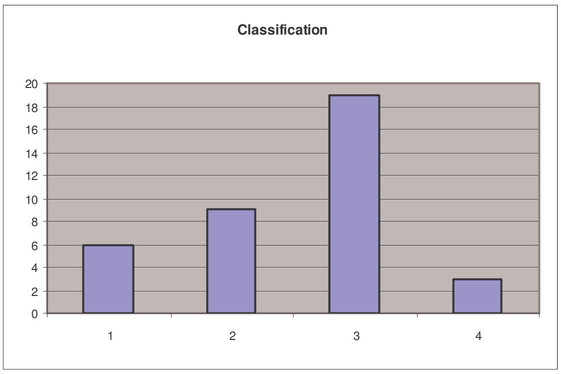
1.1.2 Graphical presentation qualitative variables

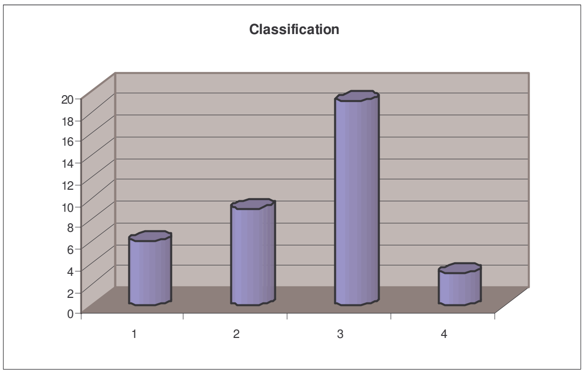
The statistics often use **graphs** for better plasticity of variables analysis. They are these two types for nominal variable:

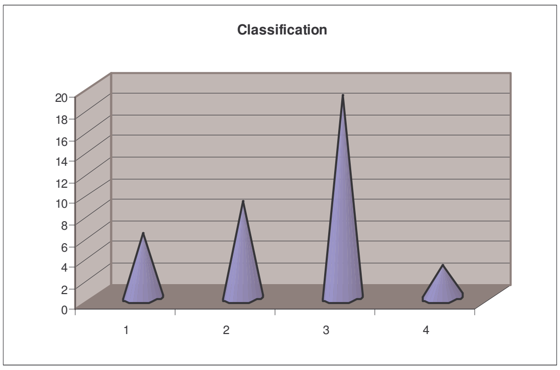
- **Histogram** (bar chart)
- **Pie chart**

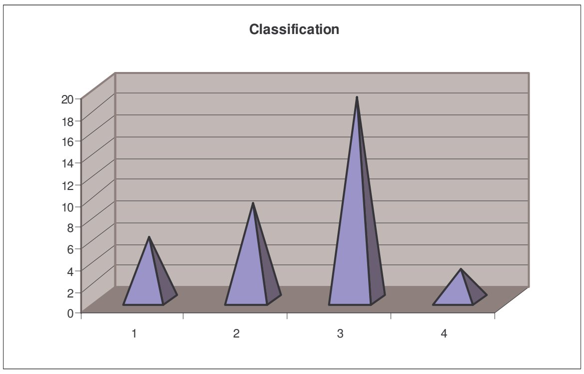
Histogram is a classical graph whereof we take variants of the variable on one axis and variable frequencies on the second one. Individual values of the frequency are then displayed as bars (boxes or vectors, squared logs, cones ...)

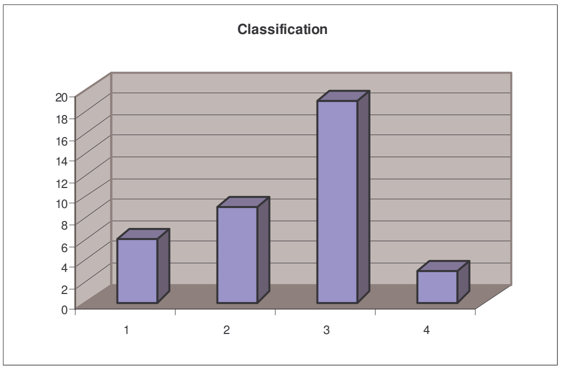
Examples:



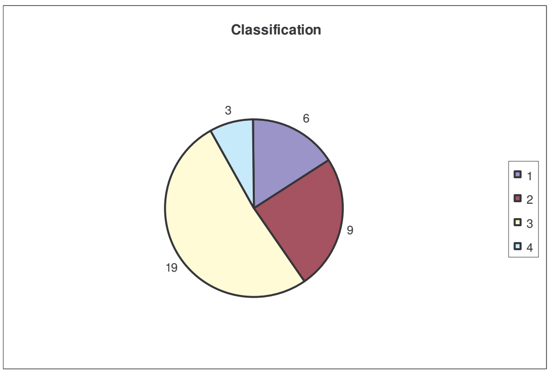


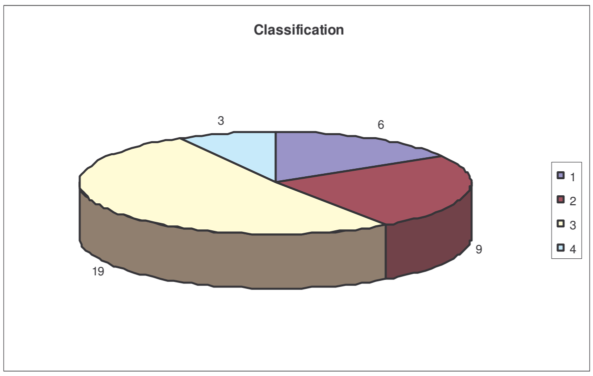


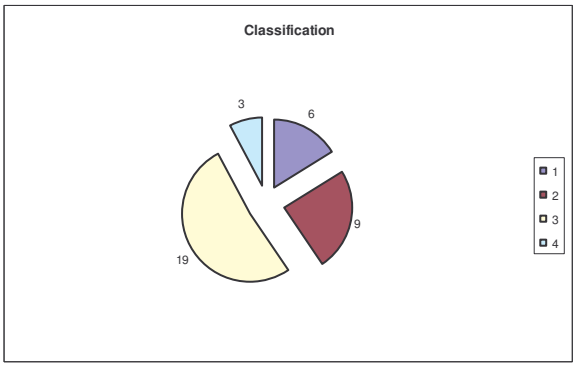


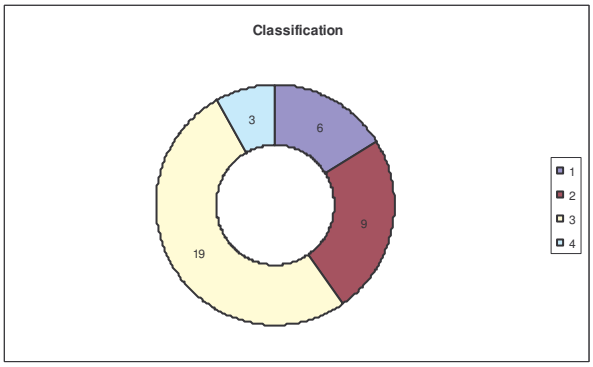


Pie graph represents relative frequencies of the individual variants of the variable. Individual relative frequencies are proportionally represent as a sector of a circle (when we change a circle to an ellipse we obtain three-dimensional effect).



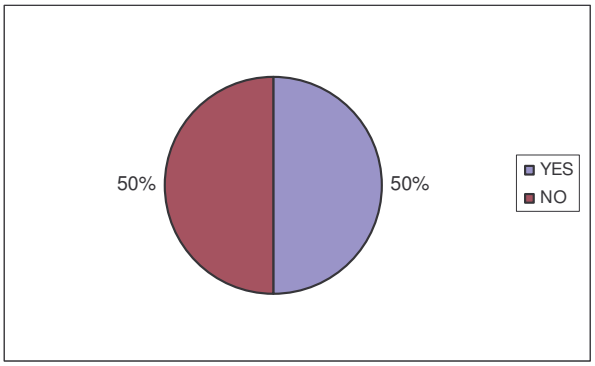




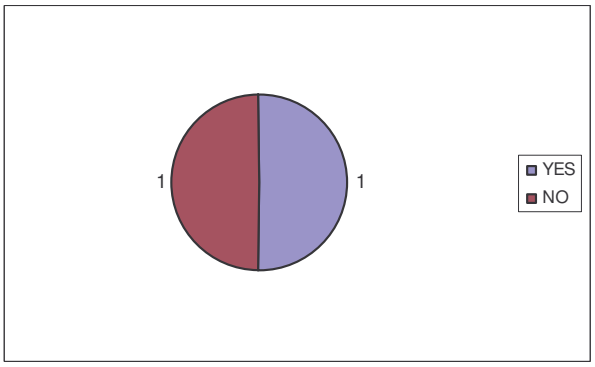


ATTENTION!!! There we must take care of graph description in the pie charts. Marking individual sectors by relative frequencies without adding their absolute frequencies is not sufficient.

Example: We executed enquiry appertain to implementation high school fee. Following chart presents results:

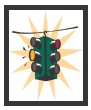


These are interesting results, aren't they? But they are true. Now we modify the chart the way it was recommended:



What do you think now? From the second chart we see that we asked two people - the first one said YES and the second one NO. So what have we discovered? Create only such charts

as their interpretation was absolutely perceptible. If we obtain a pie chart without absolute frequencies ask whether it is an author nescience or it is his purpose.



Solved example

We made crossroad usage research. Obtained data are in following table. They represent color of cars that pass through crossroad. Analyze these data and represent results in graphical form.

red	blue	red	green
blue	red	red	white
green	green	blue	red

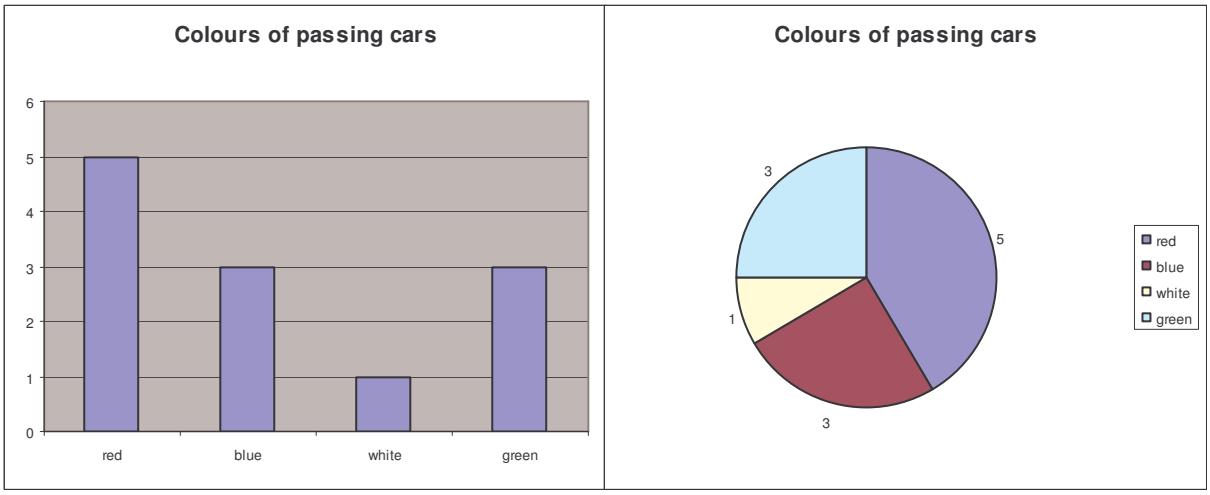
Solution:
It's obvious that it's a qualitative (lexical) variable and considering the fact that there is no point in ordering or comparing colors of cars we can say it's a nominal variable.

For its descriptions we choose frequency table we determine mode and color of passing vehicles we represent by histogram and pie graph.

FREQUENCY TABLE		
Colors of passing cars	**Absolute frequency**	**Relative frequency**
	n_i	p_i
red	5	5/12 = 0,42
blue	3	3/12 = 0,25
white	1	1/12 = 0,08
green	3	3/12 = 0,25
Total	12	1,00

We observed 12 cars total.

Mode = red (i.e. in our sample there were mostly red cars)



1.1.3 Ordinal variable

Now we will continue to ordinal variable description. Ordinal variable (as well nominal variable) has various lexical variants into group but these variants can be sort i.e. we can define which variant is "smaller" or "bigger".

For description ordinal variable we use same statistical characteristics and graphs such as for description nominal variable (frequency, relative frequency, mode + histogram, pie graph) extended two others characteristics (cumulative frequency and cumulative relative frequency) expressing sorting of ordinal variable.

- **Cumulative frequency of i-th variant m_i**

  - it's a number of values of variable showing the frequency of variants less or equal i-th variant

  E.g. we have a variable "classification from statistics". That has these variants: "1", "2", "3" or "4". Then for example cumulative frequency for variant "3" will be equal number of students who got classification "3" or better.

  If there are individual variants sort by their "size" ("$x_1 < x_2 < \ldots < x_k$") then it must holds true:

  $$m_i = \sum_{j=1}^{i} n_j$$

  So it's obvious that cumulative frequency k-th ("the highest") variant is equal measure of the variable - n.

  $$m_k = n$$

The second special characteristic for ordinal variable is cumulative relative frequency.

- **Cumulative relative frequency of i-th variant F_i**

  - a part of group are values gaining i-th and lower variant. It is expressed by this characteristic.

  $$F_i = \sum_{j=1}^{i} p_j$$

  This is nothing else then relative expression of the cumulative frequency:

  $$F_i = \frac{m_i}{n}$$

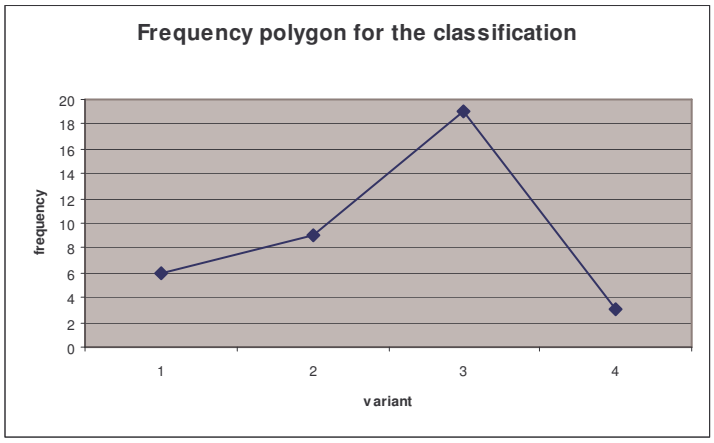
As well as at nominal variable we can present statical characteristics using frequency table at ordinal variable. It contains comparing with frequency table of nominal variable also values of cumulative and cumulative relative frequencies.

FREQUENCY TABLE				
Values x_i	**Absolute frequency**	**Cumulative frequency**	**Relative frequency**	**Relative cumulative frequency**
	n_i	**m_i**	**p_i**	**F_i**
x_1	n_1	$m_1 = n_1$	p_1	$F_1 = p_1$
x_2	n_2	$m_2 = n_1 + n_2 = m_1 + n_2$	p_2	$F_2 = p_1 + p_2 = F_1 + p_2$
⋮	⋮	⋮	⋮	⋮
x_k	n_k	$m_k = n_{k-1} + n_k = n$	p_k	$F_k = F_{k-1} + p_k = 1$
Total	$\sum_{i=1}^{k} n_i = n$	-----	$\sum_{i=1}^{k} p_i = 1$	-----

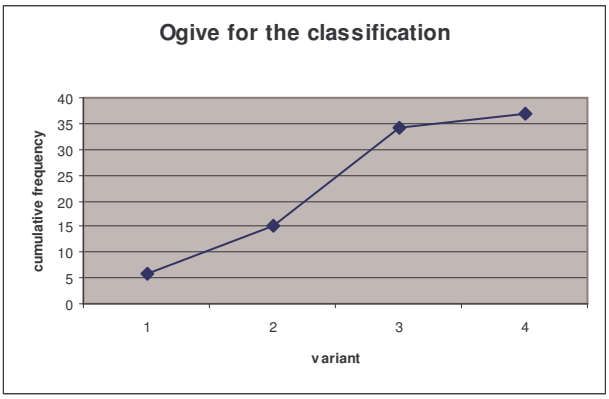
1.1.4 Graphical presentation ordinal variables

We made a mention of the histogram and the pie graph for graphical presentation of the ordinal variable. But these graphs don't reflect sorting of the individual variants. With this we have at command frequency polygon (or ogive) and Pareto graph.

Frequency polygon
- it's a line graph. The frequency is placed along the vertical axis and the individual variants of the variable are placed along the horizontal axis (from "the smallest" till "the highest"). These points are connected with lines.

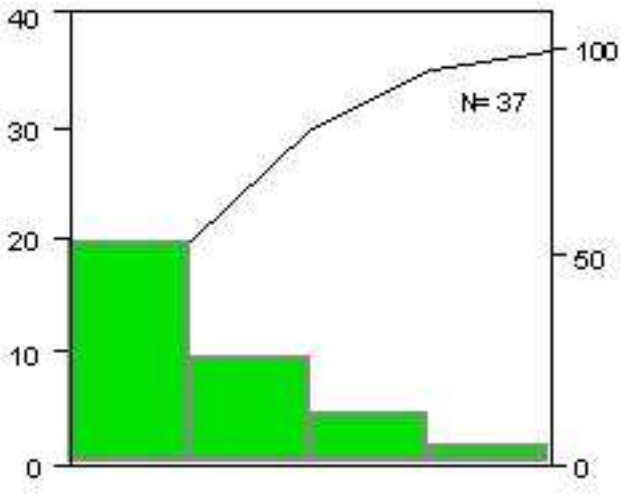


Ogive (cumulative frequency polygon)
- it's a frequency polygon of the cumulative frequency or the relative cumulative frequency. The vertical axis is the cumulative frequency or relative cumulative frequency. The horizontal axis represents possible variants. The graph always starts at zero at the lowest variant and will end up at the total frequency (for a cumulative frequency) or 1.00 (for a relative cumulative frequency).

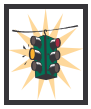


Pareto graph

- it's a bar graph for qualitative variable with the bars arranged according to frequency
- there are particular variants on horizontal axis ordered from the one with "the biggest" importance to "the smallest one"



Consider the decline of cumulative frequency polygon. It's lower as frequency on individual variables drops.



Solved example

Following data represent size of the t-shirts that were sell in sale of the company CLOTHES.

S, M, L, S, M, L, XL, XL, M, XL, XL, L, M, S, M, L, L, XL, XL, XL, L, M

a) Analyze these data and represent results in graphical form.
b) Determine how much percent of people bought t-shirt L maximal value.

Solution:

a) The variable is qualitative (lexical) and t-shirts size can be sort therefore it's an ordinal variable. For its description we use frequency table for the ordinal variable and we determine a mode.

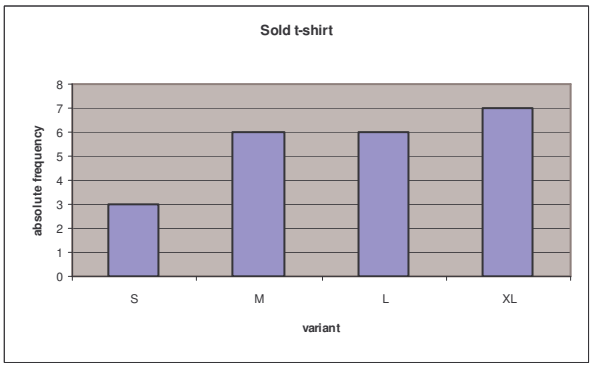
FREQUENCY TABLE				
T-shirt size	**Absolute frequency**	**Cumulative frequency**	**Relative frequency**	**Relative cumulative frequency**
	n_i	m_i	p_i	F_i
S	3	3	$3/22 = 0{,}14$	$3/22 = 0{,}14$
M	6	$3+6=9$	$6/22 = 0{,}27$	$9/22 = 0{,}41$
L	6	$9+6=15$	$6/22 = 0{,}27$	$15/22 = 0{,}68$
XL	7	$15+7=22$	$7/22 = 0{,}32$	$22/22 = 1{,}00$
Total	22	-----	1,00	-----

Mode = XL (the most people bought t-shirt XL value)

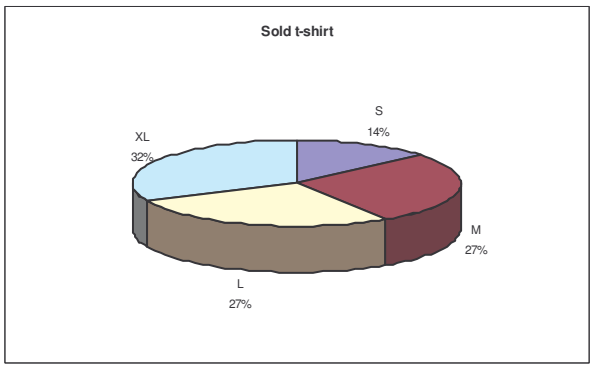
Graphical output will be histogram, pie graph and cumulative frequency polygon (we don't create pareto graph because we haven't got a technical data).

Graphical output:

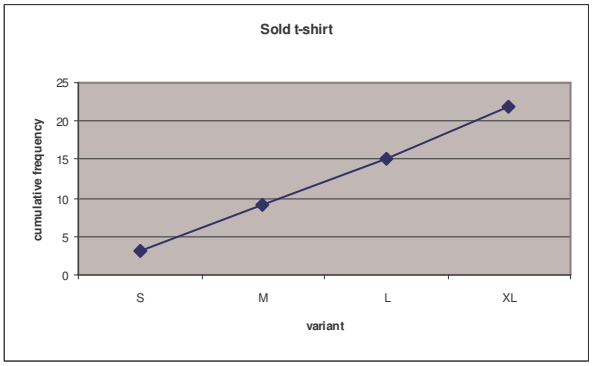
histogram



pie graph



cumulative frequency polygon



Total selling was 22 t-shirts.

b) Answer to this question we get from value of the relative cumulative frequency for variant L. We see that 68% of people bought t-shirt L size and smaller.

1.2 Statistical characteristics of quantitative variables

For description quantitative variable we can use most of the statistical characteristics that are used for ordinal variable description (frequency, relative frequency, cumulative frequency and cumulative relative frequency). With these characteristics we add another two characteristics:

- **measures of position** – those indicate a typical distribution of the variable values (dislocation on the numerical axis)

  and

- **measures of variability** – those indicate a variability (variance) of the values round their typical position

1.2.1 Measures of position and variability

The most used measure of position is a mean of variable. The mean represents average or typical value of the sampling population. The most famous mean for quantitative variable is:

- **Arithmetical mean** $\bar{x}$

  Its value we obtain by means of this formula:

$$\bar{x} = \frac{\sum_{i=!}^{n} x_i}{n}$$

  where: x_i ... particular values of the variable
  n ... size of the sampling population (number of the values of the variable)

  Properties of the arithmetical mean:

1. $\sum_{i=!}^{n} (x_i - \bar{x}) = 0$,

   - sum of all diversions of variable values from their arithmetical mean is equal to zero what means that arithmetical mean compense influence of random errors on variable

2. $\forall (a \in \Re): \left(\bar{x} = \frac{\sum_{i=!}^{n} x_i}{n} \Rightarrow \frac{\sum_{i=!}^{n} (a + x_i)}{n} = a + \bar{x} \right)$

- if we add a same number to all values of the variable, the arithmetical mean increase about this number too

3. $\forall (b \in \Re): \left(\bar{x} = \frac{\sum_{i=!}^{n} x_i}{n} \Rightarrow \frac{\sum_{i=!}^{n} (bx_i)}{n} = b\bar{x} \right)$

- if we multiple all values of the variable a same number, the arithmetical mean increase the same way

For calculation of sampling population mean the arithmetical mean is not always the best solution. For example if we work with a variable representing relative changes (cost indexes,...) we use so-called geometrical mean. For calculation of mean in cases when variable has a character of unit's part (problems about common work ...) we use harmonical mean.

Considering that mean is set from all variable values it carries maximum information about sampling population. On the other hand it's very sensitive to so-called **outlier observations** what are values which are extraordinary different from others and they can diverge mean as much that it's not representing sampling population any more. To identify outlier observations we shall return later.

Among measures of position that are less dependent on outlier observations belong:

- **Mode** $\hat{x}$

  In case of mode we will discern between discrete and continuous quantitative variable. **For discrete variable** we define **mode** $\hat{x}$ as value of the most frequency of the variable (analogous to by the qualitative variable).

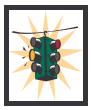
  But by the **continuous variable** we think of mode $\hat{x}$ as value around which is the most concentration of variable values.

  For assessment of this value we use **shorth**. Short is the shortest interval whereof lies at least 50% of variable values. In case of sample large as $n = 2k\ (k \in \mathrm{N})$ (even number of values) k values lies in short - what is n/2 (50%) variable values. In case of sample large as $n = 2k+1\ (k \in \mathrm{N})$ (odd number of values) $k+1$ values lies in short - what is about 1/2 more then 50% variable values (n/2+1/2).

  Then we define **mode** $\hat{x}$ as centre of the short.

  From said it results that short length (top boundary - bottom boundary) is unambiguously given but that's not applied to its location nor its mode.

If mode can be determined unambiguously we talk about **unimode variable** when variable has two modes we call it **bimode**. When there are two or more modes in a sample it usually signalizes a heterogenity of variable values. This heterogenity can be removed by dividing sample into more subsamples (for example bimode mark person's height can be divided according to sex into two unimode marks - women's height and men's height).



Solved example

The following data represent age of the musicians which played on the concert. The variable age is a continuous. Determine mean, short and mode for the variable.

22	82	27	43	19	47	41	34	34	42	35

Solution:

a) Mean:

In this case we use arithmetical mean:

$$\bar{x} = \frac{\sum_{i=!}^{n} x_i}{n} = \frac{22+82+27+43+19+47+41+34+34+42+35}{11} = 38{,}7 \ \ year$$

Average age is 38,7 year for musician played on the concert.

b) Shorth:

Our sample population has 11 values. 11 is odd number of values. 50% of this is 5,5 and the nearest higher natural number is 6 - otherwise: n/2+1/2 = 11/2+1/2 = 12/2 = 6. There out imply that 6 values will be lies in the shorth.

And following advance?

- we sort variable
- we determine size of all intervals (having 6 elements) in which $x_i < x_{i+1} < \ldots < x_{i+5}$
- the shortest of these intervals is shorth (size of the interval = $x_{i+5} - x_i$)

Original data	Sorting data	Size of intervals (having 6 elements)
22	19	16 (= 35 – 19)
82	22	19 (= 41 – 22)
27	27	15 (= 42 – 27)
43	***34***	***9*** (= 43 – 34)
19	***34***	13 (= 47 – 34)
47	***35***	47 (= 82 – 35)
41	***41***	
34	***42***	
34	***43***	
42	47	
35	82	

From table we see that the shortest interval has size 9. The only one interval corresponds to this size: $\langle 34;43 \rangle$.

Shorth = $\langle 34;43 \rangle$. This we will interpret as half of musicians are 34 to 43 years old.

c) Mode:

Mode is defined as center of shorth:

$$\hat{x} = \frac{34 + 43}{2} = 38{,}5$$

Mode = 38,5 year, i.e. typical age is 38,5 year for musician played on the concert.

Other characteristics for description quantitative variable are **quantiles**. Those serve for more detailed illustration of distribution of the variable values within the scope of the population.

- **Quantiles**
  Quantiles are characteristics which describe of location of individual values (within the scope variable). The quantiles are resistant to outlier observation analogous to mode. Generally the quantile is define as value which divide sample into two parts - the first one contain values that are less than given quantile and the second one contain values that are bigger or equal than given quantile. We must have got sorted data (from the least to the biggest value).

  Quantile of variable x which separates 100% lesser values from rest of sample (i.e. from 100(1-p)% values) we call **100p % quantile** and we mark it $\mathbf{x_p}$.

  In work we most often meet these quantiles:

  - **Quartiles**

  When division is into four parts the values of the variate corresponding to 25%, 50% and 75% of the total distribution are called quartiles.

Lower quartile $x_{0,25}$ = 25% quantile (it divides a sample of data so that 25% of values is less than this quartil, i.e. 75% is bigger (or equal))

Median $x_{0,5}$ = 50% quantile (it divides a sample of data so that 50% of values is less than median and 50% of values is bigger (or equal))

Upper quartile $x_{0,75}$ = 75% quantile (it divides a sample of data so that 75% of values is less than this quartil, i.e. 25% is bigger (or equal))

Example:

Data	6 47 49 15 43 41 7 39 43 41 36
Ordered Data	6 7 15 36 39 41 41 43 43 47 49
Median	41
Upper quartile	43
Lower quartile	15

The difference between the 1st and 3rd quartiles is called the **inter-quartile range (IQR)**.

$$IQR = x_{0,75} - x_{0,25}$$

Example:

Data	2 3 4 5 6 6 6 7 7 8 9
Upper quartile	7
Lower quartile	4
IQR	7 - 4 = 3

- **Deciles** – $x_{0,1;}\ x_{0,2;}\ \ldots\ ;\ x_{0,9}$

  The deciles divide the data into 10 equal regions.

- **Percentiles –** $x_{0,01};\ x_{0,02};\ \ldots;\ x_{0,99}$

  The percentiles divide the data into 100 equal regions.

  For example, the 80th percentile is the number which has 80% below it and 20% above it. Rather tan counting 80% from the bottom, count 20% from the top.

  Note: The 50th percentile is the median.

- **Minimum x_{min} and Maximum x_{max}**

  $x_{\min} = x_0$, i.e. 0% of values are less than minimum

$x_{max} = x_1$, i.e. 100% of values are less than maximum

The quantiles we determine by means of the following process:

1. The sample population we order by size

2. Of the individual values we assign the sequence so that the least value will be at first place and the highest value will be at n-th place (n is number of values)

3. 100p% quantile is equal of variable value with sequence z_p where: $z_p = n \cdot p + 0{,}5$
   We round z_p to integer number !!!!!

ATTENTION!!!!
When we have even number of data median is not uniquely defined. *Any number between two middle values (including these values) can be taken as median. The most often we take middle of these values.*

Now we talk about **relation** between **quantiles and cumulative relative frequency**. The value p denotes cumulative relative frequency of quantile x_p i.e. relative frequency of those variable values that are lesser than quantile x_p. Quantile and cumulative relative frequency are inverse notions.

Graphical or tabular representation of the ordered variable and appropriate cumulative frequencies is designated as **distribution function of the cumulative frequency** or **empirical distribution function**.

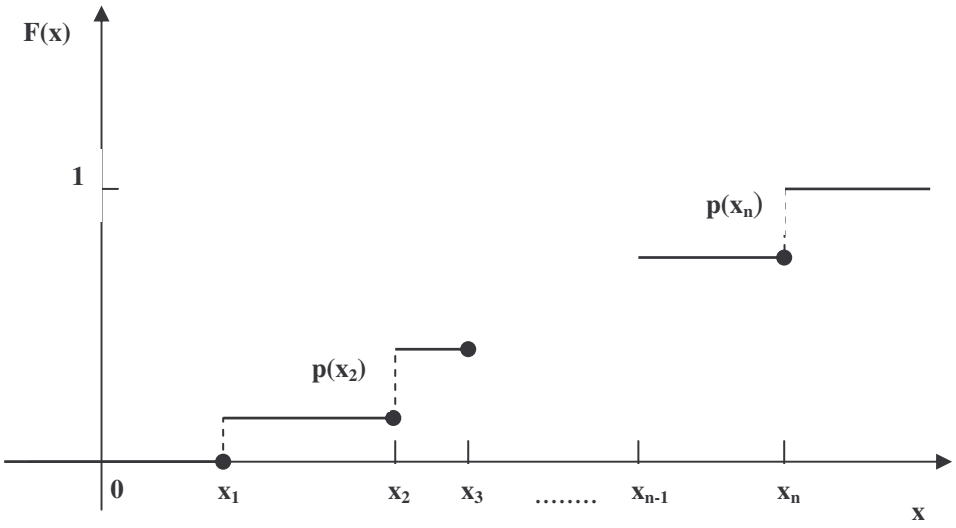
- **Empirical distribution function F(x) for the quantitative variable**

  We have ordered sample population ($x_1 < x_2 < \ldots < x_n$) and we denote $p(x_i)$ as relative frequency of the value x_i. Then it must hold true for empirical distribution function F(x):

$$F(x) = \begin{cases} 0 & for\ x \le x_1 \\ \sum_{i=1}^{j} p(x_i) & for\ x_j < x \le x_{j+1}, 1 \le j \le n-1 \\ 1 & for\ x_n < x \end{cases}$$

  The empirical distribution function is monotonous increasing function and it is continuous from the left.

$$p(x_i) = \lim_{x \to x_{i+}} F(x) - F(x_i)$$

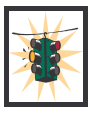


- **MAD**

  MAD is a name for **m**edian **a**bsolute **d**eviation from the median.

  We determine MAD in this way:

  1. we order the sampling population by size
  2. we determine a median of the sampling population
  3. for each value we determine absolute value of its deviation from the median
  4. the absolute deviations from the median we order by size
  5. now we determine a median of the absolute deviations from the median i.e. MAD



Solved example

We have these data: 22, 82, 27, 43, 19, 47, 41, 34, 34, 42, 35 (these are the same data as the previous solved example).
Determine:

a) all quartiles
b) inter-quartile range
c) MAD
d) draw in an empirical distribution function

Solution:

a) We must determine lower quartile $x_{0,25}$; median $x_{0,5}$ and upper quartile $x_{0,75}$. At first we order the data by size and we assign a sequence to them.

Original data	Ordered data	Sequence
22	19	1
82	22	2
27	**27**	**3**
43	34	4
19	34	5
47	**35**	**6**
41	41	7
34	42	8
34	**43**	**9**
42	47	10
35	82	11

Now we can assign a sequence of the variable values for individual quartiles i.e. also their values:

Lower quartil $x_{0,25}$: $p = 0,25; n = 11 \Rightarrow z_p = 11.0,25 + 0,5 = 3,25 \cong 3 \Rightarrow x_{0,25} = 27$,

i.e. 25% musicians is younger then 27 years (75% of them have 27 years and more).

Median $x_{0,5}$: $p = 0,5; n = 11 \Rightarrow z_p = 11.0,5 + 0,5 = 6 \Rightarrow x_{0,5} = 35$

i.e. a half of the musician is younger then 35 years (50% of them have 35 years and more).

Upper quartil $x_{0,75}$: $p = 0,75; n = 11 \Rightarrow z_p = 11.0,75 + 0,5 = 8,75 \cong 9 \Rightarrow x_{0,75} = 43$

i.e. 75% musicians is younger then 43 years (25% of them have 43 years and more).

b) **Inter-quartile range IQR:**

IQR = $x_{0,75} - x_{0,25} = 43 - 27 = 16$

c) **MAD**

If we want determine this characteristic we must act upon the definition (a median of absolute deviations from the median).

$x_{0,5} = 35$

Original data x_i	**Ordered data y_i**	Absolute values of deviations of the ordered data from their median $\lvert y_i - x_{0,5}\rvert$	Ordered absolute values **M_i**
22	19	$16 = \lvert 19 - 35\rvert$	0
82	22	$13 = \lvert 22 - 35\rvert$	1
27	27	$8 = \lvert 27 - 35\rvert$	1
43	34	$1 = \lvert 34 - 35\rvert$	6
19	34	$1 = \lvert 34 - 35\rvert$	7
47	35	$0 = \lvert 35 - 35\rvert$	**8**
41	41	$6 = \lvert 41 - 35\rvert$	8
34	42	$7 = \lvert 42 - 35\rvert$	12
34	43	$8 = \lvert 43 - 35\rvert$	13
42	47	$12 = \lvert 47 - 35\rvert$	16
35	82	$47 = \lvert 82 - 35\rvert$	47

$$MAD = M_{0,5}$$

$$p = 0{,}5;\ n = 11 \Rightarrow z_p = 11.0{,}5 + 0{,}5 = 6 \Rightarrow M_{0,5} = 8$$

(MAD is a median absolute deviation from the median i.e. 6th value of ordered absolute deviations from the median)

MAD = 8.

d) The last thing is draw in an empirical distribution function. Here's its definition:

$$F(x) = \begin{cases} 0 & for\ x \le x_1 \\ \sum_{i=1}^{j} p(x_i) & for\ x_j < x \le x_{j+1}, 1 \le j \le n-1 \\ 1 & for\ x_n < x \end{cases}$$

- we write ordered variable values their frequencies and relative frequencies into the table and of them we derive an empiric distribution function:

Original data x_i	**Ordered data a_i**	Absolute frequencies of the ordered values **n_i**	Relative frequencies of the ordered values **p_i**	Empirical distribution function $F(a_i)$
22	19	1	1/11	0
82	22	1	1/11	1/11
27	27	1	1/11	2/11
43	34	2	2/11	3/11
19	35	1	1/11	5/11
47	41	1	1/11	6/11
41	42	1	1/11	7/11
34	43	1	1/11	8/11
34	47	1	1/11	9/11
42	82	1	1/11	10/11
35				

From definition of the empirical distribution function F(x) results that F(x) is equal 0 for all x <19, F(x) is equal 1/11 for 22≥x>19, F(x) is equal 1/11 + 1/11 for 27≥x> 22, etc.

x	$\langle -\infty;19\rangle$	$\langle 19;22\rangle$	$\langle 22;27\rangle$	$\langle 27;34\rangle$	$\langle 34;35\rangle$
F(x)	0	1/11	2/11	3/11	5/11

x	$\langle 35;41\rangle$	$\langle 41;42\rangle$	$\langle 42;43\rangle$	$\langle 43;47\rangle$	$\langle 47;82\rangle$	$\langle 82;\infty)$
F(x)	6/11	7/11	8/11	9/11	10/11	11/11

Means, mode and median (i.e. measures of position) represent imaginary centre of the variable. But a distribution of the individual values of the variable round of this centre (i.e. measures of variability) interested us too.

The following three statistical characteristics allow a description of sampling population variability. Short and inter-quartile range we include among measures of variability.

- **Sample variance s^2**

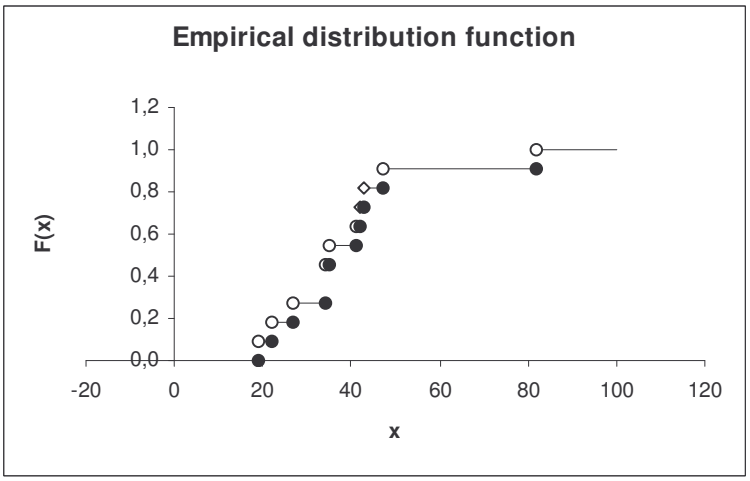
  - it is the most frequently measure of variability

  The sample variance is given by:

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

  - the sample variance is the sum of the squared deviations from their mean divided by one less than the sample size

  General **properties of the sample** variance are for example:



  - The sample variance of a constant number is equal

otherwise: if all variable values are the same the sampling has zero diffusenesses

- $$\forall a \in \Re : \left[\left(s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \right) \wedge (y_i = a + x_i) \right] \Rightarrow \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = s^2$$

otherwise: if we add a same constant number to all variable values the sample variance won't be change

- $$\forall b \in \Re : \left[\left(s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \right) \wedge (y_i = bx_i) \right] \Rightarrow \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = b^2 s^2$$

otherwise: if we multiple all variable values an arbitrary constant number (b) the sample variance increase about square of this constant number (b^2)

Disadvantage for use the sample variance as a measure of variability is that a size of this characteristic is square of the variable size. For example: if the variable is cash in EUR than the sample variation of this variable will be in EUR^2. That is why we use other measure of variability namely a standard deviation.

- **Standard deviation s**

  - it is calculated by taking the square root of the variance

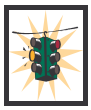
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Other disadvantage for use the sample variation and the standard deviation is that we can't compare variability of the variable that are express in various units. Which variable has bigger variability - height or weight of an adult? Coefficient of variance will give us answer for this question.

- **Coefficient of variation V_x**

  - it represents relative measure of variability of the variable x and it is often expressed as a percentage
  - it is the ratio of the sample standard deviation to the sample mean:

$$V_x = \frac{s}{\bar{x}}$$



Solved example

Firm producing the table glass developed less expensive technology for improving glass resistant against fire. For testing there was selected and cut in half 5 table glasses. One half was treated by a new technology while the other one was left for control. Both halves were tested for increasing effect of fire till they crack. These results were obtained:

Critical temperature (glass cracked) [^oC]	
Old technology x_i	**New technology y_i**
475	485
436	390
495	520
483	460
426	488

Compare both technologies by means of basic characteristics of the exploratory analysis (mean, variation,...).

Solution:

- at first we try compare both technologies with the help of the mean:

Mean for the old technology:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{475 + 436 + \ldots + 426}{5} = 463{,}0 \quad [^oC]$$

Mean for the new technology:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{485 + 390 + \ldots + 488}{5} = 468{,}6 \quad [^oC]$$

Based on calculated means we could say that we recommend new technology because critical temperature is almost 6^{o}C higher using it.

- now we determine measures of variability

The old technology:

Sample variance:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{(475 - 463{,}0)^2 + (436 - 463{,}0)^2 + \ldots + (426 - 463{,}0)^2}{5-1} = 916{,}3 \quad [^oC^2]$$

Standard deviation:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{s_x^2} = \sqrt{916{,}3} = 30{,}3 \quad [^oC]$$

New technology:

Sample variance:

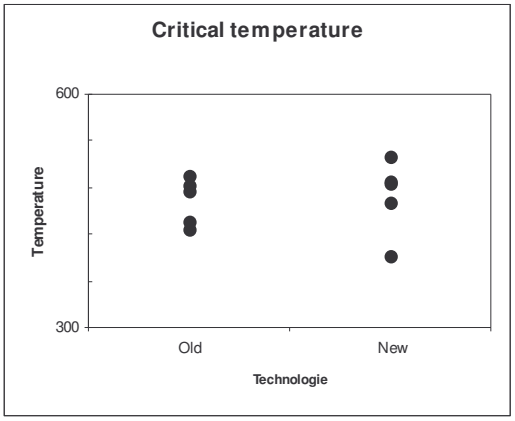
$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{(485-468{,}6)^2 + (390-468{,}6)^2 + \ldots + (488-468{,}6)^2}{5-1} = 2384{,}4 \quad [^oC^2]$$

Standard deviation:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{s_y^2} = \sqrt{2384{,}4} = 48{,}8 \quad [^oC]$$

Sample variance (standard deviation) is much bigger for new technology. What does it mean? Look at the graphical representation of measured data. Critical temperatures are much more spread what mean this technology is not well managed yet and its use can't guarantee higher quality of production. In this case it can come to significant improving as well as significant reducing of critical temperature. That's why the new technology should be subjected to continuous research.

These conclusions are based only on exploratory analysis. Statistics provides us more exact methods for analysis of such problems (hypothesis testing).



And now we go back to exploratory analysis as such. We made a mention of outliers. For now we know that as the outliers we specify these variable values which are extraordinary different from others and that influence for example representatives of mean. How to identify these values?

- **Identification of the outliers**

  In the statistical practice we can meet with a few methods of the outliers' identification. We'll show three of them.

  1. The outlier can be such value x_i that is far more then 1,5 IQR from lower (or upper) quantile.

$$\left[\left(x_i < x_{0,25} - 1{,}5IQR\right) \vee \left(x_i < x_{0,75} + 1{,}5IQR\right)\right] \Rightarrow x_i \text{ is an outlier}$$

2. The outlier can be such value x_i where absolute value of z-axis is greater then 3.

$$z - axis._i = \frac{x_i - \bar{x}}{s}$$

$$\left(\left|z - axis._i\right| > 3\right) \Rightarrow x_i \text{ is an outlier}$$

3. The outlier can be such value x_i where absolute value of median-axis is greater then 3.

$$median - axis._i = \frac{x_i - x_{0,5}}{1{,}483.MAD}$$

$$\left(\left|median - axis._i\right| > 3\right) \Rightarrow x_i \text{ is an outlier}$$

For outliers identification in a concrete problem we can choose any of these three rules. Z-axis is "less strict" than median-axis to outliers. It's caused by z-axis is determine on the basis of mean and standard deviation and they are strongly influence of outliers values. While median-axis is determine on the basis of median and MAD and they are immune to outliers.

When we decide that any value is an outlier we must distinguish a type of that outlier. In case that outlier is caused by:

- blunders, typing errors, evincible failure of the people or the technology ...
- effects of faults or wrong measurement, ...

It comes to this if we know the outlier cause and if assume that will not occur again we can cast out this outlier from other process. In the others cases we must consider if we can cast out the outliers and at the same time won't get about important information any events which are with low frequencies.

The others characteristics which describe qualitative variable are **skewness** and **kurtosis**. Formulas for calculation of these characteristics are rather complicated that is why we determine these characteristics by means of some statistical program.

- **Skewness**

  - Skewness is defined as asymmetry in the distribution of the variable values. Values on one side of the distribution tend to by further from the "middle" than values on the other side.

  - Its value we obtain by means of this formula:

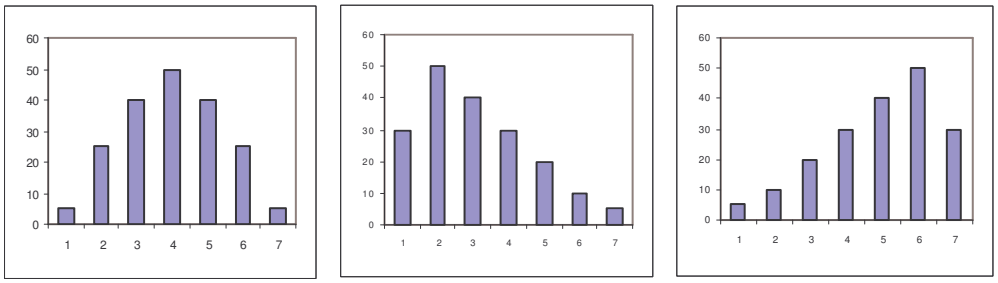
$$\alpha = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

Skewness interpretation:

$\alpha = 0$ ... variable values are distributed symmetrically round the mean

$\alpha > 0$ ... there predominate values less then mean by the variable

$\alpha < 0$ ... there predominate values greater then mean by the variable



α=0 α>0 α<0

- **Kurtosis**

- Kurtosis represents concentration of variable values round their mean.

- Its value we obtain by means of this formula:

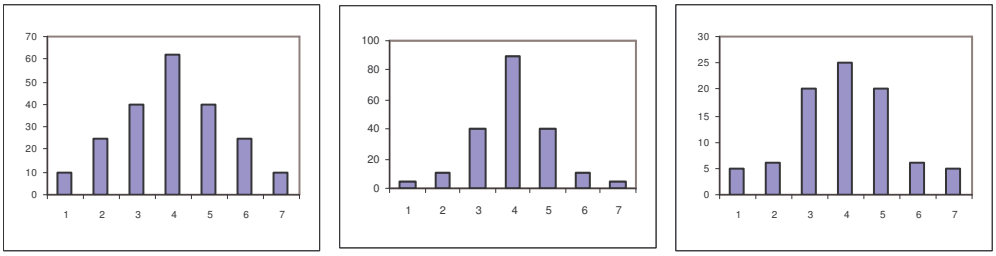
$$\beta = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

Kurtosis interpretation:

$\beta = 0$ ... Kurtosis corresponds to normal distribution

$\beta > 0$ ... "peaked" distribution of the variable

$\beta < 0$ ... "flat" distribution of the variable



ß=0 ß>0 ß<0

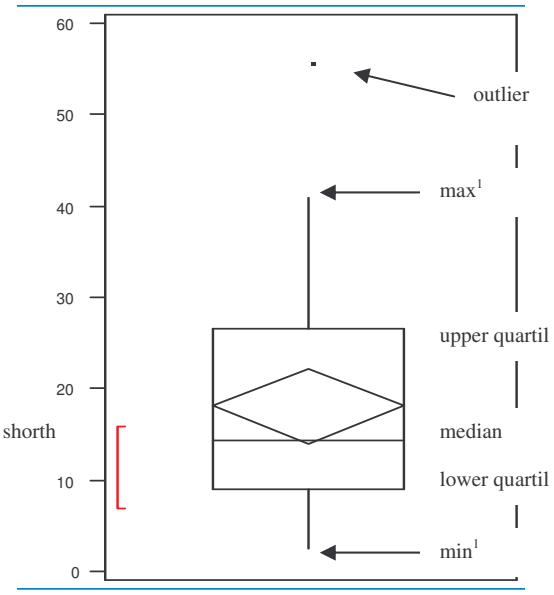
Now we have defined all numerical characteristics for description of the quantitative variable. We have show left how we can graphically represent quantitative variable.

1.2.2 Graphical presentation quantitative variable

Box plot

A box plot is a way of summarizing a set of data measured on an interval scale. It is often used in exploratory data analysis. It is a type of graph which used to show the shape of the distribution, its central value, and variability. The picture produced consists of the most extreme values in the data set (maximum and minimum), the lower and upper quartiles, and the median.

A box plot is especially helpful for indicating whether a distribution is skewed and whether there are any unusual observations (outliers) in the data set.



Notice.: *A box plot construction begins drawing in outliers and until then we mark the others characteristics (min[1], max[1], quartiles and shorth).*

Stem and leaf plot

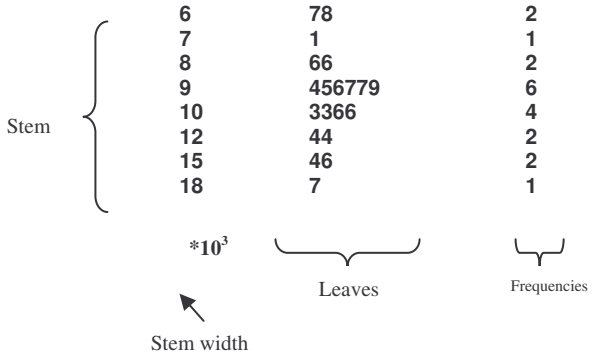
We saw it that simplicity is an advantage of box plot. But information's about concrete values of variable are missing us sometimes. We would digestedly inscribe the numeric values. To it we use stem and leaf plot.

We have a variable which represent average month pay of bank employees in Czech Republic.

Average month pay [CZK]									
10 654	9 765	8 675	12 435	9 675	10 343	18 786	15 420	8 675	7 132
6 732	6 878	15 657	9 754	9 543	9 435	10 647	12 453	9 987	10 342

Average month pay [CZK] - ordered data									
6 732	6 878	7 132	8 675	8 675	9 435	9 543	9 675	9 754	9 765
9 987	10 342	10 343	10 647	10 654	12 435	12 453	15 420	15 657	18 786

How we have to inscribe these data. The information about "unimportant" places we neglect and we inscribe ordered data only pursuant to higher places. For our information are interesting values from third place. The values that are on a fourth place we write down sorted. Herewith they create a **stem**. Under the graph we adduce a **stem width**. This width denotes coefficient whereby we multiply values in the graph.

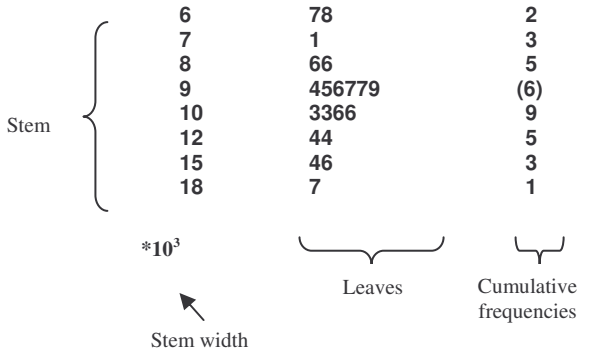


The second column in the graph -**leaves** - is numbers which represent "important" place. These numbers we write in appropriate rows.

The third column is **absolute frequency** for particular rows.

For example: the first row in the graph represents two values - (6.7 and 6.8)*10^3 CZK i.e. 6700 CZK and 6800 CZK, the sixth row represents two values too - (12.4 and 12.4)*10^3 CZK, i.e. two employees have average month pay 12400 CZK, etc.

It exist various modifications of this graph. For example in the third column could be cumulative frequencies whereas in the row whereof is a median we show absolute frequency (in parentheses) and towards this row the frequencies cumulate both from the least values and from the highest values - see picture.



Finally you can take exception that you can make different types of construction of the stem and leaf plot for one problem. Nowhere is it said which place of variable is important and which one is not important. This conclusion depends to you. We can say one tip - the long stem with the short leaves and the short stem with long leaves indicate of incorrect choice of scale. Look at picture.

0	66788999999	11
1	000022558	9

*10^4

Questions

1. What is exploratory statistics concerned with?

2. Characterize the base types of variables.

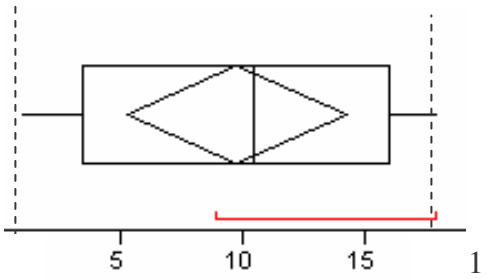
3. Which statistical characteristics can be contained in frequency table (for what type of variable)?

4. What are the outliers and how we define them?

5. Which characteristics is sensitive on the outliers occurrence:
   a) Median
   b) Arithmetical mean
   c) Upper quartil

6. How we graphically represent the qualitative (quantitative) variables?

7. This box plot describes profits of the students during holiday.



Denote assertions which do not correspond with displayed reality.

a) A student earned max 19 thousands CZK
b) Inter-quartile range is cca 10 thousands CZK
c) Half of students earned less than 11 thousands CZK
d) Shorth is cca (5;15) thousands CZK

Problems

Example 1: The following data represent country of the car production. Analyze these data (frequency, relative frequency, cumulative frequency and cumulative relative frequency, mode) and represent them in graphical form (histogram, pie graph).

USA	USA	Germany	Czech Rep.
Germany	Germany	Germany	Czech Rep.
Czech Rep.	Czech Rep.	USA	Germany

Example 2: The following data represent waiting time (min) of the customer to the service. Draw box plot and stem and leaf plot.

120	80	100	90
150	5	140	130
100	70	110	100

Example 3: During a traffic survey there was an utilization of crossroad enterance observed. Student making research always wrote down a number of cars waiting in queue when green light jumped on. These are his outcomes:

3 1 5 3 2 3 5 7 1 2 8 8 1 6 1 8 5 5 8 5 4 7 2 5 6 3 4 2 8 4 4 5 5 4 3 3 4 9 6 2 1
5 2 3 5 3 5 7 2 5 8 2 4 2 4 3 5 6 4 6 9 3 2 1 2 6 3 5 3 5 3 7 6 3 7 5 6

Draw box plot, empirical distribution function and calculate mean, standard deviation, shorth, mode and inter-quartile range.

2. PROBABILITY THEORY

Study time: 70 minutes

Aim - you will be able to

- characterize probability theory
- explain general notions of probability theory
- explain and use general relations between events
- explain a notion of probability
- define probability by basic axioms
- define properties of probability function
- use a conditional probability
- explain theorem of total probability and Bayes theorem

2.1. General notions

Study time: 20 minutes

Aim

- characterize probability theory
- explain general notions of probability theory

Explication

Probability Theory is the deductive part of statistics. Its purpose is to give a precise mathematical definition or structure to what has been thus far an intuitive notion of randomness. Making randomness more precise will allow us to make exact probability statements. For example when discussing association, we could only make rough statements in terms of tendencies.
Mathematically, probability is a set function. That is, it is a function defined on some domain of sets. Therefore, we begin this discussion my considering the fundamental nature of sets and the basic operations performed on sets, the elements of the domain of our probability function.

- **General notions of the probability theory**

Definition of Set - set A is a collection of elements. Elements are basic intuitive mathematically undefined entities. To define a set, it is necessary to be able to determine whether any element is included or not included in the set. The notion of inclusion is also an intuitive undefined concept.

Definition of Elementary Events - In the case of probability theory, the elements of sets on which probability measures are defined are called elementary events. In practice, these elementary events may be measurement units, cases, or sample points.

Example:

{reverse, obverse} –when tossing the coin

{1,2,3,4,5,6} – when tossing the dice

We denote a set of all results Ω. This set we call **sample space** (of the elementary events)**.** The elementary event {ω} is a subset of the Ω set which contains one element ω from Ω set, $\omega \in \Omega$.
Then the event A will be an arbitrary subset of Ω, $A \subset \Omega$.

From statistics data we can easily determine that share of boys born in particular years with respect to all born children is moving around 51,5%. Despite the fact that in individual cases we can't foretell a sex of a child we can relatively exactly guess how many boys we find among 10 000 born children.

From this example imply that relative frequencies of some events are stabilized with increase repetition number on certain values. We shall call this phenomenon *a stability of the relative frequencies*. This stability of relative frequencies is an empiric basis of the probability theory. **Relative frequency** is number n(A)/n where n is a total number of experiments and n(A) is a number of experiment realizations in which event A became.

Summary of notions

Probability theory is mathematical branch whose logical structure is created axiomatically.
Mathematic statistics is a science which is concerned with questions of data mining data analyzing and results forming.
Random experiment is every finite process whose result is not given in advance by conditions upon whose is runed.
Sample space Ω is a set of all possibly results of the experiment.
Relative frequencies of some events with increase repetition number show certain **stability**.

2.2. Operations with the elementary events

Study time: 20 minutes

Aim

- types of the elementary events
- general relations between events

Explication

What are types of the elementary events?

If the elementary event ω ∈ Ω (ω ∈ A) came then we can say that an event A came with the experiment realization. We denote this result ω ∈ A as **result favourable to the event A**.

Certain event
- is the event which become with every realization of the experiment. It is equivalent with the Ω set.
Certain event is for example: *we toss one of these numbers 1,2,3,4,5,6* (while tossing a dice)

Impossible event
- is the event which can never become in the experiment. We will denote it as $\varnothing$.
Impossible event is for example: *we toss number 8* (while tossing a dice).

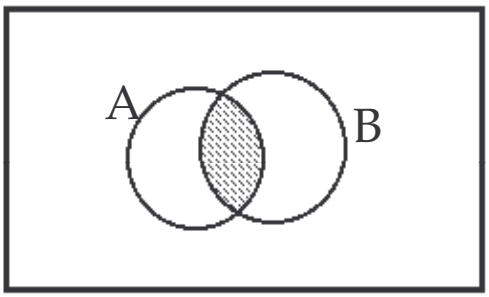
What are relations between events?

Operations on Sets - The operations of union, intersection, complementation (negation), subtraction, the concept of subset, and the null set and universal set or sample space make up the algebra of sets.

Intersection A ∩ B
The set of all elements that are both in A and in B.

Graphic example:

$$A \cap B = \{\omega \mid \omega \in A \wedge \omega \in B\}$$

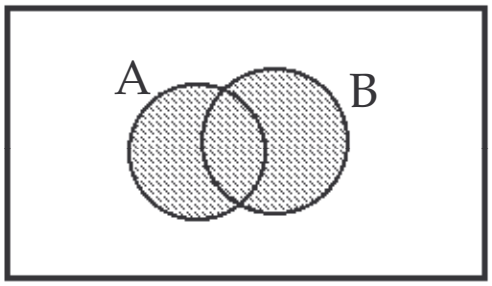


Example – tossing a dice: event A – we toss a number 2 ,3 or 4 and event B – we toss a even number. It is obvious that A ∩ B ={2,4}.

Union A ∪ B
The set of all elements that are either in A or in B.

Graphic example:

$$A \cup B = \{\omega \mid \omega \in A \vee \omega \in B\}$$



Example – tossing a dice: Event A = {1,3,4} and event B is when we toss even number. It's obvious that A ∪ B ={1,2,3,4,6}.

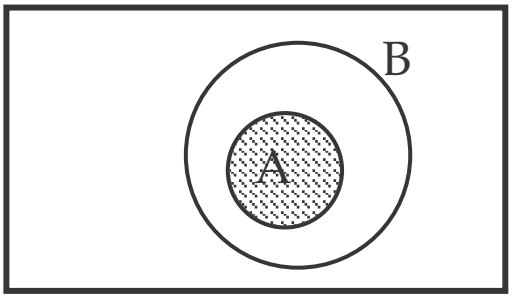
Disjoint events A ∩ B = ∅
Two events A and B can't become together. They have none common result.

Example – tossing a dice: Event A – we toss even number and even B – we toss odd number. These events never have a same result. If event A become than event B can't become.

Subsets (Subevent) A ⊂ B
A is a subset of B if every element of A is also an element of B. It's mean if event A become than event B become too.

Graphic example:

$$A \subset B \Leftrightarrow \{\omega \in \mathrm{A} \Rightarrow \omega \in B\}$$



Example – tossing a dice: Event A – we toss number 2 and event B – we toss even number. The event A is subevent of the event B.

Events A and B are equivalent A = B if $A \subset B$ and at the same time $B \subset A$.

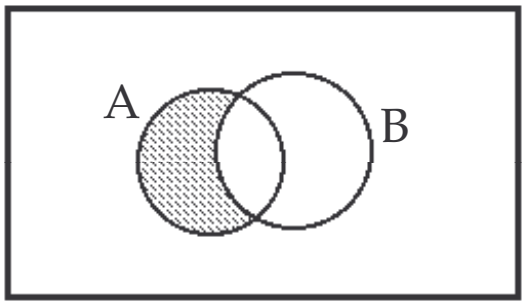
Example – tossing a dice: Event A – we toss even number, event B – we toss number what is divide of number 2. These events are equivalent.

Subtraction A-B
The set of all elements that are in A but not in B

$$A - B = A \cap \overline{B}$$

$$A - B = \{\omega | \omega \in A \wedge \omega \notin B\}$$

Graphic example:

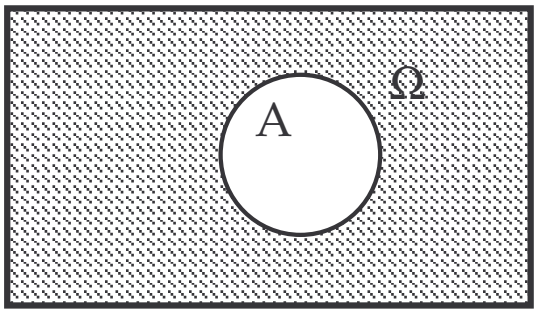


Example – tossing a dice: Event A – we toss a number greater than two and event B – we toss an even number. Subtraction of the events A and B is an event A – B ={3,5}.

Complement of the event A (opposite event)
The set of all elements that are not in A.

$$\overline{A} = \{\omega | \ \omega \notin A\}$$

Graphic example:



Example – tossing a dice: Event A – we toss an even number, then an event $\overline{A}$ - we toss an odd number.

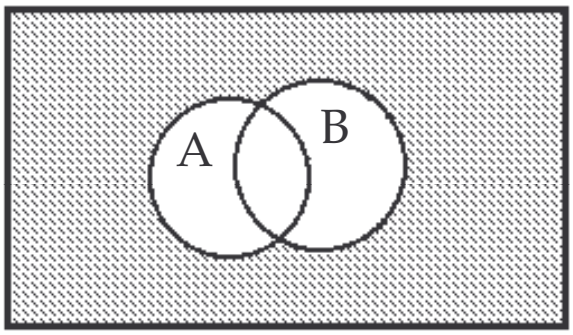
DeMorgan's Laws

- DeMorgan's Laws are logical consequences of the fundamental concepts and basic operations of sets

1. law

The set of all elements that are neither in A nor in B.

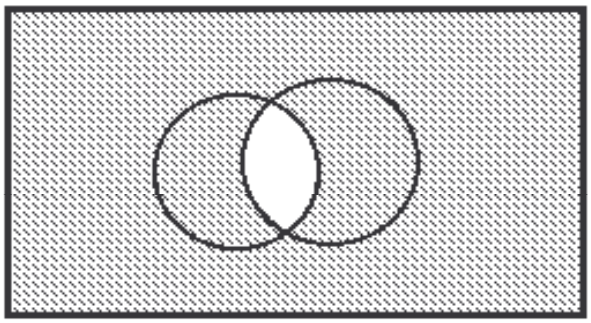
$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$



2. law

The set of all elements that are either not in A or not in B.

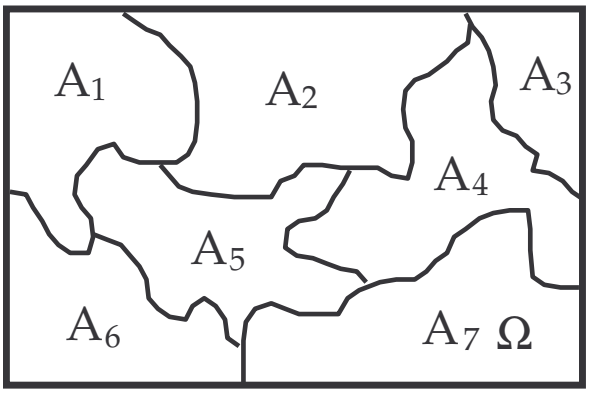
$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$



Mutually disjoint sets and partitioning the sample space

The collection of sets $\{A_1, A_2, A_3, \ldots\}$ partition the sample space Ω: $A_i \bigcap A_j = \varnothing$ for $i \neq j$

$$\Omega = \bigcup_{i=1}^{n} A_i$$



2.3. Probability theory

Study time: 30 minutes

Aim

- notion of probability
- basic theorems and axioms of probability
- types of probability
- conditional probability
- theorem of total probability and Bayes theorem

Explication

Notion of probability

Probability of the event A is a number P(A) which has a property that a relative frequency of the event A with increase realizations number is approaching to the number P(A).

This probability definition is known as *classic probability definition.*

Now we introduce axiomatic probability definition.

Axiomatic probability definition

Probability space is a triad (Ω, S, P) where

(i) Ω is sample space (elements of Ω are elementary events)

(ii) S is a set of subsets of Ω that it holds:

a) $\Omega \in S$;

b) if $A \in S$ then $\overline{A} = \Omega - A \in S$;

c) if $A_1, A_2, A_3, \ldots \in S$ then $\bigcup_{i=1}^{\infty} A_i \in S$

Elements of S we denote as **events**.

(iii) P is function from S to $< 0,1 >$ such that it holds:

a) $P(\Omega) = 1$- probabilities are scaled to lie in the interval [0,1];

b) $P(\overline{A}) = 1 - P(A)$ for every $A \in S$;

c) For a collection of mutually disjoint sets, the probability of their union is equal to the sum of their probabilities.

$$\text{If } A \cap B = \varnothing, \text{ then}$$
$$P\{A \cup B\} = P\{A\} + P\{B\}$$

In general,

$$A_i \cap A_j = \varnothing, \ \forall 1 \le i, j \le \infty; i \ne j,$$
$$P\{\bigcup_{i=1}^{\infty} A_i\} = \sum_{i=1}^{\infty} P\{A_i\}$$

Function P is called **probability measure** or shorter **probability**.

Example – tossing a dice:
$\Omega = \{1,2,3,4,5,6,\}$,
S is a set of subsets of Ω (sometimes we denote $S = \exp \Omega$) and probability is defined by $P(A) = \frac{\text{card}A}{6}$ where card A is number of set A elements.

Basic Theorems of Probability

The following theorems are the logical consequences of the three basic probability axioms we have postulated.

1. For disjoint events A and B hold:
   $A \cap B = \varnothing$ then
   $P\{A \cup B\} = P\{A\} + P\{B\}$

2. If for two events A,B hold:
   $B \subset A$ then $P\{B\} \le P\{A\}$
   - note that A is partitioned by B and its complement, and hence $P\{A\}$ is sum of these two parts

3. For every event A holds: $P\{\overline{A}\} = 1 - P\{A\}$
   - the union of the two sets is the sample space, the intersection is the null sets

4. It holds: $P\{\varnothing\} = 0$

5. It hold: $P\{B - A\} = P\{B\} - P\{B \cap A\}$
   - note that B-A and B intersection A are two disjoint sets whose union is B

6. Especially if $A \subset B$ then $P\{B - A\} = P\{B\} - P\{A\}$

7. For arbitrary events A,B hold:
   $P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$

8. Follows from de Morgan's laws

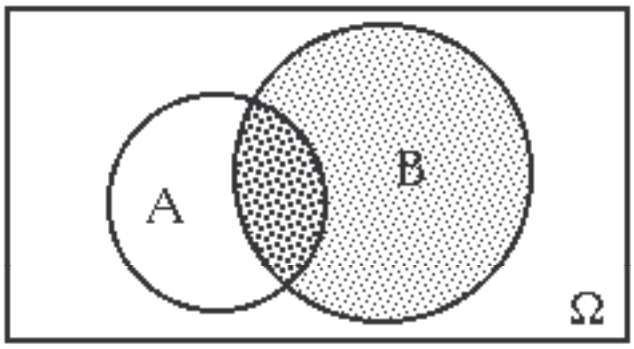
$$P\{A \cup B\} = 1 - P\{\overline{A \cup B}\}$$
$$= 1 - P\{\overline{A} \cap \overline{B}\}$$

Definition of Conditional Probability

The definition of conditional probability determines how probabilities adjust to changing conditions. When we say that the condition B applies, we mean that the set B is known to have occurred and therefore the rest of the sample space in the complement of B has zero probability. Under these new circumstances, the revised probability of any other event, A, can be determined from the following definition of conditional probability:

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

By this formula, the probability of that part of the event A which is in B or intersects with B is revised upwards to reflect the condition that B has occurred and becomes the new probability of A. It is assumed that the probability of B is not zero.



$P\{A|B\}$ - *probability of the event A conditional by the event B*

Conditional Probability Definition of Independence

If the condition that B has occurred does not affect the probability of A, then we say that A is independent of B.

$$P\{A|B\} = P\{A\}$$

From the definition of conditional probability, this implies

$$P\{A\} = \frac{P\{A \cap B\}}{P\{B\}}$$

and hence,

$$P\{A \cap B\} = P\{A\} \cdot P\{B\}$$

It is clear from this demonstration that if A is independent of B, then B is also independent of A.

Example – tossing a dice:
For events A - "we toss 1 in the first toss" and B - "we toss 1 in the second toss" and event C = A ∩ B - "we toss 1 in the both tosses" then it holds:

$$P\{C\} = P\{A \cap B\} = P\{A\}.P\{B\} = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$$

Theorem of Total Probability

If a collection of sets $\{B_1, B_2, B_3, \ldots, B_n\}$ partition the sample space Ω, that is,

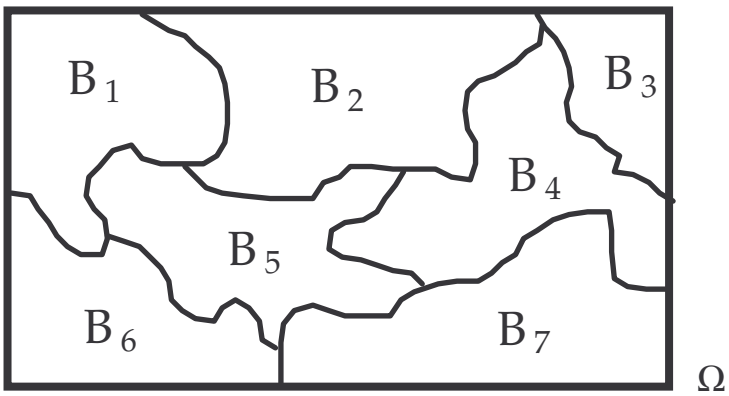
$$B_i \cap B_j = \varnothing; \forall\ i \neq j$$

$$\bigcup_{i=1}^{n} B_i = \Omega$$

then for any set A ($P\{A\} \neq 0$) in the sample space Ω,

$$P\{A\} = \sum_{i=1}^{n} P\{A|B_i\}\cdot P\{B_i\}$$

n=7



Proof: Since the collection of sets $\{B_1, B_2, B_3, \ldots, B_n\}$ partitions the sample space Ω,

$$P\{A\} = \sum_{i=1}^{n} P\{A \cap B_i\}$$

From the definition of conditional probability

$$P\{A \cap B_i\} = P\{A|B_i\}\,P\{B_i\}$$

Bayes Theorem

If the collection of sets $\{B_1, B_2, B_3, \ldots, B_n\}$ partitions the sample space Ω, then

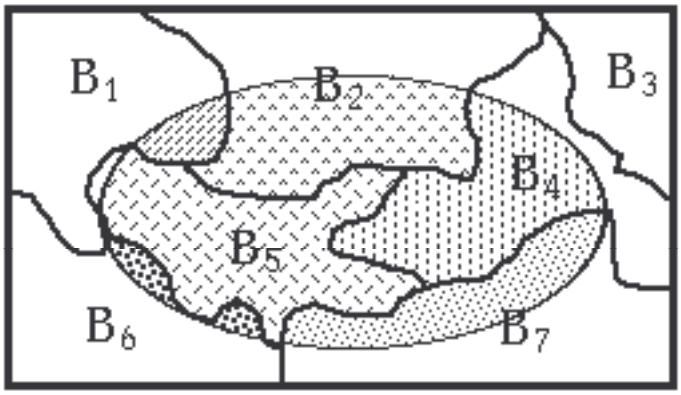
$$P\{B_k|A\} = \frac{P\{A|B_k\}P\{B_k\}}{\sum_{i=1}^{n} P\{A|B_i\}P\{B_i\}}$$

Proof: From the definition of conditional probability,

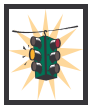
$$P\{B_k|A\} = \frac{P\{B_k \cap A\}}{P\{A\}} = \frac{P\{A\}P\{B_k\}}{P\{A\}}$$

Substituting for $P\{A\}$ from the Theorem of Total Probability, the proof follows.

Graphical representation of Bayes theorem (vyšrafovaná plocha znázorňuje jev A):



PROBABILITY THEORY – SOLVED EXAMPLES



Solved example

Probability of failing of the extinguishing system is 20%. Probability that alarming system fails is 10% and probability that both systems fail is 4%. What is a probability that:
a) at least one system will be working?
b) both systems will be working?

Solution:

We denote: H ... extinguishing system works
S ... alarming system woks

We know that: $P(\overline{H}) = 0{,}20$
$P(\overline{S}) = 0{,}10$
$P(\overline{H} \cap \overline{S}) = 0{,}04$

We must find:

ada) $P(H \cup S)$

We have two possibilities for solving:

By the definition: Events H and S are not the disjoint events and hence:

$$P(H \cup S) = P(H) + P(S) - P(H \cap S),$$

but would be a problem determine a $P(H \cap S)$

By the opposite event: From de Morgan's laws we can write:

$$P(H \cup S) = 1 - P(\overline{H \cup S}) = 1 - P(\overline{H} \cap \overline{S}),$$

$$\underline{\underline{P(H \cup S)}} = 1 - 0{,}04 = \underline{\underline{0{,}96}}$$

The probability (that at least one system will be working) is 96%.

adb) $P(H \cap S)$

We can't solve it by the definition**:**

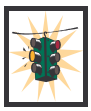
$$(P(H \cap S) = P(H|S) \cdot P(S) = P(S|H) \cdot P(H)),$$

because we have no information about dependency of the failures of the individual systems. Hence we try to use the opposite event:

$$P(H \cap S) = 1 - P(\overline{H \cap S}) = 1 - P(\overline{H} \cup \overline{S}) = 1 - [P(\overline{H}) + P(\overline{S}) - P(\overline{H} \cap \overline{S})],$$

$$\underline{\underline{P(H \cap S) = 1 - [P(\overline{H}) + P(\overline{S}) - P(\overline{H} \cap \overline{S})] = 1 - [0{,}20 + 0{,}10 - 0{,}04] = 0{,}74}}$$

The probability (that both systems will be working) is 74%.



Solved example

120 students passed mathematics and physics exams. 30 of them failed to pass both exams. 8 failed to pass only math exam and 5 failed to pass only physics exam. What is probability that random student:

a) passed math exam if we know that he failed to pass physics exam
b) passed physics exam if we know that he failed to pass math exam
c) passed math exam if we know that he passed physics exam

Solution:

We denote: M ... he passed mathematics exam
F ... he passed physics exam

We know that:

$$P(\overline{M} \cap \overline{F}) = \frac{30}{120}$$
$$P(\overline{M} \cap F) = \frac{8}{120}$$
$$P(M \cap \overline{F}) = \frac{5}{120}$$

We must find:

ada) $P(M|\overline{F})$

by the definition of conditional probability:

$$P(M|\overline{F}) = \frac{P(M \cap \overline{F})}{P(\overline{F})} = \frac{P(M \cap \overline{F})}{P(M \cap \overline{F}) + P(\overline{M} \cap \overline{F})},$$

$$\underline{\underline{P(M|\overline{F}) = \frac{P(M \cap \overline{F})}{P(M \cap \overline{F}) + P(\overline{M} \cap \overline{F})} = \frac{\frac{5}{120}}{\frac{5}{120} + \frac{30}{120}} = \frac{5}{35} = \frac{1}{7} \cong 0{,}14}}$$

The probability (that he passed math exam if we know that he failed to pass physics exam) is 14%.

adb) $P(F|\overline{M})$

the same way as ada):

$$P(F|\overline{M}) = \frac{P(F \cap \overline{M})}{P(\overline{M})} = \frac{P(F \cap \overline{M})}{P(F \cap \overline{M}) + P(\overline{F} \cap \overline{M})},$$

$$\underline{\underline{P(F|\overline{M}) = \frac{P(F \cap \overline{M})}{P(F \cap \overline{M}) + P(\overline{F} \cap \overline{M})} = \frac{\frac{8}{120}}{\frac{8}{120} + \frac{30}{120}} = \frac{8}{38} = \frac{4}{19} \cong 0{,}21}}$$

The probability (that he passed physics exam if we know that he failed to pass math exam) is 21%.

adc) $P(M|F)$

from the definition:

$$P(M|F) = \frac{P(M \cap F)}{P(F)},$$

we have two possibilities:

1)

$$\underline{\underline{P(M|F) = \frac{P(M \cap F)}{P(F)} = \frac{1 - P(\overline{M \cap F})}{1 - P(\overline{F})} = \frac{1 - P(\overline{M} \cup \overline{F})}{1 - [P(\overline{F} \cap M) + P(\overline{F} \cap \overline{M})]} = \frac{1 - [P(\overline{F}) + P(\overline{M}) - P(\overline{F} \cap \overline{M})]}{1 - [P(\overline{F} \cap M) + P(\overline{F} \cap \overline{M})]} =}}$$

$$= \frac{1 - [[P(\overline{F} \cap M) + P(\overline{F} \cap \overline{M})] + [P(F \cap \overline{M}) + P(\overline{F} \cap \overline{M})] - P(\overline{F} \cap \overline{M})]}{1 - [P(\overline{F} \cap M) + P(\overline{F} \cap \overline{M})]} =$$

$$= \frac{1 - [P(\overline{F} \cap M) + P(F \cap \overline{M}) + P(\overline{F} \cap \overline{M})]}{1 - [P(\overline{F} \cap M) + P(\overline{F} \cap \overline{M})]} = \frac{1 - \left[\frac{5}{120} + \frac{8}{120} + \frac{30}{120}\right]}{1 - \left[\frac{5}{120} + \frac{30}{120}\right]} = \frac{\frac{77}{120}}{\frac{85}{120}} = \frac{77}{85} \cong 0{,}91$$

2)
We write given data into the table:

	They passed math exam	**They failed to pass math exam**	**Total**
They passed physics exam		8	
They failed to pass physics exam	5	30	35
Total		38	**120**

and we calculate remaining data:

How much students passed physics exam? It is total number(120) minus number of students who failed to pass physics exam (35) and that is 85. Analogously for number of students who

passed math exam: 120 – 38 = 82. And for number of students who passed both exams: 82 – 5 = 77.

	They passed math exam	**They failed to pass math exam**	**Total**
They passed physics exam	77	8	85
They failed to pass physics exam	5	30	35
Total	82	38	**120**

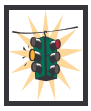
Finding probabilities are:

$$P(M \cap F) = \frac{77}{120}; \quad P(F) = \frac{85}{120},$$

from that imply:

$$P(M|F) = \frac{P(M \cap F)}{P(F)} = \frac{\frac{77}{120}}{\frac{85}{120}} = \frac{77}{85} \cong 0{,}91$$

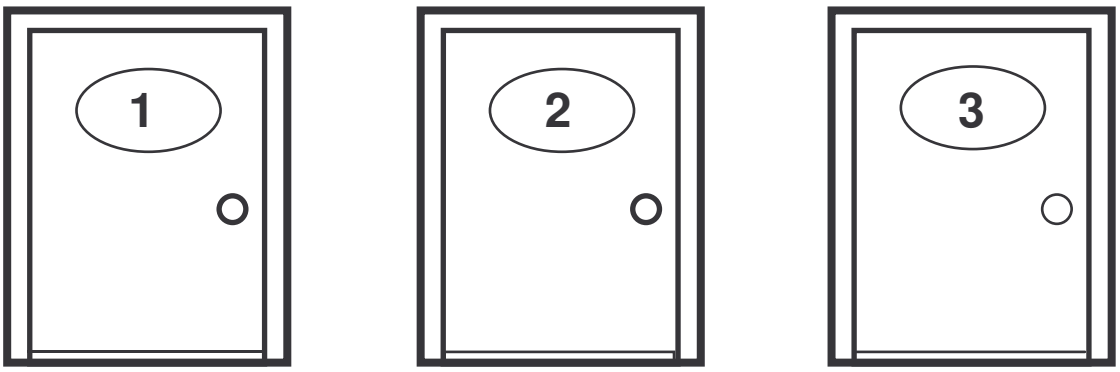
The probability (that he passed math exam if we know that he passed physics exam) is 91%.



Solved example (Example of an Application of Bayes's Theorem)

In a famous television game show, the winner of the preliminary round is given the opportunity to enhance his winnings. The contestant is presented with three closed doors and told that behind one of the doors is a new automobile while behind the other two doors are goats. If the contestant correctly selects the door which conceals the automobile, he will win the automobile.

The game show host asks the contestant to make a preliminary selection, after which the host opens one of the other two doors to reveal a goat. The contestant is then given the option of switching his choice to the other door which remains closed. Should he change his choice?



Solution:

The sample space consists of three possible arrangements {AGG, GAG, GGA}.

Assume that each of the three arrangements have the following probabilities:

$p_1 = P\{AGG\}$ $p_2 = P\{GAG\}$ $p_3 = P\{GGA\}$

where $p_1 + p_2 + p_3 = 1$.

Assume without loss of generality that the contestant's preliminary choice is Door #1 and the host opens Door #3 to reveal a goat. One the basis of this information we must revise our probability assessments. It is clear that the host cannot open Door #3 if it conceals the automobile.

$$P\{\text{Door \#3} \mid GGA\} = 0$$

Also, the host must open Door #3 if Door #2 conceals the automobile since he cannot open Door #1, the contestant's choice.

$$P\{\text{Door \#3} \mid GAG\} = 1$$

Finally if the automobile is behind the contestant's first choice, Door #1, the host can choose to open either Door #2 or Door #3. Suppose he chooses to open Door #3 with some probability q.

$$P\{\text{Door \#3} \mid AGG\} = q$$

Then by Bayes' Theorem, we can compute the revised probability that the automobile is behind Door #2 as

$$P\{GAG \mid \text{Door \#3}\} = \frac{P\{\text{Door \#3} \mid GAG\} \cdot P\{GAG\}}{P\{\text{Door \#3}\}}$$

Substituting known values into this equation we obtain,

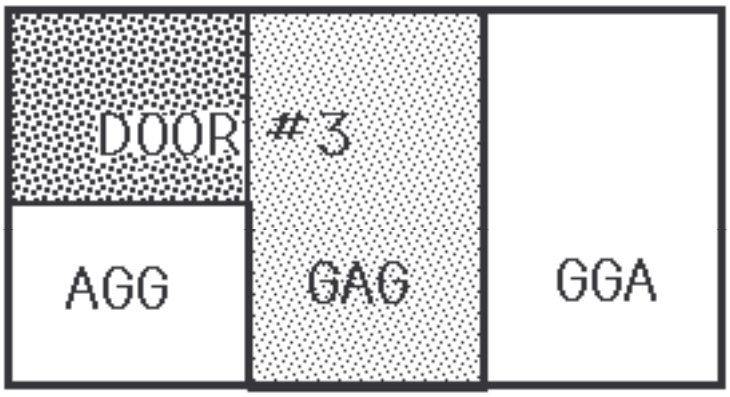
$$P\{GAG \mid \text{Door \#3}\} = \frac{1 \times p_2}{(q \times p_1) + (1 \times p_2) + (0 \times p_3)} = \frac{p_2}{qp_1 + p_2}$$

Thus the probability that the automobile is behind Door #2 after the host has opened Door #3 is greater than one half if,

$$qp_1 < p_2 .$$

In this case, the contestant should change his choice. In the normal case where the original probabilities of the three arrangements, p_i, are equal and the host chooses randomly between Door #2 and Door #3, the revised probability of Door #2 concealing the automobile will be greater than one half. Therefore, unless the contestant has a strong a priori belief that Door #1

conceals the automobile, and/or believes that the host will prefer to open Door #3 before Door #2, he should switch his choice.



As the above diagram illustrates, if the original probabilities of all three arrangements are equal and the host chooses randomly which door to open, then of the one half of the sample space covered by opening Door #3, two thirds falls in the region occupied by arrangement GAG. Therefore, if the host opens Door #3, Door #2 becomes twice as likely as Door #1 to conceal the automobile.

Summary of notions

Random experiment is every finite process whose result is not determined in advance by conditions upon whose it runs and which is at least theoretically infinitely repeatable.
Possible results of random experiment are called **elementary events.**
A set of all elementary events we call **a sample space**.
Probability measure is real function defined upon subset system of the sample space which is non-negative normed and σ-aditive.
Conditional probability is a probability of event with conditional that some other (not impossible) event happened.
A and B events are **independent** if intersection probability of these two events is equal to a product of individual event probabilities.
Total probability theorem gives us a way how to determine probability of some event A while presuming that complete set of mutual disjoint events is given.
Bayes's theorem allows us to determine conditional probabilities of individual events in this complete set while presuming that A event happened.

Questions

1. How we determine probability of two events union?
2. How we determine probability of two events intersection?
3. When are two events independent?

Problems

Example 1: Suppose that a man and a woman each have a pack of 52 playing cards. Each draws a card from his/her pack. Find the probability that they each draw the ace of clubs.

{*Answer:* independent events - 0.00037}

Example 2: A glass jar contains 6 red, 5 green, 8 blue and 3 yellow marbles. If a single marble is chosen at random from the jar, what is the probability of choosing a red marble? a green marble? a blue marble? a yellow marble?

{*Answer:* P(red)=3/11, P(green)=5/22, P(blue)=4/11, P(yellow)=3/22}

Example 3: Suppose there are two bowls full of cookies. Bowl #1 has 10 chocolate chip cookies and 30 plain cookies, while bowl #2 has 20 of each. Fred picks a bowl at random, and then picks a cookie at random. We may assume there is no reason to believe Fred treats one bowl differently from another, likewise for the cookies. The cookie turns out to be a plain one. How probable is it that Fred picked it out of bowl #1?

{*Answer*: Conditional probability - 0.6}

Example 4: Suppose a certain drug test is 99% accurate, that is, the test will correctly identify a drug user as testing positive 99% of the time, and will correctly identify a non-user as testing negative 99% of the time. This would seem to be a relatively accurate test, but Bayes's theorem will reveal a potential flaw. Let's assume a corporation decides to test its employees for opium use, and 0.5% of the employees use the drug. We want to know the probability that, given a positive drug test, an employee is actually a drug user.

{*Answer:* Bayes's theorem - 0.3322}

3. RANDOM VARIABLES

Study time: 80 minutes

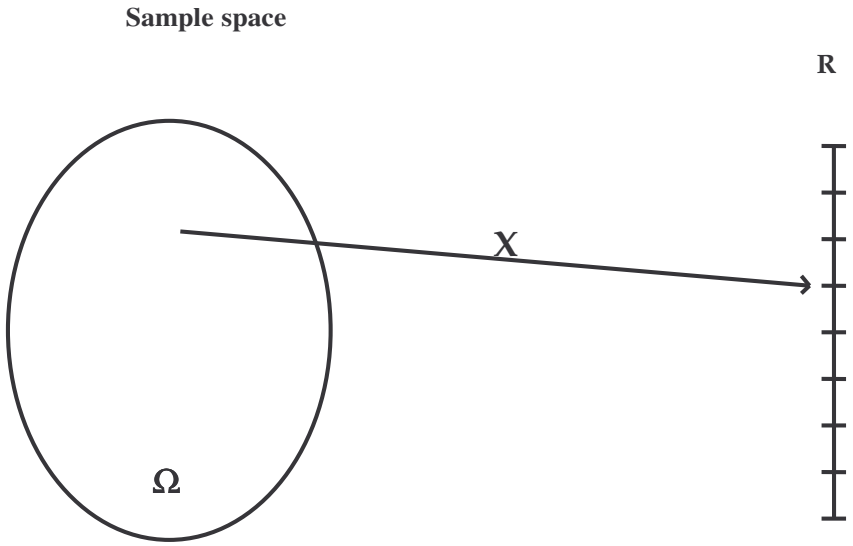
Aim - you will be able to

- describe the random variable by the distribution function
- characterize a discrete and a continuous random variable
- understand the hazard rate function
- determine the numerical characteristics of the random variable
- transform the random variable

Explication

3.1. Definition of a random variable

Let us consider a probability space (Ω, S, P). A **random variable X** (RV) on a sample space Ω is such real function $X(\omega)$ that for each real $x \in R$ is the set $\{\omega \in \Omega \mid X(\omega < x)\} \in S$, i.e. it is a random event. Therefore, the random variable is such function $X: \Omega \rightarrow R$ that for each $x \in R$ holds: $X^{-1}((-\infty, x)) = \{\omega \in \Omega \mid X(\omega < x)\} \in S$. From definition implies that we can determine a probability of $X(\omega) < x$ for any $x \in R$.



A group of all values $\{x = X(\omega),\ \omega \in \Omega\}$ is called **sample space**.

3.2. Distribution function

Definition: The distribution function of a random variable X is written $F(t)$ and, for each $t \in R$ has the value:

$$F(t) = P\{X \in (-\infty, t)\} = P(X \leq t).$$

Properties of the probability distribution function:

1. $0 \leq F(x) \leq 1$ for $-\infty < x < +\infty$

2. the distribution function is a monotonic increasing function of x, i.e. $\forall x_1, x_2 \in R: x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$

3. the distribution function F(x) is left-continuous

4. $\lim_{x \to +\infty} F(x) = 1;\ \lim_{x \to -\infty} F(x) = 0$

5. $\forall$ a, b $\in$ R; a < b : $P(a \leq X < b) = F(b) - F(a)$

6. $P(x = x_0) = \lim_{x \to x_0 +} F(x) - F(x_0)$

If the range of the random variable function is discrete, then the random variable is called a discrete random variable. Otherwise, if the range includes a complete interval on the real line, the random variable is continuous.

3.3. Discrete random variable

We speak about discrete random variable if a random variable is from some finite and enumerable set. The most often it is an integer random variable e.g. a number of student that entered the main building of VŠB TUO before midday (0,1,2,...), a number of house members (1,2,3,...), a number of car accidents during one day on a Prague - Brno highway (0,1,2,...), etc..

Definition
We say that a random **variable X has a discrete probability distribution** when:

1. $\exists$ finite or enumerable set of real numbers M={ x_1, ... , $x_{n, ...}$} that
   $P(X = x_i) > 0$ i = 1, 2, ...
2. $\sum_i P(X = x_i) = 1$

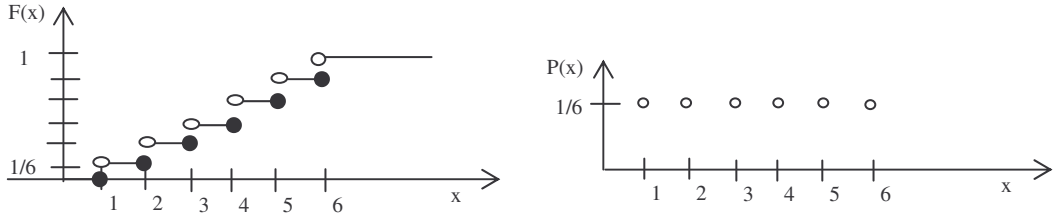
Function P(X = x_i) $\Leftrightarrow$ P(x_i) is called **probability function of random variable** X.
A distribution function of such distribution is a step function with steps in x_1, ... , $x_{n, ..}$
For a distribution function of discrete random variable it holds:

$$F(x) = \sum_{x_i < x} P(X = x_i)$$

Example

A throwing dice, X … a number of obtained dots

x_i	$P(X = x_i)$	$F(x_i)$
1	1/6	0
2	1/6	1/6
3	1/6	2/6
4	1/6	3/6
5	1/6	4/6
6	1/6	5/6



3.4. Continuous random variable

If a random variable may have any value from a certain interval we speak about a random variable with continuous distribution. As an example we can name a service life of a product (0, ∞), the length of specific object etc. In such case, we can use a density function as well as distribution function to describe a distribution of random variable.

Definition
Random variable has a **continuous probability distribution** when a function f(x) exists that

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad \text{pro } -\infty < x < \infty$$

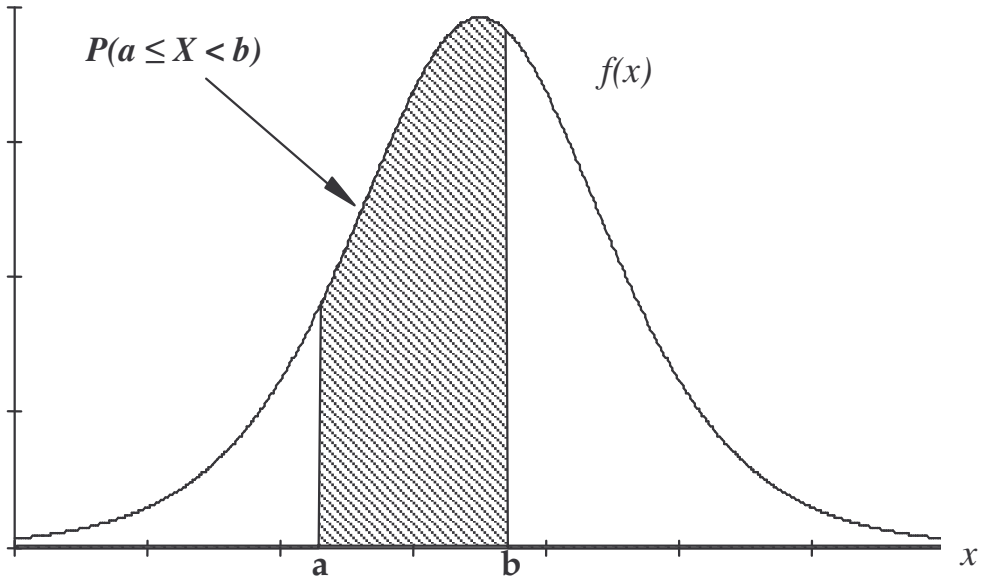
Function *f(x)* is called a **probability density function** of continuous random variable X. It is non-negative real function.
We can show that in all points where a derivation of distribution function exists it holds:

$$f(x) = \frac{dF(x)}{dx}$$

If we know a distribution function we can easily determine a probability density function vice versa.

The area below *f(x)* spline for $x \in < a;b >;\ (a,b \in R)$ in any interval is the probability that X will gain the value of this interval. It also fully corresponds with our density definition.

$$P(a \le X < b) = F(b) - F(a) = \int_{-\infty}^{b} f(t)\,dt - \int_{-\infty}^{a} f(t)\,dt = \int_{a}^{b} f(t)\,dt$$



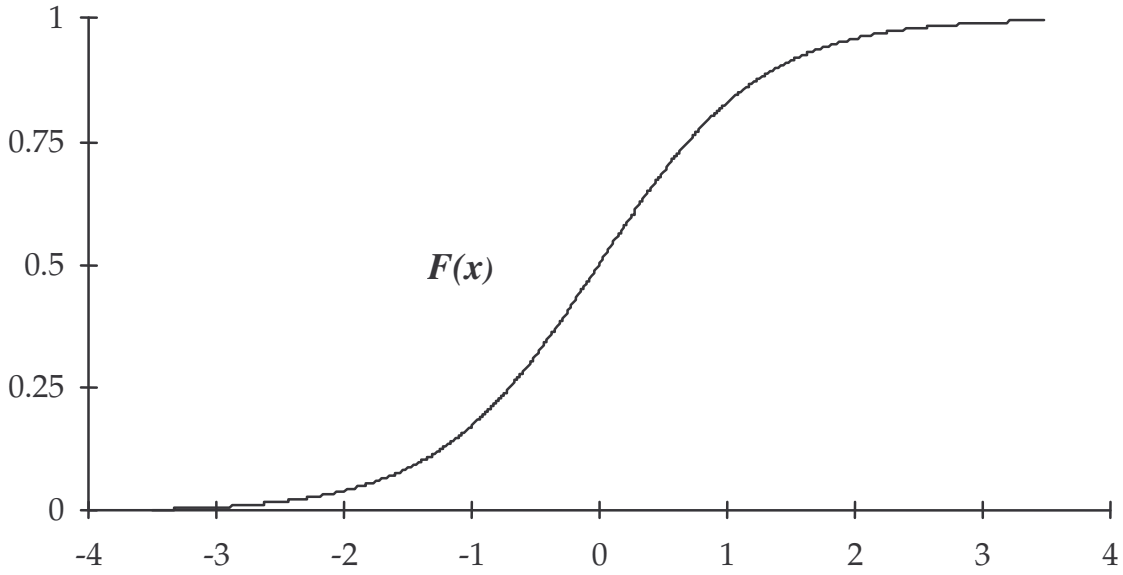
One of attributes for each density probability is the fact that the whole area under curve is equal to one. It is analogical to a discrete random variable where the sum of probabilities for all possible results is also equal to one. We can describe this attribute by the following equation:

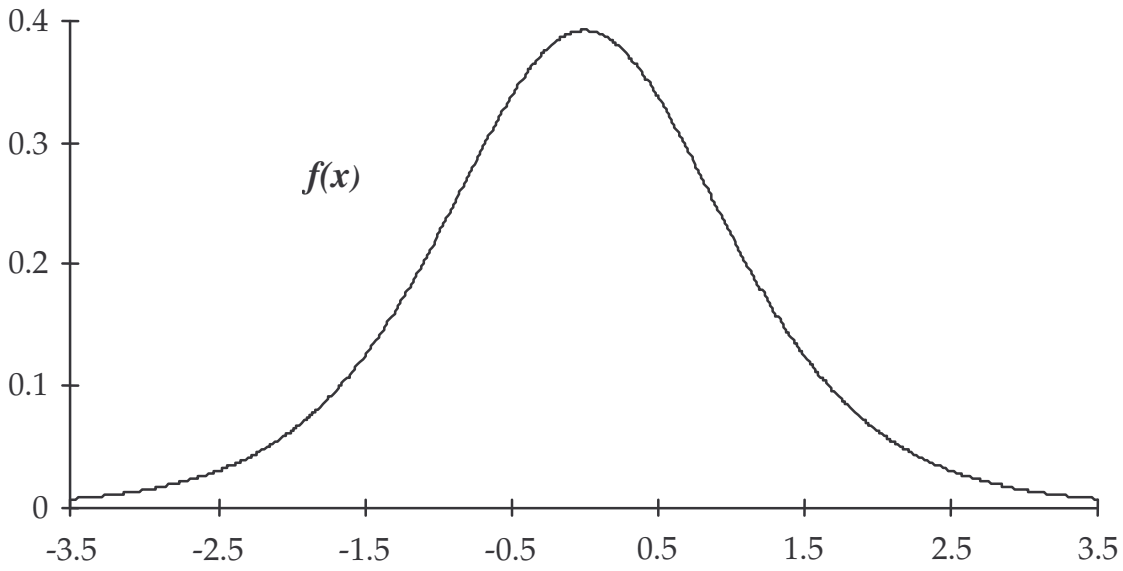
$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

Example

Logistic probability distribution has a following distribution function *F(x)* and a probability density *f(x):*

$$F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \qquad f(x) = \frac{\beta_1 e^{-(\beta_0+\beta_1 x)}}{(1+e^{-(\beta_0+\beta_1 x)})^2}$$





3.5. Failure rate

Let X be a non-negative random variable with continuous distribution. Then, we define a **failure rate** for $F(t) < 1$

$$\lambda(t) = \frac{f(t)}{1 - F(t)}.$$

We can easily derive the following formula:

$$\lambda(t) = \lim_{\Delta t \to 0+} \frac{P(t < X \le t + \Delta t | X > t)}{\Delta t} = \frac{f(t)}{1 - F(t)}$$

Let X be a **mean time to failure** of any system. Then, the failure rate expresses that if in the t-time there was no failure the probability of failure in a small following time Δt is approximately $\lambda(t).\Delta t$:

$$P(t < X \le t + \Delta t | X > t) \approx \frac{f(t)}{1 - F(t)} \Delta t = \lambda(t).\Delta t$$

The failure rate characterizes the probability distribution of non-negative random variable.

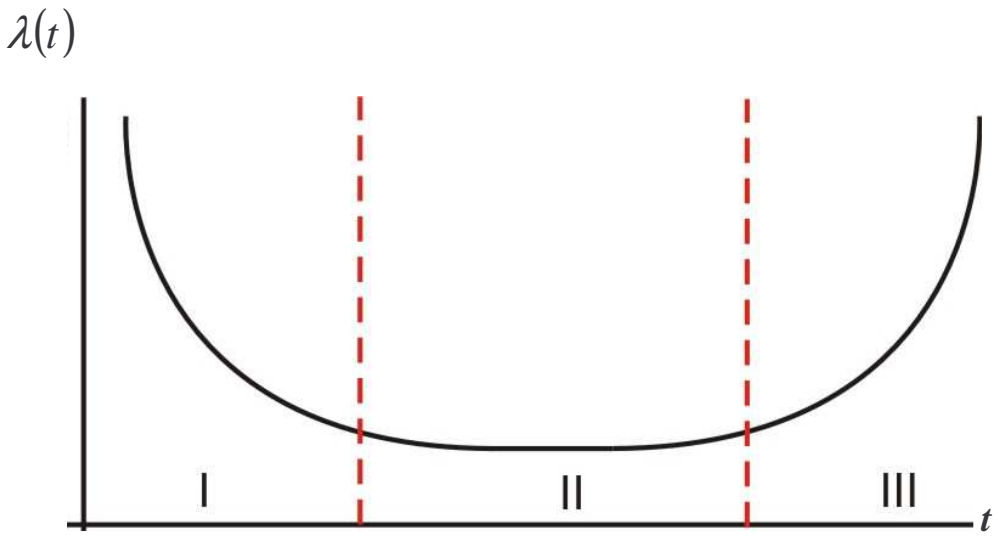
Table 1 shows the mutual conversions between $f(t), F(t), \lambda(t)$:

	$F(t)$	$f(t)$	$\lambda(t)$
$F(t)$	$F(t)$	$\int_0^t f(x)dx$	$1 - \exp\left[- \int_0^t \lambda(x)dx \right]$
$f(t)$	$\frac{dF(t)}{dt}$	$f(t)$	$\lambda(t) \cdot \exp\left[- \int_0^t \lambda(x)dx \right]$
$\lambda(t)$	$\frac{\frac{dF(t)}{dt}}{1 - F(t)}$	$\frac{f(t)}{1 - \int_0^t f(x)dx}$	$\lambda(t)$

Table 1

- **The most often graphical interpretation of failure rate**

Let a random variable X be a **mean time to failure** of any system. Then, a typical form of failure rate is shown in the following figure. The curve in this figure is called the **bathtub curve.**



I … The first part is a decreasing failure rate, known as early failures or **infant mortality.**

II … The second part is a constant failure rate, known as random failures.

III … The third part is an increasing failure rate, known as wear-out failures.

3.6. Numerical characteristics of random variable

The probability distribution of each random variable X is fully described by its distribution function F(x). In many cases we can summarize the total information to several numbers. These numbers are called the **numerical characteristics of the random variable X.**

1. **Moments**
   r-th general moment is denoted $\mu_r' = EX^r$ r = 0,1,2, …

   discrete RV: $\mu_r' = \sum_i x_i^r \, . \, P(x_i)$

   continuous RV: $\mu_r' = \int_{-\infty}^{\infty} x^r . f(x) \, dx$ r = 0,1,2, …

if stated progression or integral tend absolutely.

r-th central moment is denoted $\mu_r = E[X - EX]^r$ r = 0,1,2, …

discrete RV: $\mu_r = \sum_i [x_i - EX]^r \, . \, P(x_i)$

continuous RV: $\mu_r = \int_{-\infty}^{\infty} (x - EX)^r \, . f(x) \, dx$

if stated progression or integral tend absolutely.

2. Expected value (mean) $EX = \mu_1'$

discrete RV: $EX = \sum_i x_i \,.\, P(X = x_i)$

continuous RV: $EX = \int_{-\infty}^{\infty} x.\, f(x)dx$

Properties:

1. $E(aX + b) = a.\, EX + b \quad a,b \in R$
2. $E(X_1 + X_2) = EX_1 + EX_2$
3. X_1, X_2 ... independent RV $\Rightarrow E(X_1.\, X_2) = EX_1 \,.\, EX_2$
4. $Y = g(X);$ g(X) is a continuous function: $EY = E(g(X))$

   Y is a continuous RV: $EY = \int_{-\infty}^{\infty} g(x).\, f(x)\, dx$

   Y is a discrete RV: $EY = \sum_i g(x_i) \,.\, P(X = x_i)$

3. Variance $DX = \mu_2 = E(X - EX)^2 = EX^2 - (EX)^2$

discrete RV: $DX = \sum_i x_i^2 . P(x_i) - (\sum_i x_i .\, P(x_i))^2$

continuous RV: $DX = \int_{-\infty}^{\infty} x^2 .\, f(x)dx - (\int_{-\infty}^{\infty} x.\, f(x)dx)^2$

Properties::

1. $D(aX + b) = a^2 .\, DX$
2. X_1, X_2 ... independent $\Rightarrow D(X_1 + X_2) = DX_1 + DX_2$

4. **Standard deviation** $\sigma_x = \sqrt{DX}$

5. **Skewness** $a_3 = \mu_3 / \sigma_x^3$

Is a level of symmetry for the given probability distribution and it is hold:

$a_3 = 0$ … symmetrical distribution

$a_3 < 0$ .... negative skewed set

$a_3 > 0$ .... positive skewed set

6. **Kurtosis** $a_4 = \mu_4 / \sigma_x^4$

Is a level of kurtosis (flatness):

$a_4 = 3$ .... normal kurtosis (i.e. kurtosis of a normal distribution)

$a_4 < 3$ .... lower kurtosis than normal distribution one (flatter)

$a_4 > 3$ .... greater kurtosis than normal distribution one (sharper)

7. Quantiles

$p \in (0,1)$

x_p ... 100p% quantile $\quad x_p = \sup\{x \mid F(x) \leq p\}$

continuous RV: $\quad F(x_p) = p$

Special types of the quantiles:

$x_{0,5}$… 50% quantile is called a median
$x_{0,25}$ and $x_{0,75}$… 25% quantile is called a lower quartile and 75% quantile is called an upper quartile
$x_{k/10}$ … k =1,2,… 9 the k-th decile
$x_{k/100}$ … k =1,2,… 99 the k-th percentile

8. Mode

The mode $\hat{x}$ of a discrete RV X is such value that holds:

$$P(X = \hat{x}) \geq P(X = x_i) \qquad i = 1, 2, ...$$

It means that the mode is a value in which the discrete RV comes with the biggest probability.

The mode $\hat{x}$ of a continuous RV X is such value that holds:

$$f(\hat{x}) \geq f(x) \qquad \text{pro } -\infty < x < \infty$$

It is a value where the probability density has a maximum value.

4. RANDOM VECTOR

Study time: 60 minutes

Aim - you will be able to

- describe a random vector and its joint distribution
- explain the notions of marginal and conditional probability distribution
- explain a stochastic independence of random variables

Explication

4.1. Random vector

For continuous random variables, the joint distribution can be represented either in the form of a distribution function or of a probability density function.

$$F(x_1, x_2, ..., x_n) = P(X_1 < x_1, X_2 < x_2, ..., X_n < x_n), \quad F: R^n \to R$$

$$f(x_1, x_2, ..., x_n) = \frac{\partial^n F(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}$$

The two forms are again equivalent. In terms of the joint probability density function, the joint distribution function of $X_1,...,X_n$ is

$$F(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_n} ... \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(t_1, t_2, ..., t_n)\, dt_1\, dt_2\, ...dt_n.$$

Although in theory the joint distribution of a discrete variable with a continuous variable does exist, there is no practical algebraic formulation of such a distribution. Such distributions are only represented in conditional form.

4.2. Marginal distribution

Definition

Let $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ be a random vector. The random vector $\boldsymbol{Y} = (X_{i_1}, X_{i_2}, ..., X_{i_k})$, where $k < n$, $i_j \in \{1,2,...,n\}$, $i_u \neq i_v$ *pro* $u \neq v$, we called the **marginal random vector.** Especially,

X_i is the **marginal random variable** for every $i=1,2,\ldots,n$. The probability distribution of $\boldsymbol{Y}$ we called the **marginal probability distribution**.

Let $\boldsymbol{X} = (X_1, X_2)$ be a bivariate random variable with given distribution function $F(x_1, x_2)$.

$$F_1(x_1) = \lim_{x_2 \to +\infty} F(x_1, x_2) = F(x_1, +\infty) \ldots \text{marginal distribution function of random variable } X_1$$

$$F_2(x_2) = \lim_{x_1 \to +\infty} F(x_1, x_2) = F(+\infty, x_2) \ldots \text{marginal distribution function of random variable } X_2$$

For continuous random variables, the marginal probability density of one jointly distributed variable is found by integrating the joint density function with respect the other variable.

$$f_1(x_1) = \int_{x_2} f(x_1, x_2) dx_2 \quad \text{for } X_1$$

$$f_2(x_2) = \int_{x_1} f(x_1, x_2) dx_1 \quad \text{for } X_2 .$$

For discrete random variables, the marginal distributions are given by:

$$P_1(x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2) \ldots \text{ marginal distribution function of } X_1$$

$$P_2(x_2) = \sum_{x_1} P(X_1 = x_1, X_2 = x_2) \ldots \text{ marginal distribution function of } X_2$$

4.3. Conditional distribution

The conditional distribution is the distribution of one variable at a fixed value of the other jointly distributed random variable. For two discrete variables, the conditional distribution is given by the ratio of the joint probabilities to the corresponding marginal probability.

$$p(x_1 \mid x_2) = P(X_1 = x_1 \mid X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P_2(x_2)} = \frac{p(x_1, x_2)}{P_2(x_2)} .$$

For continuous random variables, the conditional densities are given analogously by the ratio of the joint density to the corresponding marginal density.

$$f(x_1 | x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} .$$

$f_2(x_2) = \int_{-\infty}^{+\infty} f(x_1, x_2) dx_1$ is the corresponding marginal density of X_2.

4.4. Independence of Random Variables

Definition

$X_1 \ldots X_n$ are **mutually independent** $\Leftrightarrow$ the random events $\{X_i < x_i\}$, ($i=1,2,\ldots,n$, where $x_i \in R$) are mutually independent.

Therefore, $X_1 \ldots X_n$ are mutually independent $\Leftrightarrow$ $F(x_1,\ldots,x_n) = F_1(x_1) \ldots\ldots F_n(x_n)$.

It is true, because
$F(x_1,\ldots,x_n) = P(X_1<x_1, \ldots, X_n<x_n) = P(X_1< x_1) \,.\, P(X_2<x_2) \ldots\ldots P(X_n<x_n) = F_1(x_1) \,.\, F_2(x_2) \ldots\ldots F_n(x_n)$.

This implies the following rule::
$X_1 \ldots X_n$ are mutually independent $\Leftrightarrow f(x_1,\ldots,x_n) = f_1(x_1) \,.\, \ldots\ldots f_n(x_n)$.

Example:
X_1, X_2 are mutually independent. Determine the variance $X_1 + X_2$.

In general, if X_1 and X_2 are not independent, the variance of their sum is given by

$$D(X_1 + X_2) = D(X_1) + D(X_2) + 2\,Cov(X_1, X_2)$$

where the covariance of X_1 and X_2 is defined by

$$Cov(X_1, X_2) = E[(X_1 - E(X_1))(X_2 - E(X_2))]$$

When X_1 and X_2 are independent, the covariance is zero.
An alternate expression for the covariance similar to that for the variance and simpler for computation is

$$Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2).$$

Correlation coefficient

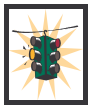
The correlation coefficient measures the strength of the relation between two random variables, X_1 and X_2. The correlation coefficient is defined by

$$\rho_{X_1X_2} = \frac{Cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}.$$

The correlation coefficient properties are:

1. $-1 \le \rho \le 1$
2. $\rho(X,Y) = \rho(Y,X)$

The correlation assumes values between -1 and 1. A value close to 1 implies a strong positive relationship, a value close to -1 implies a strong negative relationship, and a value close to zero implies little or no relationship.



Solved example

Imagine that we will repeat the trial for three times (we known the success probability, e.g. coin throws).

Lets write all possible combinations: (S - success, F - false):

{ FFF; SFS; SSF; FSS; FSF; FFS; SFF; SSS }

Now we specify the following random variables:

Y ... a number of attempts to the first success
Z ... a number of the following successes

a) determine the probability function P(Y), P(Z)
b) set the joint probability function Y, Z
c) determine the marginal distribution function and P(Y | Z), P(Z | Y)

Solution:

ada) Y and Z are the discrete RV and that is why Y and Z can gain the values: 0, 1, 2, 3
Let's name all element events of the sample space:

A1 ... FFF	$P(A1) = (1 - p)^3$
A2 ... SFS	$P(A2) = p^2.(1 - p)$
A3 ... SSF	$P(A3) = p^2.(1 - p)$
A4 ... FSS	$P(A4) = p^2.(1 - p)$
A5 ... FSF	$P(A5) = p.(1 - p)^2$
A6 ... FFS	$P(A6) = p.(1 - p)^2$
A7 ... SFF	$P(A7) = p.(1 - p)^2$
A8 ... SSS	$P(A8) = p^3$

For our calculation we use the fact that the F and S variables are independent.

Y ... a number of attempts to first success			
0	1	2	3
SFS, SSF, SFF, SSS	FSS, FSF	FFS	FFF

Z ... a number of following successes			
0	1	2	3
FFF	SFS, FSF, FFS, SFF	SSF,FSS	SSS

Since A1, ... , A8 events are disjoint we can simply determine the probablity function (p=0.5).

Y ... a number of attempts to first success			
P(Y= 0)	P(Y= 1)	P(Y= 2)	P(Y= 3)
0.5	0.25	0.125	0.125

Z ... a number of following successes			
P(Z = 0)	P(Z = 1)	P(Z = 2)	P(Z = 3)
0.125	0.5	0.25	0.125

adb) we will proceed in the same way like we did in probability function finding
Z

Y	0	1	2	3
0	-	SFS, SFF	SSF	SSS
1	-	FSF	FSS	-
2	-	FFS	-	-
3	FFF	-	-	-

		Z		
Y	0	1	2	3
0	0	0.25	0.125	0.125
1	0	0.125	0.125	0
2	0	0.125	0	0
3	0.125	0	0	0

adc) marginal probability functions - P(Y), P(Z)

			Z		
Y	0	1	2	3	P(Y)
0	0	0.25	0.125	0.125	0.5
1	0	0.125	0.125	0	0.25
2	0	0.125	0	0	0.125
3	0.125	0	0	0	0.125
P(Z)	0.125	0.5	0.25	0.125	1

$P(Y \mid Z) = P(Y \wedge Z) / P(Z)$

			Z	
Y	0	1	2	3
0	0	0.5	0.5	1
1	0	0.25	0.5	0
2	0	0.25	0	0
3	1	0	0	0

$P(Z \mid Y) = P(Y \wedge Z) / P(Y)$

			Z	
Y	0	1	2	3
0	0	0.5	0.25	0.25
1	0	0.5	0.5	0
2	0	1	0	0
3	1	0	0	0

Summary of notions

Random variable X is a real function which can be characterized by a **distribution function** *F(t).*

Distribution function is a function that assigns to each rea number a probability that the random variable will be less then this real number. Distribution function has some general properties like $\forall$ a, b $\in$ R; a < b platí $P(a \leq X < b) = F(b) - F(a)$.

According values the random variable may become we distinguish **continuous** and **discrete variable.**

The discrete random variable is also characterized by a **probability function**, the continuous one by a **density function**.

In many cases, it is useful to cover the whole information about random variable into several numbers that characterize some properties of random variable while allowing the comparison of different random variables. These numbers are called the **numerical characteristics** of random variable.

A **random vector** is a vector consisted of random variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ that is characterized by **joint distribution function.**

From joint distribution function of random vector we can easily determine a **marginal probability distribution** of particular random variables the vector is composed of.

Questions

1. What is a mutual relationship between a distribution function and probability function of discrete random variable?
2. What is a mutual relationship between a distribution function and probability density function of continuous random variable?
3. What is a median and a mode?
4. Explain term of conditional probability distribution.
5. Explain term of stochastic independence of random variables.
6. What does a value of correlation coefficient tell us?

Problems

Example 1: Let Y be a continuous variable defined by a probability density function:

$$f(y) = \begin{cases} c.(1+y).(1-y); & -1 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find a constant c, a distribution function, an expected value and a variance of this variable.

{*Answer:* c=0.75; $F(y) = 0.25\,(3y - y^3+2)$; EY = 0; DY = 0.2}

Example 2: Let random variable W is defined as a linear transformation of random variable Y, defined in previous example.

$$W = 5Y + 6$$

Find a probability density function, a distribution function, an expected value and a variance of random variable W.

{*Answer:* $f(w) = -\frac{3}{500}\,(w^2-12w+11)$; $F(w) = 0.25\,[3(\frac{w-6}{5}) - (\frac{w-6}{5})^3+2]$; EW = 6; DW = 5}

Example 3: Let random variable Z be defined as:

$$f(z) = 1 / [\,(1+e^{z}).(1+e^{-z})\,]; \qquad -\infty < z < \infty$$

Find a distribution function of random variable Z.

{*Answer:* $F(z) = \frac{e^z}{1+e^z}$ }

5. SOME IMPORTANT PROBABILITY DISTRIBUTIONS

5.1. Discrete Probability Distributions

Study time: 50 minutes

Aim - you will be able to

- characterize Bernoulli trials and types of discrete distributions
- characterize Poisson process and Poisson distribution
- describe contexture in between discrete distributions

Explication

A lot of discrete random variables exist and now we summarize basic information about the most common discrete variables.

Bernoulli trials:

- a sequence of Bernoulli trials is defined as a sequence of random events which are mutually independent and which have only two possible outcomes (e.g. success-nonsuccess, 1-0)
- probability of event occurrence (a success) p is constant in any trial

$$P\{\text{Trial 'i'} = \text{"Success"}\} = p$$

Binomial random variable:

The most natural random variable to define on the sample space of Bernoulli trials is the number of successes. Such a random variable is called a binomial random variable. If X is the number of successes in n Bernoulli trials where the probability of success at each trial is p, then we represent the distribution of X by the short-hand notation:

$$X \rightarrow B(n, p)$$

where B indicates that X has a binomial distribution and n and p are the parameters determining which particular distribution from the binomial family applies to X.

The probability distribution of a binomial random variable can be expressed algebraically as:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k};\ 0\le k\le n$$

$$EX=\sum_{k=0}^{n}k\cdot P(X=k)=\sum_{k=0}^{n}k\cdot\frac{n!}{(n-k)!k!}\cdot p^k(1-p)^{n-k}=n\cdot p\sum_{k=0}^{n}\frac{(n-1)!}{(n-k)!(k-1)!}\cdot p^{k-1}\cdot(1-p)^{n-k}=n\cdot p$$

$DX = EX^2 - (EX)^2$

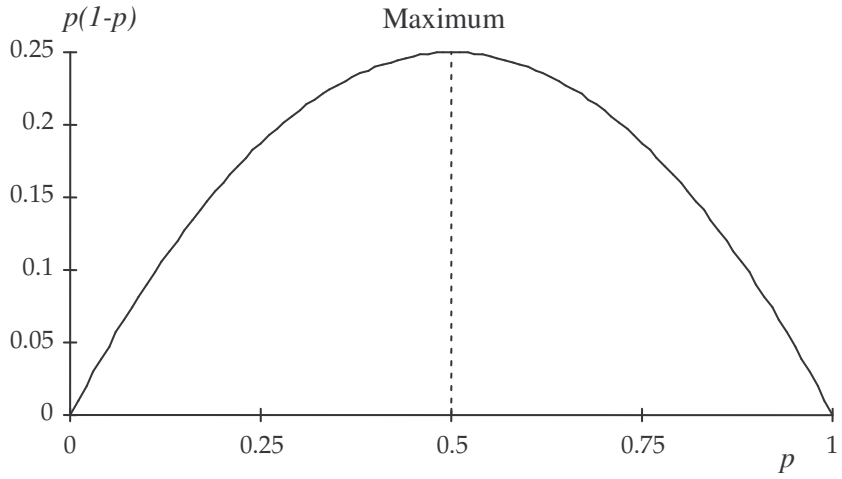
$$EX^2=\sum_{k=1}^{n}k^2\cdot P(X=k)=\sum_{k=1}^{n}k\cdot(k-1)\cdot\frac{n!}{(n-k)!k!}\cdot p^k\cdot(1-p)^{n-k}+EX=$$

$$=n\cdot(n-1)\cdot p^2\cdot\sum_{k=2}^{n}\frac{(n-2)!}{(n-k)!(k-2)!}\cdot p^{k-2}\cdot(1-p)^{n-k}+EX=$$

$$=n\cdot(n-1)\cdot p^2+n\cdot p=(np)^2-np^2+np$$

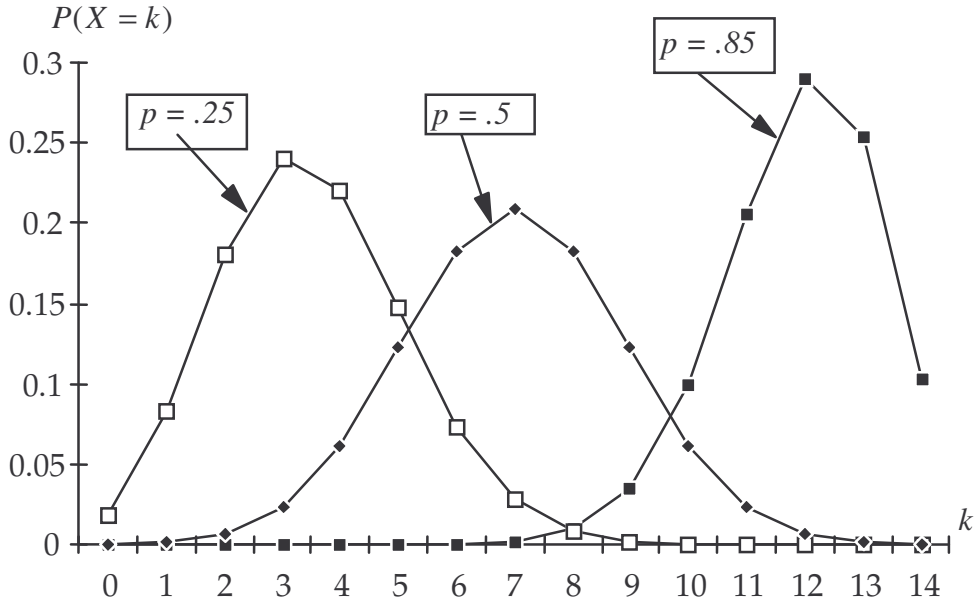
$DX = EX^2 - (EX)^2 = n.p.(1 - p)$

Notice that the variance of the binomial distribution is maximum when $p = 0.5$.



Example:

Some examples of binomial distributions for $n = 14$ trials are illustrated below. Notice that as p, the probability of success at each trial increases, the location of the distribution shifts to higher values of the random variable. Also notice that when p = 0.5, the distribution is symmetric around 7.5.



Geometric distribution:

This distribution has a single parameter, p, and we denote the family of geometric distributions by

$$X \rightarrow G(p)$$

G(p) … the geometric random variable is defined as the number of trials until a success occurs or until the first success

The probability distribution for a geometric random variable is:

$$P(X = k) = p(1-p)^{k-1};\ 1 \le k < \infty$$

The expression for the mean of the geometric distribution is

$$EX = \sum_{k=1}^{\infty} k \cdot P(X = k) = \sum_{k=1}^{\infty} k \cdot p \cdot (1-p)^{k-1} = p \cdot \sum_{k=1}^{\infty} k \cdot (1-p)^{k-1} = p \cdot \frac{\partial}{\partial(1-p)} \cdot \left(\sum_{k=1}^{\infty} (1-p)^{k} \right) = \frac{1}{p}$$

Note: By first evaluating the series and then taking its derivative the result is obtained.

The mean number of Bernoulli trials until the first success is the inverse of the success probability at each trial, again an entirely intuitive result. That is, if 10% of the trials are successful, on average it will take ten trials to obtain a success.

To find the variance, we first evaluate the expected value of X^2 and then modify the expression using the same technique as for the binomial case. We note that the expression now has the form of the second derivative of the same geometric series we evaluated for the mean. Taking derivatives of this evaluated expression, we obtain

$$E(X^2) = \sum_{k=1}^{\infty} k^2\, p(1-p)^{k-1}$$

$$= p(1-p)\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} + p\sum_{k=2}^{\infty} k(1-p)^{k-1}$$

$$= p(1-p)\frac{\partial^2 \sum_{k=1}^{\infty}(1-p)^k}{\partial^2(1-p)} + p\frac{\partial \sum_{k=1}^{\infty}(1-p)^k}{\partial(1-p)}$$

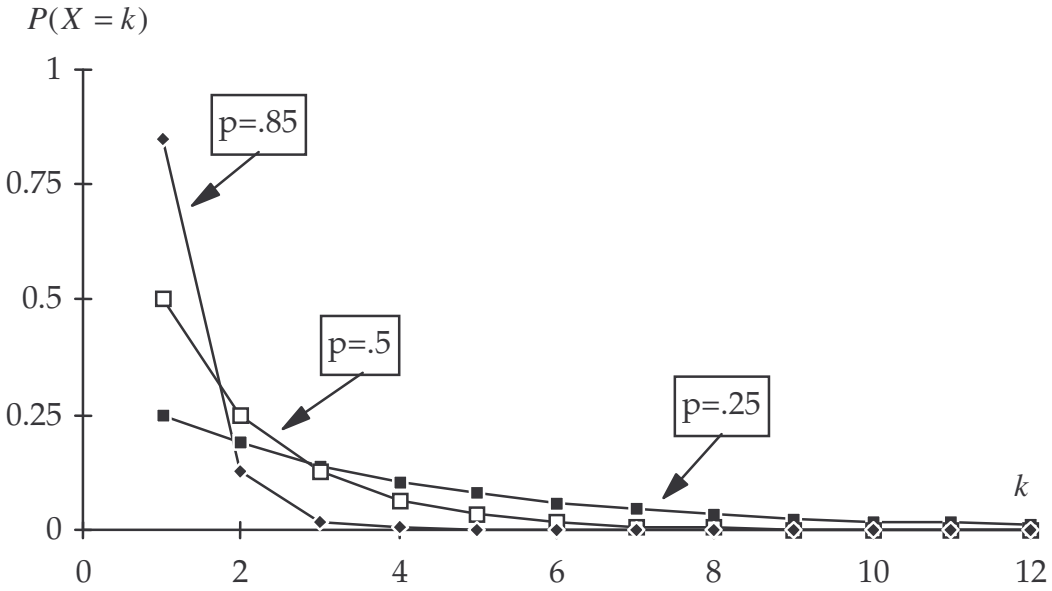
$$= p(1-p)\frac{\partial^2\left(\frac{1-p}{p}\right)}{\partial^2(1-p)} + p\frac{\partial\left(\frac{1-p}{p}\right)}{\partial(1-p)} = \frac{2(1-p)}{p^2} + \frac{1}{p}$$

From the mean and expected value of X^2 we can derive the variance.

$$DX = EX^2 - (EX)^2 = \frac{1-p}{p^2}$$

Example:

Some examples of geometric distributions are illustrated below. Not surprisingly, the probability of long sequences without success decreases rapidly as the success probability, p, increases.



Negative binomial random variable

The negative binomial distribution has two parameters, k and p and is denoted by

$$X \rightarrow NB(k, p)$$

The negative binomial is the number of Bernoulli trials until the k^{th} success.

The negative binomial distribution is:

$$P(X = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k};\ k \le n < \infty$$

The mean and variance of the negative binomial distribution can be computed easily by noting that a negative binomial random variable with parameters k and p is just the sum of k independent geometric random variables with parameter p. Independence the geometric random variables follow from the independence assumption of sequences of Bernoulli trials. Thus if

$$W_i \rightarrow G(p); 1 \le i \le k$$

then

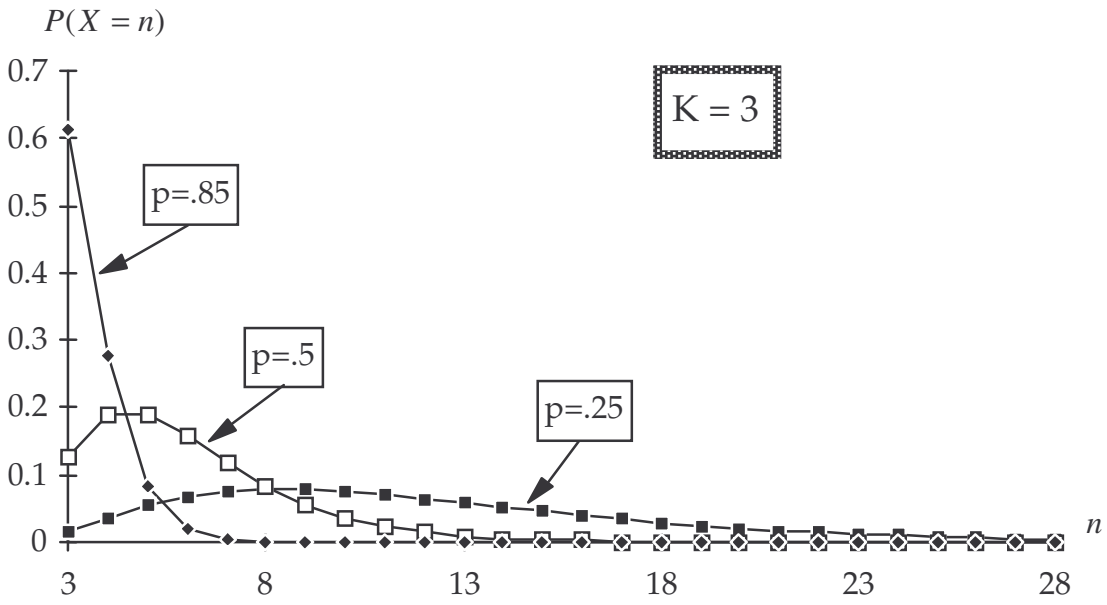
$$X = \sum_{i=1}^{k} W_i$$

$$E(X) = \sum_{i=1}^{k} E(W_i) = \frac{k}{p}$$

$$D(X) = \sum_{i=1}^{k} D(W_i) = \frac{k(1-p)}{p^2}$$

Example:

The following chart illustrates some examples of negative binomial distributions for k=3. Notice that for values of p near 0.5, the distribution has a single mode near 5. This mode moves away from the origin and diminishes in magnitude as p decreases indicating an increase in variance for small p. The negative binomial distribution has a shape similar to the geometric distribution for large values of p.



***Note*:**
Comparison of Binomial and Negative Binomial Distributions

It is interesting to compare the distributions of binomial and negative binomial random variables. Notice that except for the combinatorial term at the beginning of each distributional expression, the portion contributed by the probability of single sample space elements is identically $p^k (1 - p)^{n-k}$.

Binomial distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}; 0 \le k \le n$$

For the binomial, the number of trials (n) is fixed and the number of successes (k) is variable.

Negative binomial distribution

$$P(X = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}; k \le n < \infty$$

For the negative binomial, the number of successes (k) is fixed and the number of trials (n) is variable.

Poisson process

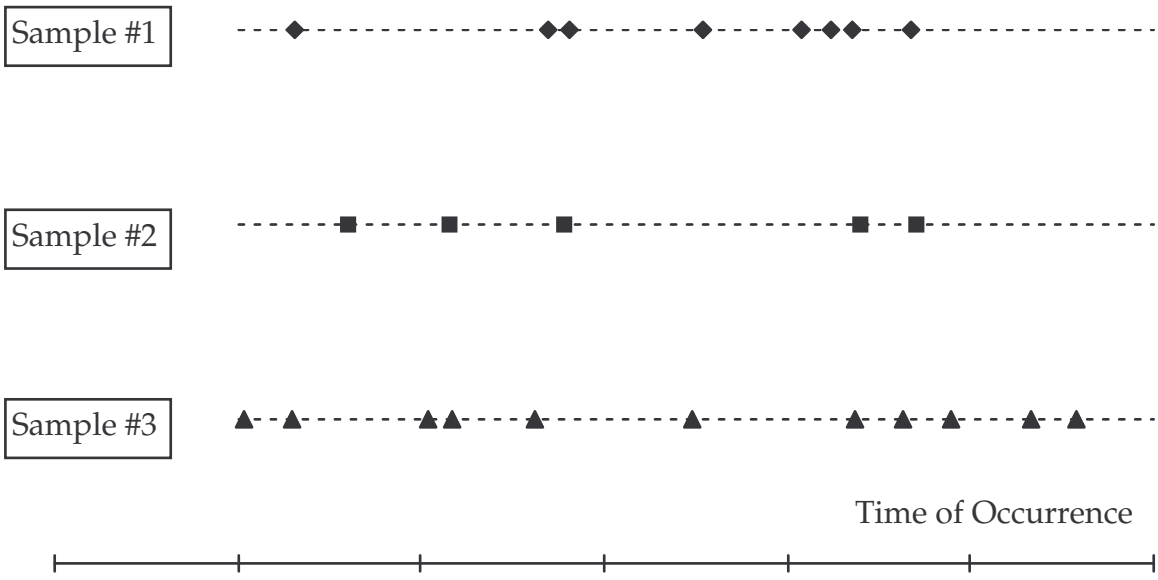
The Poisson process is a second general type of sample space model which is widely applied in practice. The Poisson process may be viewed as the continuous time generalization of a sequence of Bernoulli trials, sometimes called the Bernoulli process. The Poisson process describes the sample space of randomly occurring events in some time interval. The Poisson process assumes that the **rate at which events occur is constant** throughout the interval or region of observation and those events occur **independently** of each other.
λ … lambda $\rightarrow$ the rate at events occur

However, events occurring over some region can also be modeled by a Poisson process. The appearance of defects in some product, or mold on the surface of a leaf under certain conditions could follow a Poisson process.

Examples:

- Customers arriving at a bank to transact business.
- patients arriving at a clinic for treatment
- telephone inquiries received by a government office, *etc.*

Some examples of elements of the sample space for a Poisson process are illustrated below. This is a complex difficult to characterize sample space. The number of elements in the sample space is uncountable infinite and in this sense continuous. Probabilities cannot be assigned to individual elements of this sample space, only to subsets.



The Poisson process describes events which occur randomly over some time interval or spatial region.

Poisson distribution

The Poisson distribution has a single parameter and therefore we denote this random variable by the symbolic notation,

$$X \rightarrow P(\lambda t)$$

Consider a Poisson process that is observed for a time period t. Suppose the rate of occurrence of events is λ during the time period. Then the total rate of occurrence over the entire observation interval is λt. Now divide the interval t into n subintervals of equal length t/n. Occurrence of events in each of these intervals will be mutually independent at constant rate $\lambda t/n$. If n becomes large enough, the interval lengths, t/n, will become small enough that the probability of two events in one interval is effectively zero and the probability of one event is proportional to $\lambda t/n$. Then the distribution of the number of events in the total interval t can be approximated by the distribution of a binomial random variable with parameters n and $\lambda t/n$. Thus,

$$P(X = k) = \binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{n-k}$$

Taking the limit as n goes to infinity, this expression becomes

$$\lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{n-k} = \frac{(\lambda t)^k}{k!}\lim_{n\to\infty}\frac{n!}{(n-k)!\,n^k}\left(1-\frac{\lambda t}{n}\right)^{n-k} = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \quad ,$$

We can express the distribution of a Poisson random variable as:

$$P(X=k)=\frac{(\lambda t)^k e^{-\lambda t}}{k!};\ 0\le k<\infty$$

$$\sum_{k=0}^{\infty}P(X=k)=\sum_{k=0}^{\infty}\frac{(\lambda t)^k e^{-\lambda t}}{k!}=e^{-\lambda t}\sum_{k=0}^{\infty}\frac{(\lambda t)^k}{k!}=1$$

Calculation of mean:

$$E(X)=\sum_{k=0}^{\infty}kP(X=k)=\sum_{k=0}^{\infty}k\frac{(\lambda t)^k e^{-\lambda t}}{k!}=\lambda t\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}e^{-\lambda t}}{(k-1)!}=\lambda t$$

The expected value of X^2 is found using the Taylor series expansion of the exponential function as well as the same algebraic manipulation as was invoked for Bernoulli random variables.

$$E(X^2)=\sum_{k=0}^{\infty}k^2P(X=k)=\sum_{k=2}^{\infty}k(k-1)\frac{(\lambda t)^k e^{-\lambda t}}{k!}+\mathrm{E}(\mathrm{X})=$$

$$=(\lambda t)^2\sum_{k=2}^{\infty}\frac{(\lambda t)^{k-2}e^{-\lambda t}}{(k-2)!}+\lambda t=(\lambda t)^2+\lambda t$$

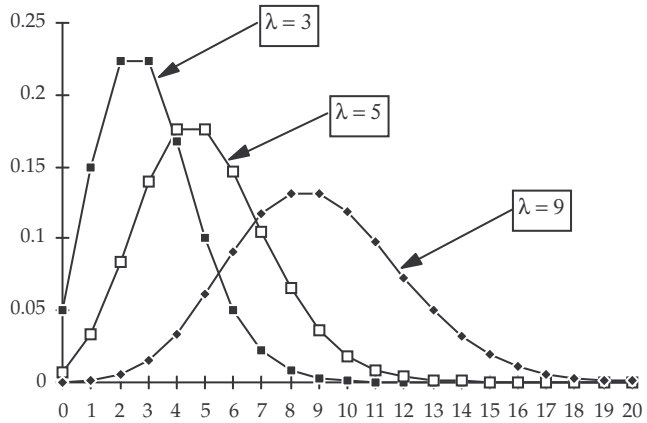
From this result, the variance follows directly.

$$D(X)=E[(X-E(X))^2]=\mathrm{E}(\mathrm{X}^2)-(\mathrm{E}(\mathrm{X}))^2=\lambda t$$

We notice that the Poisson distribution has the remarkable property that the variance is equal to the mean and by implication that the variance of the Poisson random variable will increase as the rate λ increases.

Example:

Some examples of Poisson distributions are illustrated below. Notice that at the value $\lambda = 9$, the distribution becomes almost symmetric.



5.2. Continuous Probability Distributions

Study time: 50 minutes

Aim - you should be able to

- characterize types of continuous distributions : exponential, Gamma and Weibull
- describe contexture in between continuous distributions

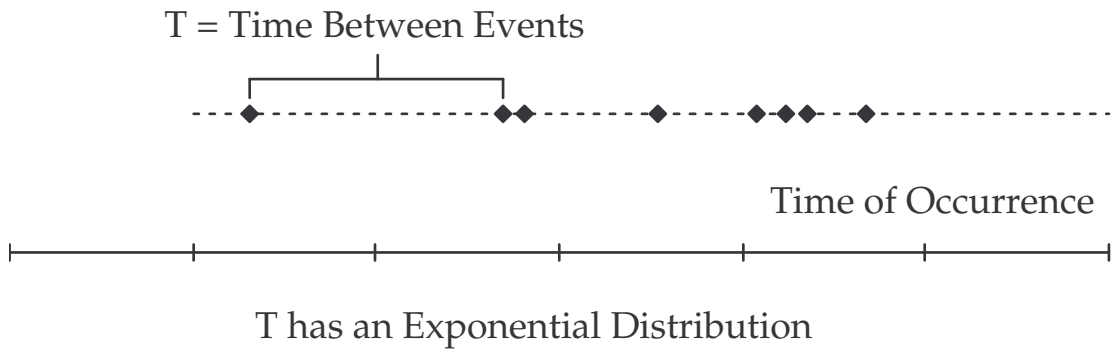
Explication

- **How does a basic description for exponential, gamma and Weibull distributions look like?**

Exponential Distribution

The exponential random variable is a second very natural variable which can be defined on the sample space generated by a Poisson process. If a continuous time process satisfies the assumptions of a Poisson process, the time between events, or equivalently because of the assumption of independence, the time until the next event will have an exponential distribution. The exponential random variable is analogous to the geometric random variable defined for a Bernoulli process.

The range of possible values for the exponential random variable is the set of non-negative numbers.



T has an Exponential Distribution

Strictly speaking, the sample space for an exponential random variable consists of intervals of varying length terminated by a single event, just as the sample space of the geometric random variable consists of a sequence of failures terminated by a success.

The probability density function and distribution function of an exponential distribution have the following simple form.

$$f(t) = \lambda e^{-\lambda t}; t \geq 0$$

$$F(t)=P(T < t) = P(N_t \geq 1) = 1 - P(N_t < 1) = 1 - e^{-\lambda t}$$

where λ is the rate at which events occur. The family of exponential random variables is identified by the single parameter, λ, the same parameter which defines the Poisson distribution.

$$T \rightarrow E(\lambda)$$

The mean of the exponential distribution is the reciprocal of the rate parameter. The result can be obtained through integrating the expected value integral by parts.

$$E(T) = \int_{t=0}^{\infty} \lambda t\, e^{-\lambda t}\, dt = \frac{1}{\lambda}$$

The variance of the exponential distribution is obtained from evaluating the following integral again through integration by parts.

$$DT = ET^2 - (ET)^2 = \ldots = \frac{1}{\lambda^2}$$

The variance equals the square of the mean and therefore the mean equals the standard deviation for an exponential distribution.

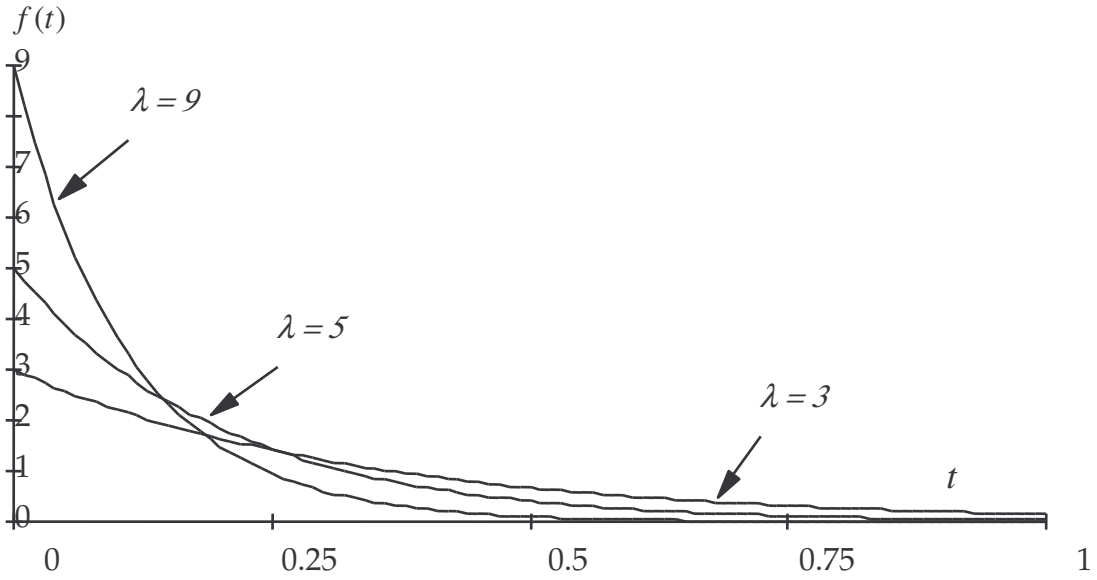
The hazard function is given by:

$$h(t) = \frac{f(t)}{1 - F(t)} \quad \text{if } F(t) < 1$$

$h(t) = \lambda = const. \Rightarrow$ the "no memory" property of the exponential distribution

Example:

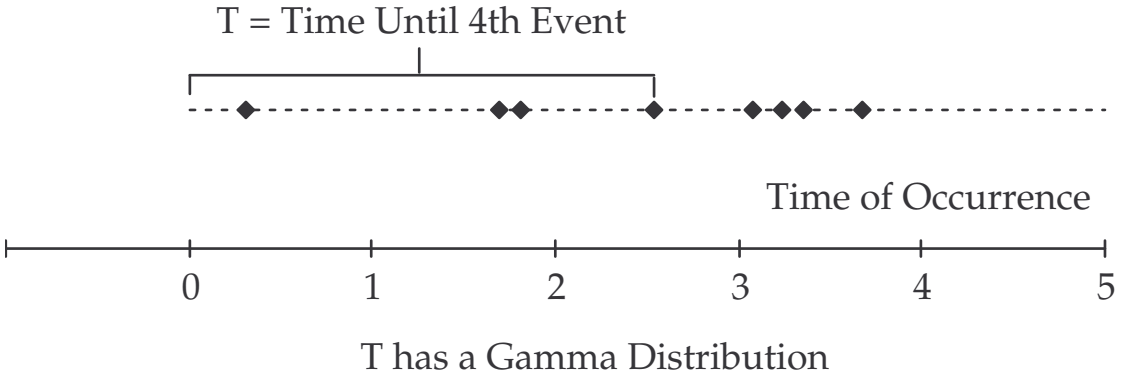
The following graph illustrates some examples of the probability density functions of exponential random variables. Notice that the shape of the exponential density is similar to the shape of the geometric probability distribution. The exponential distribution of time to next event is the continuous time equivalent of the geometric distribution which is the number of trials to next event where event may be considered a "successful" trial.



Gamma distribution

$T \rightarrow Ga(k, \lambda)$

The sample space generated by the Poisson process gives rise to a second random variable closely associated with the exponential random variable. The total time until some specified number of events, say k, occur is called a Gamma random variable and arises as a sum of k identical independent exponential random variables. If the exponential distribution is the continuous time equivalent of the geometric, then it follows that the Gamma distribution is the continuous time equivalent of the negative binomial.



The sample of a gamma random variable arising as the sum of 4 independent exponential random variables, that is as the time until the fourth event in a Poisson process will consist of intervals of varying length, all having three events and terminated by a fourth event.

The gamma distribution function for any integer value of k can be derived by the following argument. Since the gamma arises as the sum of k independent, identically distributed exponential random variables, the distribution function of the gamma is the probability that the sum of k exponentials is less than or equal to some value t. This implies that there have been at least k occurrences of a Poisson process within time t, the probability of which is

given by the cumulative distribution of a Poisson random variable with rate parameter λt, where λ is the rate of the underlying Poisson process.

$$T_k = X_1 + X_2 + X_3 + \ldots + X_k$$

$$X_i \rightarrow E(\lambda)$$

$$F(t) = P(T_k < t) = P\left(\sum_{i=1}^{k} X_i < t\right) = P(N_t \geq k) = 1 - P(N_t < k) = 1 - \sum_{j=0}^{k-1} e^{-\lambda t} \cdot \frac{(\lambda t)^j}{j!} = 1 - e^{-\lambda t}\left[\sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}\right]$$

and the probability density function is

$$f(t) = \lambda e^{-\lambda t}\left[\sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}\right] - e^{-\lambda t}\left[\sum_{j=1}^{k-1} \frac{\lambda(\lambda t)^{j-1}}{(j-1)!}\right]$$

$$= \lambda^k e^{-\lambda t}\left[\frac{t^{k-1}}{(k-1)!}\right]; \quad t > 0$$

Since the gamma random variable is the sum of *k* identical independent exponential random variables, the mean and variance will be k times the mean and variance of an exponential random variable. This same argument was used to derive the mean and variance of the negative binomial from the moments of the geometric distribution.

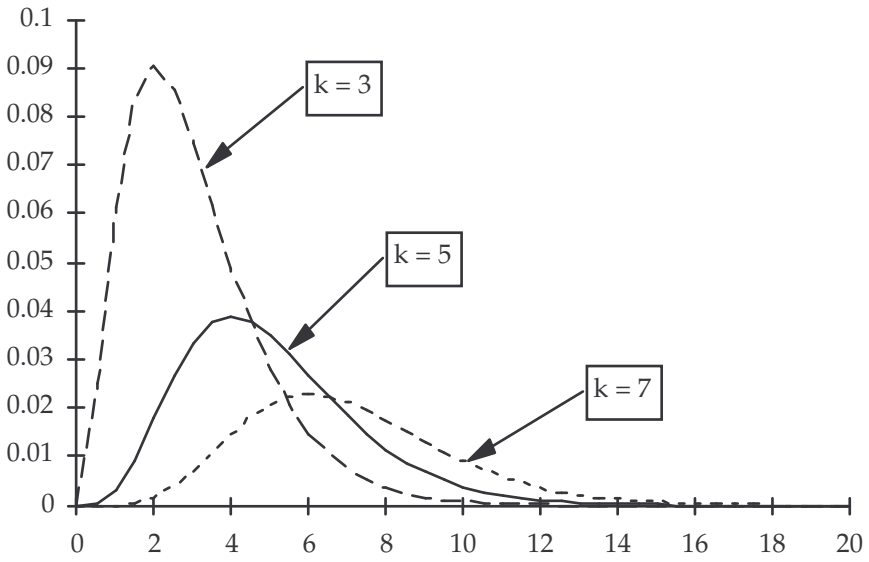
$$ET_k = EX_1 + EX_2 + \ldots + EX_k = \frac{1}{\lambda} + \ldots + \frac{1}{\lambda} = \frac{k}{\lambda}$$

$$DT_k = \ldots = \frac{k}{\lambda^2}$$

The form of the gamma distribution presented here where the parameter *k* is restricted to be a positive integer is actually a special case of the more general family of gamma distributions where *k* is a shape parameter which need only be a positive real number. The special case we have discussed is sometimes called the **Erlang distribution**.

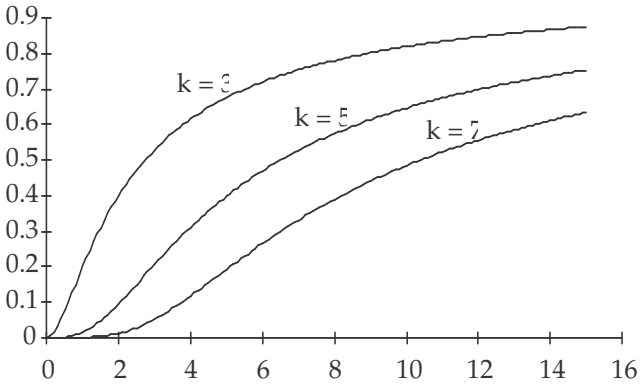
Example:

Examples of gamma probability density functions for $\lambda = 1$ are illustrated in the following chart. Notice that the gamma density has a single mode which moves away from the origin as *k* increases. Also the dispersion increases and the distribution becomes more symmetric and *k* increases.



The hazard function is given by:

$$h(x) = \frac{\lambda}{(k-1)!\sum_{j=0}^{k-1} \frac{1}{(k-1-j)!(\lambda x)^j}}$$



The hazard function of Gamma distribution, λ=1

h(x) is a sharply increasing function for k > 1 ⇒ this distribution is suitable for modeling of ageing and wear processes

Weibull distribution

The distribution function is:

$F(x) = 1 - e^{-\left(\frac{x}{\Theta}\right)^{\beta}}$, $\Theta > 0, \beta > 0, x > 0$ β ... a shape parameter, Θ ... a scale parameter.

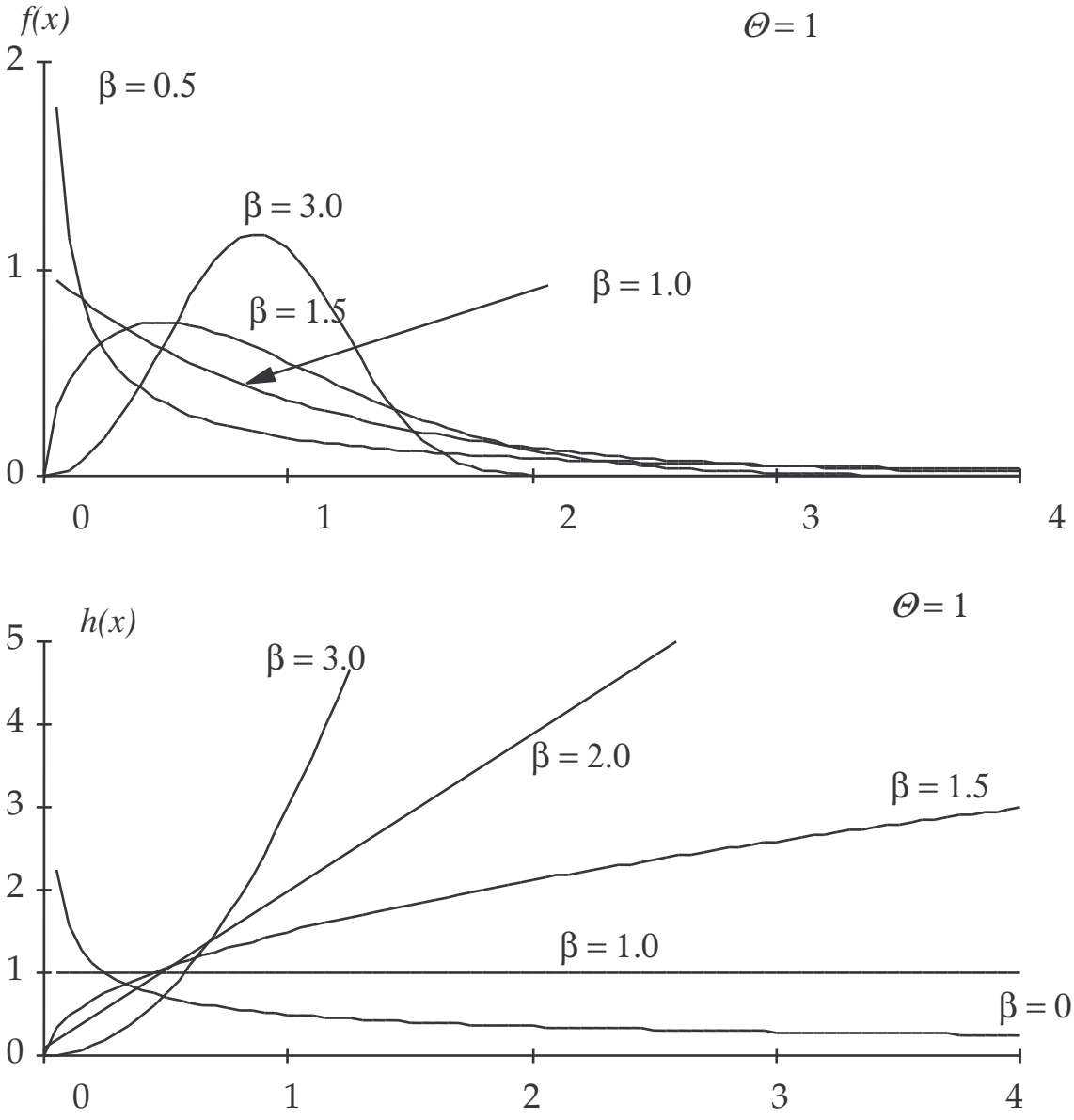
The probability density function for the Weibull is:

$$f(x) = \frac{\beta}{\Theta}\left(\frac{x}{\Theta}\right)^{\beta-1} e^{-\left(\frac{x}{\Theta}\right)^{\beta}}$$

And the hazard function for the Weibull is:

$$h(x) = \frac{\beta}{\Theta}.\left(\frac{x}{\Theta}\right)^{\beta-1}$$

Some examples of the Weibull density and the Weibull hazard function are illustrated below.



The Weibull distribution is very flexible and we use it in Reliability theory for modeling of the random variable "time to failure".

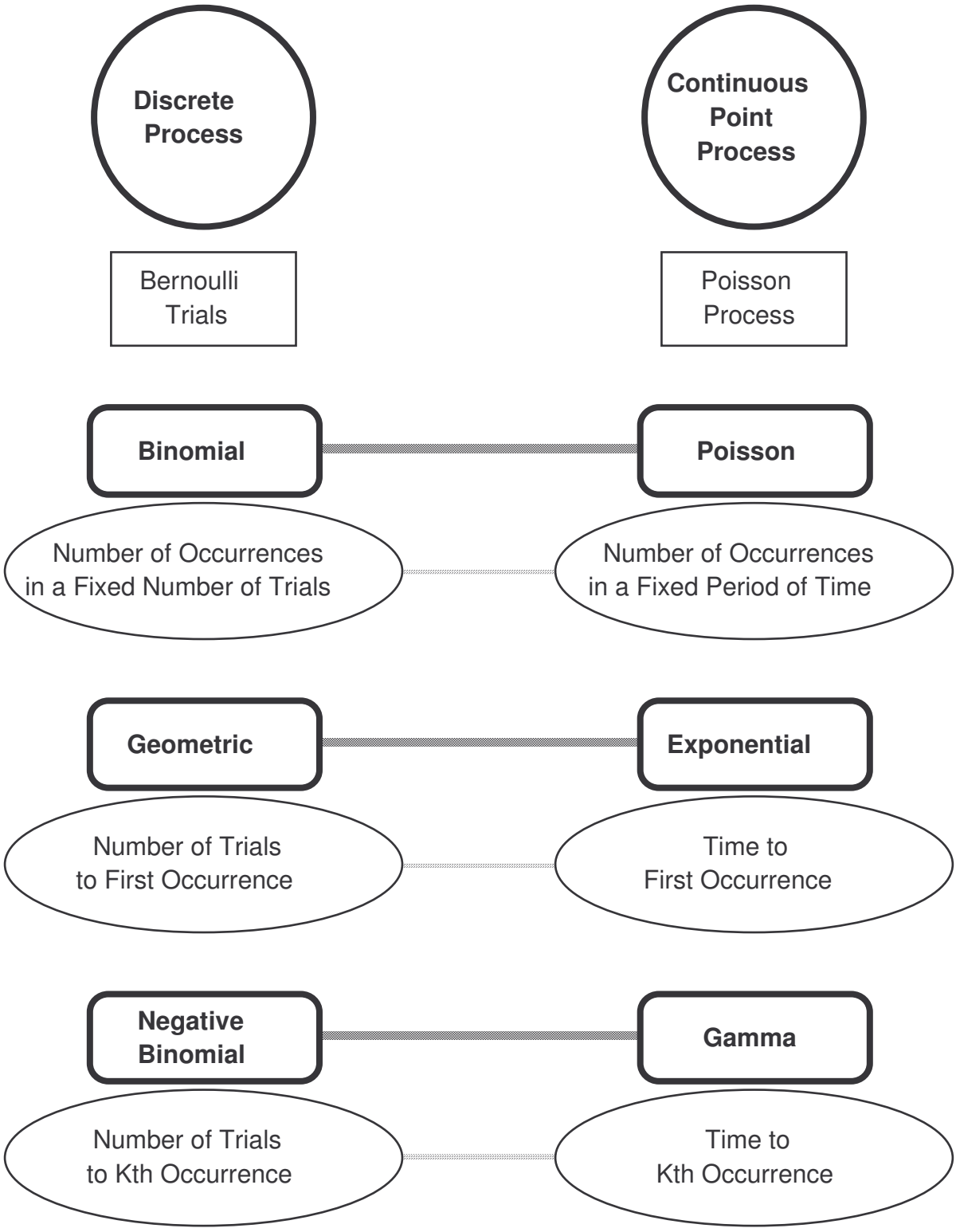
Summary of notions

A sequence of **Bernoulli trials** is defined as a sequence of random events which are mutually independent and which have only two possible outcomes (e.g. success-nonsuccess, 1-0) and the probability of event occur (a success) p is constant in any trial.
On the basis of these trials expectations we can define the following random variables: **binomial**, **geometric** and **negative binomial**.
A number of events occurrences on any deterministic interval from 0 to t can be describe (at certain expectations) by a **Poisson distribution**.
If a continuous time process satisfies the assumptions of a Poisson process, the time between events, or equivalently because of the assumption of independence, the time until the next event will have an **Exponential distribution**.
A Gamma distribution describes a time to k-th event occurrence.
A Weibull distribution is generalization of the exponential distribution and it is very flexible.

Questions

1. Which discrete and continuous distributions you know?
2. Characterize the Bernoulli trials and individual types of the discrete distributions. Determine a mean of the Binomial random variable.
3. What is Gamma distribution used for? How is it related to Exponential distribution?
4. For what β of Weibull distribution is the hazard function linear increasing?

Diagram: a contexture among distributions



Problems

Example 1*:* Suppose that a coin with probability of heads $p = 0.4$ is tossed 5 times. Let X denote the number of heads.

a) Compute the density function of X explicitly.
b) Identify the mode.
c) Find $P(X > 3)$.

{*Answer:* Let $f(k) = P(X = k) = \binom{5}{k}(0.4)^k\,(0.6)^{5-k}$ for $k = 0, 1, 2, 3, 4, 5$.

a) $f(0) = 0.0778$, $f(1) = 0.2592$, $f(2) = 0.3456$, $f(3) = 0.2304$, $f(4) = 0.0768$, $f(5) = 0.0102$.
b) mode: $k = 2$
c) $P(X > 3) = 0.9870$. }

Example 2*:* Suppose that the number of misprints N on a web page has the Poisson distribution with parameter 2.5.

a) Find the mode.
b) Find $P(N > 4)$.

{*Answer:* a) mode: $n = 2$, b)$P(N > 4) = 0.1088$}

Example 3: Message arrive at a computer at an average rate of 15 messages/second. The number of messages that arrive in 1 second is known to be a Poisson random variable.

a) Find the probability that no messages arrive in 1 second.
b) Find the probability that more than 10 messages arrive in a 1-second period.

{*Answer:* a) $3.06(10^{-7})$, b) 0.8815 }

Example 4: If there are 500 customers per eight-hour day in a check-out lane, what is the probability that there will be exactly 3 in line during any five-minute period?

{*Answer:* Poisson - 0.1288 }

6. THE NORMAL DISTRIBUTION AND THE LIMIT THEOREMS

Study time: 60 minutes

Aim - you will be able to

- characterize the normal and the standard normal distribution
- formulate and use the limit theorems
- describe special distributions

Explication

6.1. Normal Distribution

Astronomers were responsible for one of the earliest attempts to formally model the random variation inherent in the measurement process. The probability density function which was adopted at that time has been alternatively called the error function, the Gaussian curve, and today most commonly, the normal distribution. The normal distribution is the most widely used model of random variation. Its popularity is partly based on its intuitive appeal as a simple mathematical model of our instinctual notions of what constitutes random variation. However, there is also sound theoretical support for the belief that the normal distribution frequently occurs in practice.

The form of the normal density model is a simple symmetric bell-shaped curve with a single mode. This shape is achieved by the use of a negative exponential function whose argument is the square of the distance from the mode. Since squared distance makes values near zero smaller, the normal curve has a smooth rounded shape in the region of its mode. The normal density has two parameters, μ, its mode and the point about which the density is symmetric, and σ, a scale or dispersion parameter which determines the concentration of the density about the mode and the rate of decrease of the density towards the tails of the distribution. The family of normally distributed random variables is denoted by

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

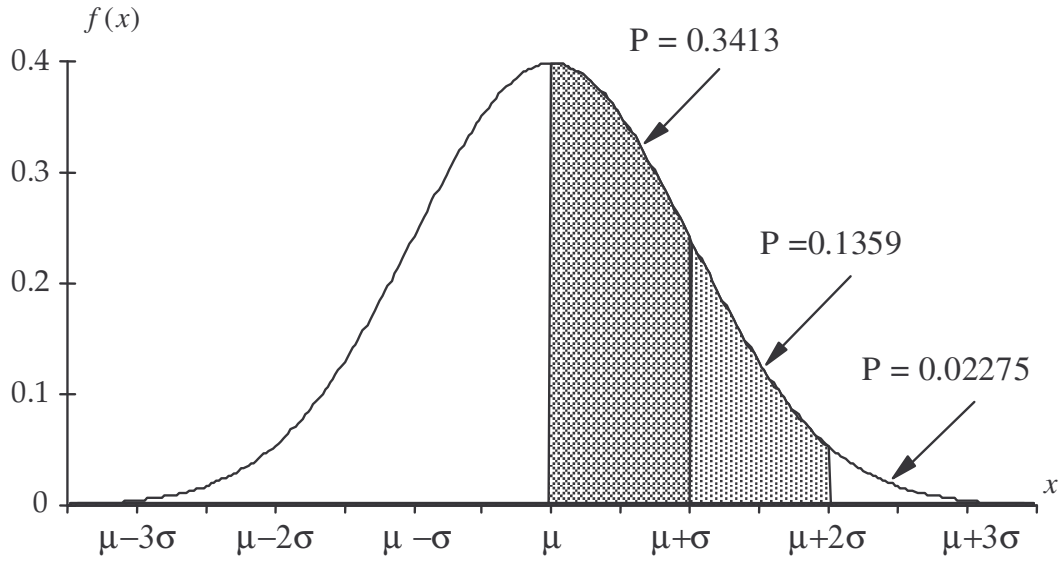
The probability density of the random variable X with the normal distribution: $X \rightarrow N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} ; -\infty < x < +\infty$$

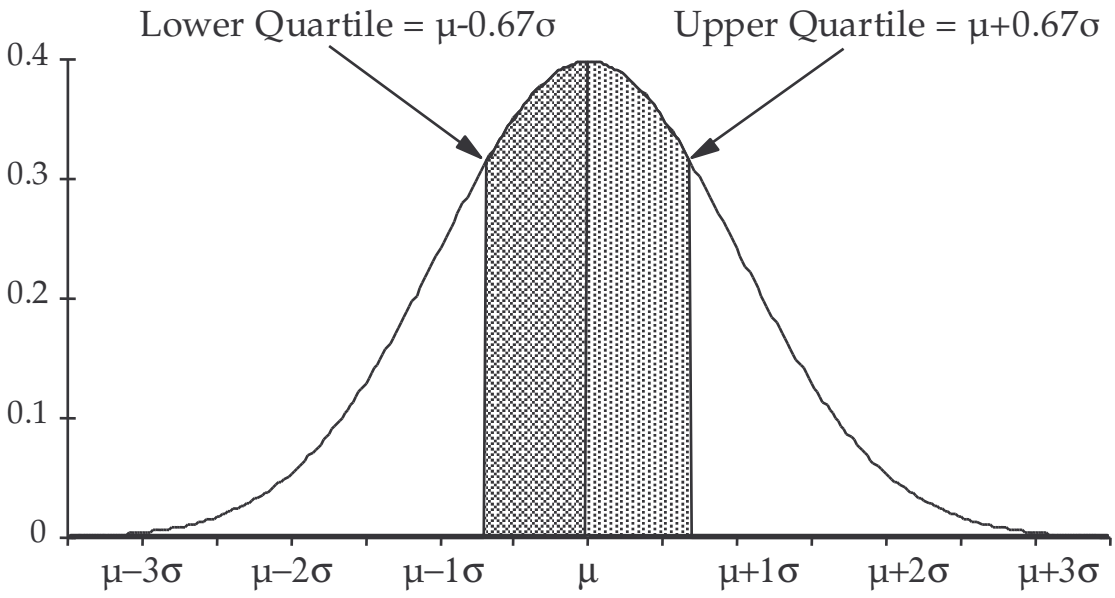
$$\text{pro - } -\infty < \mu < \infty , \qquad \sigma^2 > 0$$

This density is symmetric about μ and therefore the mean, median, and mode are all equal to μ. Also due to the symmetric bell shape of this density, the interquartile range equals the Shorth which is twice the MAD.

The following charts illustrate the distribution of probabilities for a normal random variable. The first chart shows that the probability of being between 0 and 1 standard deviation (σ) above the mean (μ) is 0.3413 or approximately one-third. Since the distribution is symmetric, the probability of being between 0 and 1 standard deviation (σ) below the mean (μ) is also approximately one third. Therefore the probability of being more than one standard deviation from the mean in either direction is again one third.



Conversely, the upper and lower quartiles are two thirds of a standard deviation above and below the mean. Thus μ+/- 0.67 σ divides the probability of the distribution into four equal parts of 25%.



The mean and variance of a normal random variable are equal to its location parameter μ, and the square of its scale parameter σ^2, respectively.

$$E(X) = \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \mu$$

$$V(X) = E[(X - E(X))^2] = \int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \sigma^2$$

The distribution function of the normal distribution:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

Standard normal distribution

A normal random variable with location parameter 0 and scale parameter 1 is called a standard normal random variable. Because of the form of the normal density, it is possible to determine probabilities for any normal random variable from the distribution function of the standard normal variable. Consequently, the standard normal random variable has been given the special symbolic destination, Z, from which the z-score derives. The standard normal distribution function is given the special symbol, Φ.

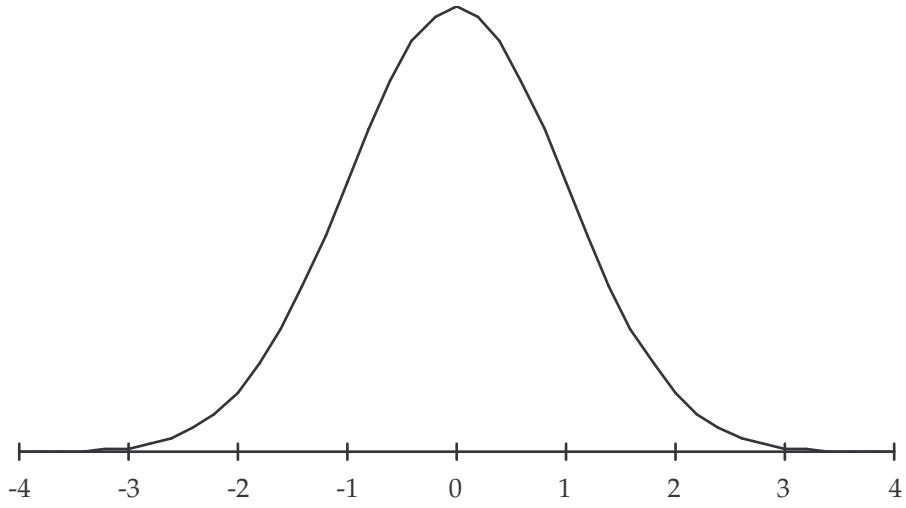
Then,

$$\Phi(z) = P(Z < z) = \int_{-\infty}^{z} \varphi(u)du = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

where φ(z) is the standard normal density

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \; ; -\infty < z < +\infty$$

Standard Normal Distribution



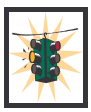
Standardization - relation between normal and standard normal distribution

Therefore if X is any normal random variable $N(\mu, \sigma^2)$ we can define a related standard normal random variable $Z = \frac{X-\mu}{\sigma}$ and it has the standard normal distribution.

$$X \dots N(\mu, \sigma^2) \;\Rightarrow\; Z = \frac{X-\mu}{\sigma} \quad , Z \dots \Phi(0, 1)$$

The distribution function of X can therefore be computed from the derived random variable Z which has a standard normal distribution:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \Phi\left(\frac{x-\mu}{\sigma}\right)$$



Solved example

X ... N(2, 25), determine P(2 <X < 8)

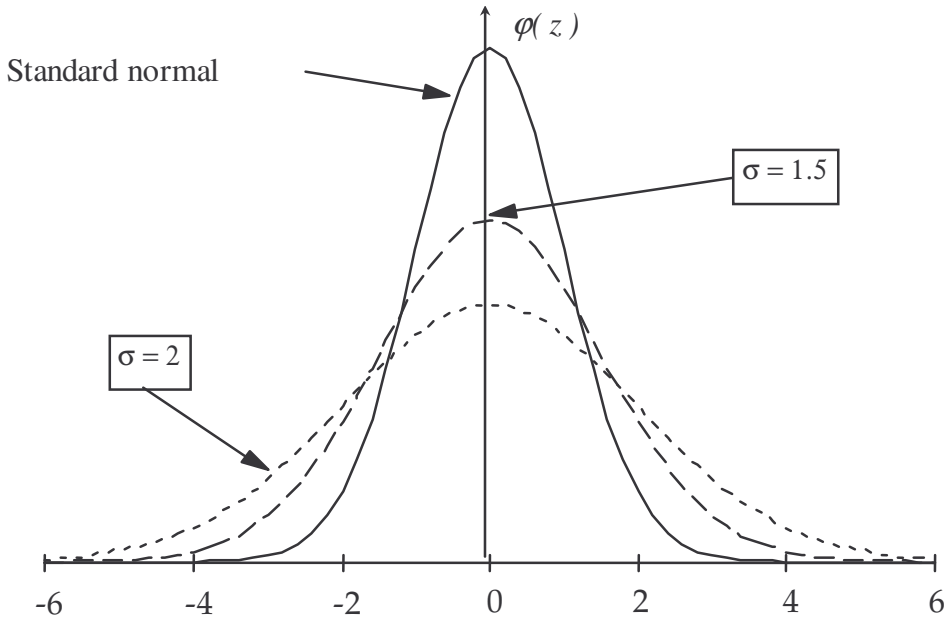
Solution:

$$P(2 < X < 8) = F(8) - F(2) = \Phi\left(\frac{8-2}{\sqrt{25}}\right) - \Phi\left(\frac{2-2}{\sqrt{25}}\right) = \Phi(1.2) - \Phi(0) = 0.885 - 0.5 = 0.385$$

In the tables or by suitable software we can find: $\Phi(1.2) = 0.885$, $\Phi(0) = 0.5$

Example:

The following chart illustrates the normal density with zero mean for selected values of σ. It is clear that the mean, median, and mode of a normal random variable are all equal, and the two parameters of the normal distribution are the embodiment of our intuitive notions about the general distributional characteristics of location and scale.



6.2. Limit Theorems

Definitions of the basic notions

Convergence in probability:

$$\lim_{n\to\infty} P(|X_n - X|<\varepsilon)=1 \quad \Rightarrow X_n \xrightarrow{p} X;$$ *i.e. a sequence of random variables* $\{X_n\}$ *converges in probability to random variable X*

Convergence in distribution:
$\{F_n(x)\}$ ... is a sequence of distribution functions corresponding to random variables $\{X_n\}$

The sequence $\{X_n\}$ *converges towards X in distribution,* if:

$$\lim_{n\to\infty} F_n(x) = F(x),$$

for every real number x at which F is continuous.

Consequence:
The sequence $\{X_n\}$ converges in distribution to distribution $N(\mu, \sigma^2)$, i.e. the random variable X_n has *asymptotical normal distribution* if $\lim_{n\to\infty} F_n(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$.

Limit theorems

Chebyshev's inequality:

X ... is an arbitrary random variable with mean *EX* and variance *DX*.
Then

$$P(\,|X - EX\,| \geq \varepsilon) \leq \frac{DX}{\varepsilon^2}, \qquad \varepsilon > 0$$

This relation results from the variance definition:

$$D(X) = \int_{-\infty}^{+\infty} (x\text{-}E(X))^2 f(x)dx = \left\{ \int_{|x-E(X)|<\varepsilon} (x\text{-}E(X))^2 f(x)dx + \int_{|x-E(X)|\geq\varepsilon} (x\text{-}E(X))^2 f(x)dx \right\}$$
$$\geq \int_{|x-E(X)|\geq\varepsilon} (x\text{-}E(X))^2 f(x)dx \geq \varepsilon^2 P(|X - E(X)| \geq \varepsilon)$$

Using of the Chebyshev's inequality (for calculation probabilities):

$$P(|X - E(X)| < k\sigma) > 1 - \frac{1}{k^2}$$

e.g. we can apply it to $X = \overline{X}$ with respect to following limit theorems:

$$P(|\overline{X} - \mu| < k\frac{\sigma}{\sqrt{n}}) > 1 - \frac{1}{k^2}$$
$$P(\frac{|\overline{X} - \mu|}{\sigma} < \frac{k}{\sqrt{n}}) > 1 - \frac{1}{k^2}$$

Law of large numbers:

$\{X_n\}$ ... is a sequence of independent random variables, each having a mean $EX_n = \mu$ and a variance $DX_n = \sigma^2$.

Define a new variable

$$\overline{X_n} = \frac{1}{n}.\sum_{j=1}^{n} X_j, \qquad n \in N$$

The sequence $\{\overline{X_n}\}$ converges in probability to μ: $\overline{X_n} \xrightarrow{p} \mu$.

Notion: This affirmation results from the Chebyshev's inequality.

Bernoulli theorem: $\{X_n\}$ ... is a sequence of the binomial independent random variables with parameters $n=1$ a $p \in (0,1)$ (so-called alternative random variable, let $X_n = 1$ in case the event

will be at one trial and $X_n = 0$ in case the event won't be; $P(X_n = 1) = p$ a $P(X_n = 0) = 1\text{-}p$).
Then we know that

$$\overline{X_n} = \frac{1}{n}.\sum_{j=1}^{n} X_j \xrightarrow{p} p$$

The expression on the left side represents a relative frequency of the event occurrence in the sequence *n* trials. That is why we can estimate a probability ingoing any occurrence by relative frequency of this event occurrence in the sequence *n* trials when we have a great number of the trials.

Central limit theorem

Lindeberg's theorem:
Let $X_1, X_2, \ldots, X_n \ldots$ be a sequence of independent random variables, $n \to \infty$.
$X_i \ldots$ have the same probability distribution, $EX_i = \mu$, $DX_i = \sigma^2$.
Then

$Y_n = \dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$ has an asymptotic normal distribution $N(0,1) \Rightarrow \lim_{n\to\infty} P(Y_n < u) = \Phi(u)$

for $-\infty < u < \infty$, it's mean that Y_n converges in distribution to distribution *N(0,1).*

For sufficiently large numbers *n* holds:

1. $X_n = \sum_{i=1}^{n} X_i \Rightarrow EX = n\mu,\ DX = n\sigma^2$, we can approximate the distribution X_n by the distribution $N(n\mu, n\sigma^2)$, i.e. X_n *has the asymptotic normal distribution*,

   $$X_n = \sum_{i=1}^{n} X_i \to N(n\mu, n\sigma^2).$$

2. Analogous for $\overline{X}$:

   $\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ *has the asymptotic normal distribution with parameters* $E\overline{X} = \mu$, $D\overline{X} = \dfrac{\sigma^2}{n}$,

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

Special case of the above theorem is Moivre-Laplace theorem:

Let $S_n \ldots$ Bi(n, p); $ES_n = np$; $DS_n = np(1\text{-}p)$
[$S_n = \sum_{i=1}^{n} X_i$, $X_i \ldots$ has the alternative distribution thus binomial Bi(1, *p*)]

then for large *n* hold that $U_n = \dfrac{S_n - np}{\sqrt{np(1-p)}} \to N(0,1)$.

Applications of the Central Limit Theorem – Normal Approximations to the binomial and Poisson distributions

Taking n observations of a Bernoulli distribution and computing the sample average $\hat{p}$ is equivalent to defining the sample proportion random variable

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{k}{n} \quad \text{... proportion of “successes“ in } n \text{ Bernoulli trials}$$

The sample proportion will have a Binomial distribution will the values re-scaled from 0 to 1. That is, if X has a binomial distribution with parameters n and p.

$$\mathrm{P}\left[\hat{p} = \frac{k}{n}\right] = \mathrm{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

Furthermore:

$$E(\hat{p}) = p; \qquad D(\hat{p}) = \frac{p(1-p)}{n}$$

Therefore by the Central Limit Theorem, we can approximate the binomial distribution by the normal distribution for large n.

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \rightarrow N(0,1)$$

$$\frac{X - np}{\sqrt{np(1-p)}} \rightarrow N(0,1)$$

Probabilities concerning ranges of value for these variables can then be calculated as

$$P(p_1 < \hat{p} < p_2) = \Phi\left(\frac{p_2 - p}{\sqrt{\frac{p(1-p)}{n}}}\right) - \Phi\left(\frac{p_1 - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$S_n = \sum_{i=1}^{n} X_i$$

$$P(k_1 < S_n < k_2) = \Phi\left(\frac{k_2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1 - np}{\sqrt{np(1-p)}}\right)$$

For smaller sample sizes a so-called continuity correction is often employed to improve the accuracy of the approximation. Thus we would compute the preceding probability as

$$P(k_1 < S_n < k_2) = \Phi\left(\frac{k_2 + 0{,}5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1 - 0{,}5 - np}{\sqrt{np(1-p)}}\right)$$

The Central Limit Theorem applies broadly to most distributions. In particular, the normal distribution can be used to approximate the Poisson distribution when the interval of observation, *t*, and hence the expected number of events, λt, is large. Let

$$\hat{\lambda} = \frac{X}{t} = \frac{count\ \ per}{unit\ \ time} = \frac{rate\ \ of}{occurrence}$$

We know that the mean and variance of the number of events during an interval *t* is λt, and therefore the mean and variance of the rate at which events occur is

$$E(\frac{X}{t}) = \lambda; \qquad D(\frac{X}{t}) = \frac{\lambda}{t}$$

Probabilities concerning Poisson counts or rates can then be calculated as

$$P(k_1 < X < k_2) = \Phi\left(\frac{k_2 - \lambda t}{\sqrt{\lambda t}}\right) - \Phi\left(\frac{k_1 - \lambda t}{\sqrt{\lambda t}}\right)$$

$$P(g_1 < \frac{X}{t} < g_2) = \Phi\left(\frac{g_2 - \lambda}{\sqrt{\frac{\lambda}{t}}}\right) - \Phi\left(\frac{g_1 - \lambda}{\sqrt{\frac{\lambda}{t}}}\right)$$

where,

$$g_1 = \frac{k_1}{t}; \qquad g_2 = \frac{k_2}{t}$$

Applying the continuity correction, we would calculate the probability as,

$$P(k_1 < X < k_2) = \Phi\left(\frac{k_2 + 0{,}5 - \lambda t}{\sqrt{\lambda t}}\right) - \Phi\left(\frac{k_1 - 0{,}5 - \lambda t}{\sqrt{\lambda t}}\right)$$

6.3. Special sampling distribution

Chi-square distribution

The chi-squared random variable arises as the sum of squared standard normal random variables. The distribution has a single parameter *n*, the number of squared normal random

variables in the sum. This parameter is called the degrees of freedom of the chi-squared distribution.

$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2$$

A χ-squared random variable with one degree of freedom χ_1^2 is simply a squared standard normal random variable. The distribution function of this random variable is

$$\begin{aligned} F_{\chi_1^2}(y) &= P(Z^2 < y) = P(-\sqrt{y} < Z < \sqrt{y}) \\ &= \Phi\left(\sqrt{y}\right) - \Phi\left(-\sqrt{y}\right) \qquad y > 0 \end{aligned}$$

The probability density function of a χ_1^2 random variable can be found from the derivative of its distribution function.

$$f_{\chi_1^2}(y) = \frac{\partial F_{\chi_1^2}(y)}{\partial y} = \left(\frac{\partial \Phi(\sqrt{y})}{\partial \sqrt{y}} + \frac{\partial \Phi(-\sqrt{y})}{\partial \sqrt{y}} \right) \frac{\partial \sqrt{y}}{\partial y}$$

$$= \frac{e^{-\frac{y}{2}}}{\sqrt{2\pi y}} \qquad y > 0$$

The general density for a χ_n^2 random variable is

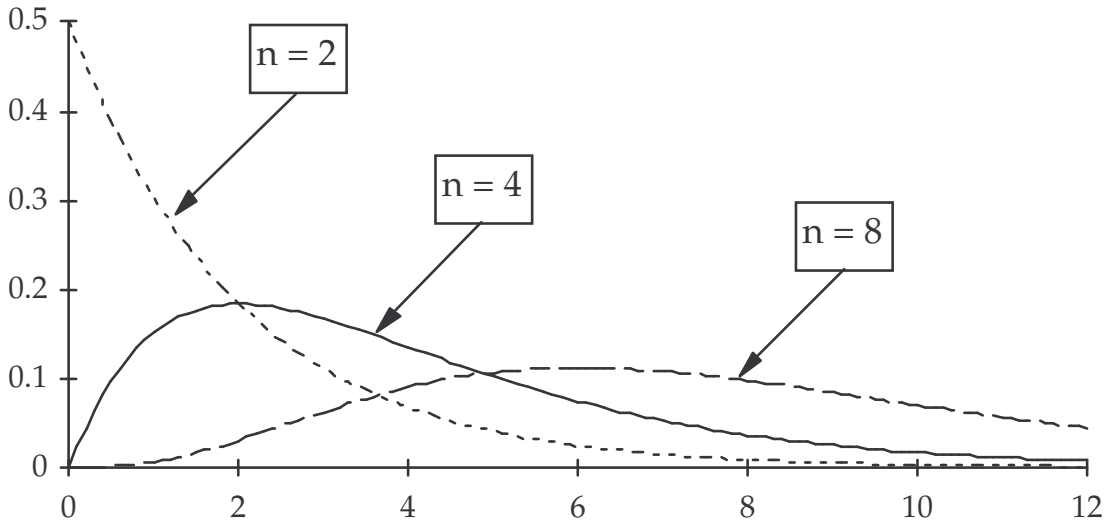
$$f_\chi(x) = \frac{x^{\left(\frac{n}{2}-1\right)} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$$

where $\Gamma(t)$ is a gamma function.

The mean and variance of the chi-squared distribution are

$$\begin{aligned} E(\chi_n^2) &= n \\ D(\chi_n^2) &= 2n \end{aligned}$$

The density function for various values of the parameter *n*:



The χ^2_n arises as the sampling distribution of the sample variance

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

If the X_i are a sample from a normal population with mean μ and standard deviation σ, then the sample variance has distribution

$$s^2 \rightarrow \chi^2_{n-1}\frac{\sigma^2}{n-1}$$

To see this, consider the sum of squared standardized observations from a normal population with mean μ and standard deviation σ.

$$\sum_{i=1}^{n}\left(\frac{x-\mu}{\sigma}\right)^2 \rightarrow \chi^2_n$$

This expression clearly has a χ^2_n distribution. Now re-express the numerator by added and subtracting the sample mean from the squared terms.

$$\frac{\sum_{i=1}^{n}(x-\bar{x}+\bar{x}-\mu)^2}{\sigma^2} \rightarrow \chi^2_n$$

Expanding and simplifying, we obtain

$$\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{\sigma^2}+\frac{(\bar{x}-\mu)^2}{\sigma^2/n}\rightarrow \chi_n^2$$

The second term is simply the squared standardized sample mean and therefore has a χ_1^2 distribution. Since the χ_n^2 is the sum of *n* squared normals, the first term must be the remaining *n-1* squared normals. Therefore,

$$\frac{(n-1)\,s^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{\sigma^2}\rightarrow \chi_{n-1}^2$$

Note that this argument is only heuristic and not a formal proof. The independence of the sample mean from the deviations about the sample mean has not been established.

This fact is important at the statistical hypothesis testing.
1. We use this distribution for verification of the random variables independence.
2. We can use chi-square distribution when we test that the random variables follow from certain distribution. This test is known as "Goodness-of-Fit Test".

Student's distribution (*t* distribution)

The Student's t distribution is the sampling distribution of the standardized sample mean when the sample variance is used to estimate the true population variance. The origin of the distribution's name, Student's t has an interesting history. An Irish statistician, W. S. Gosset first published this distribution anonymously under the pseudonym "Student" because his employer, Guiness Breweries of Dublin, prohibited its employees from publishing under their own names for fear that its competitors would discover the secret of their excellent beer. In his original paper, Gosset used the designation "t" for his statistic. Hence the name is Student's.

The Student's *t* distribution with *n* degrees of freedom is the ratio of a standard normal random variable over the square root of a chi-squared random variable divided by its degrees of freedom. The *t* distribution has a single parameter, *n* the degrees of freedom of the chi-squared random variable in the denominator.

$$t_n=\frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$$

The probability density function of this random variable is

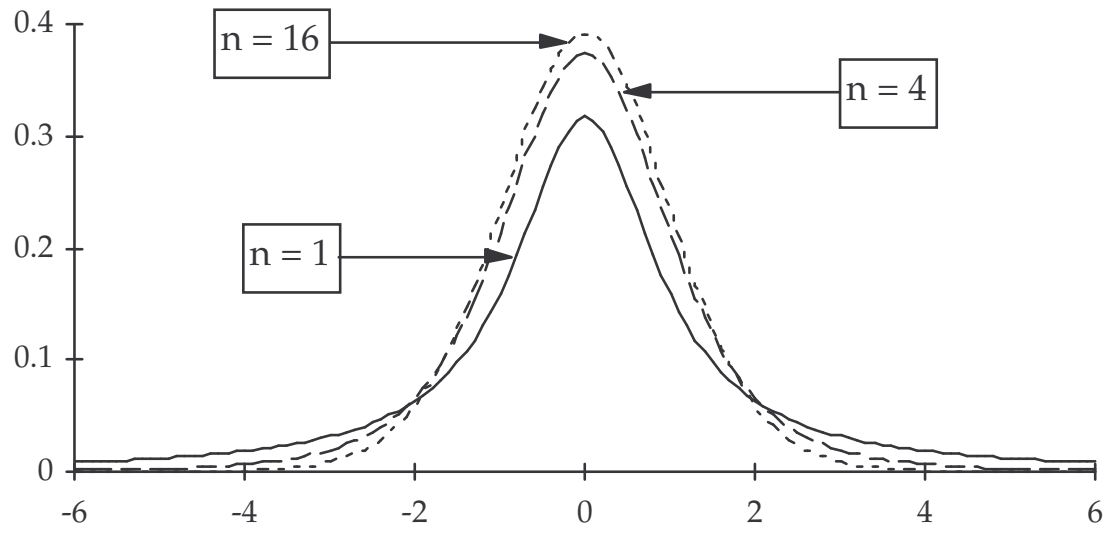
$$f(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\;\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

The mean and variance of the t-distribution are

$$E(t_n) = 0$$

$$D(t_n) = \frac{n}{n-2}$$

The following figure shows the density function for different values of number of the degrees of freedom:



If random variables $X_1, X_2, ..., X_n$ have the normal distribution $N(\mu, \sigma^2)$ and they are **independent** then we can show that

$$\overline{X} \rightarrow N(\mu, \sigma^2/n)$$

and

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

then

$$\frac{\overline{X} - \mu}{S}\sqrt{n} = \frac{\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\frac{\frac{S^2}{\sigma^2}(n-1)}{n-1}}} \rightarrow t_{n-1}$$

Student's t-distribution has a wide exercise.

Fisher-Snedecor‘s distribution - F distribution

Snedecor's F distribution arises as the ratio of two chi-squared distributions divided by their respective degrees of freedom. The F distribution has two parameters, n the degrees of freedom of the chi-squared random variable in the numerator and m the degrees of freedom of the chi-squared random variable in the denominator.

$$F_{n,m} = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_m^2}{m}}$$

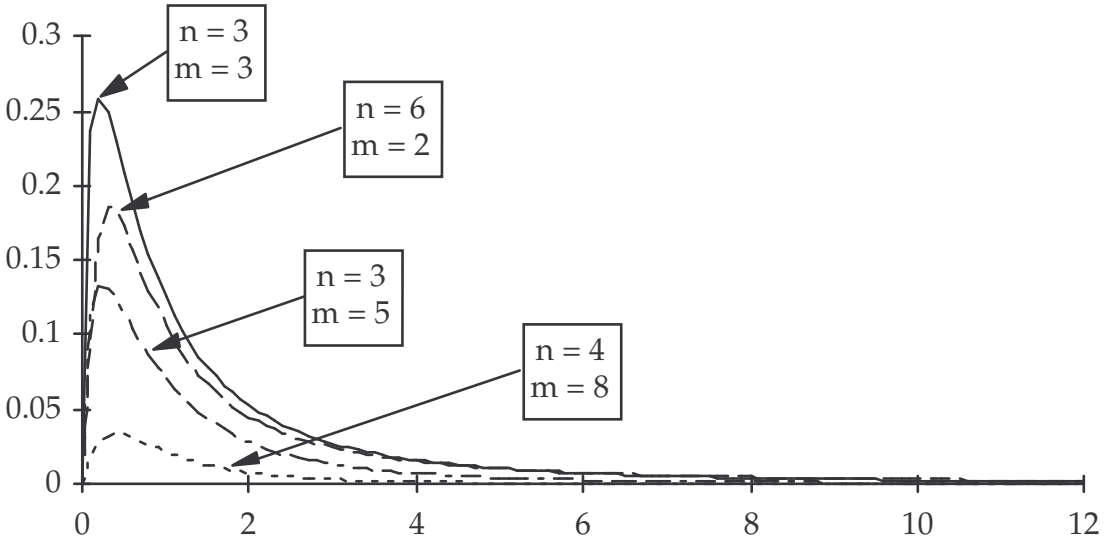
The probability density function of this random variable is

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right) n^{\frac{n}{2}} m^{\frac{m}{2}} x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)(m+nx)^{\frac{n+m}{2}}}$$

The mean and variance of the t distribution are

$$E(F_{n,m}) = \frac{m}{m-2} \qquad D(F_{n,m}) = \frac{2m^2\left(1+\frac{m-2}{n}\right)}{(m-2)^2(m-4)}$$

The following figure shows the density function for different values m a n:



Clearly such a distribution would arise as the sampling distribution of the ratio of the sample variances from two independent populations with the same standard deviation σ. The degrees of freedom represent one less than the samples sizes of the numerator and denominator sample variances respectively.

$\overline{X_1}; \overline{X_2}$

$X_{1j} \rightarrow N(\mu_1, \sigma); \quad \mathrm{j} = 1, \mathrm{n}_1$

$X_{2j} \rightarrow N(\mu_2, \sigma); \quad \mathrm{j} = 1, \mathrm{n}_2$

$$S_i^2 = \frac{\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2}{n_i - 1}; \; i = 1, 2$$

$$\frac{S_1^2}{S_2^2} \rightarrow F_{n_1, n_2}$$

We use this distribution for evaluation of statistical analysis results.

Summary of notions

One of the most important continuous distributions is a **normal distribution**. It is distribution with two parameters, when the first parameter is a mean and the second one is a variance. We get standard normal distribution for special choice of parameters (the mean is equal 0 and the variance is equal 1).
Chebyshev's inequality puts an upper bound on the probability that an observation should be far from its mean.
Chí-square distribution is a distribution derived from sum of squared standard normal random variables.
Central limit theorem describes asymptotic statistic behavior of mean. We can use it for substitution of binomial (Poisson) distribution by normal distribution.
Student's t-distribution with *n* degrees of freedom is the ratio of a standard normal random variable over the square root of a chi-squared random variable divided by its degrees of freedom.
F distribution is the ratio of two chi-squared distributions divided by their respective degrees of freedom.

Questions

1. Define relation between normal and standard normal distributions.
2. What is Chebyshev's inequality?
3. Explain law of large numbers.
4. Describe chi-square distribution.

Problems

Example 1: If the mean (μ) height of a group of students is equal to 170cm with a standard deviation (σ) of 10 cm, calculate the probability that a student is between 160cm and 180cm.

{*Answer:* 0,6828}

Example 2: Let X = "height of a randomly chosen male", and suppose that X is normally distributed with μ = 176cm and σ^2 = 25 cm^2,i.e. X is N(176,25):

(i) calculate the probability that the height of a randomly chosen male is less than or equal to 182
(ii) calculate the probability that the height of a randomly chosen male is less than or equal to 170
(iii) calculate the probability that the height of a randomly chosen male is not greater than 176
(iv) calculate the probability that the height of a randomly chosen male is between 170 and 182 cm
(v) calculate the probability that the height of a randomly chosen male is not less than 160

{*Answer:* (i) 0.885, (ii) 0.115, (iii) 0.5, (iv) 0.7698. (v) 0.9993}

Example 3: We take the heights of 9 males and we assume that the heights are i.i.d. N(176,25) as before. What is the probability that the sample mean height is between 174 and 178 cm?

{*Answer:* 0.7698}

Example 4: Premature babies are those born more than 3 weeks early. A local newspaper reports that 10% of the live births in a country are premature. Suppose that 250 live births are randomly selected and the number Y of "preemies" is determined.

(i) What is the probability that X lies in between 15 and 30 (both included)?
(ii) Find the proportion of the event fewer than 20 births are premature?

{*Answer:* (i) 0.86, (ii) 0.12}

7. INTRODUCTION TO STATISTICAL INFERENCE

Study time: 50 minutes

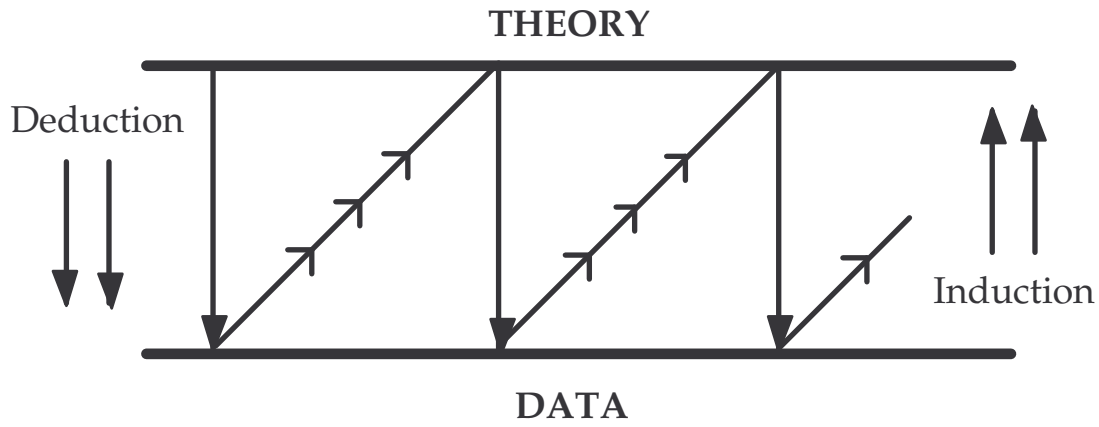
Aim - you will be able to

- understand the random sampling term
- use the sampling distribution and their properties

Explication

7.1. Introduction – The Scientific Method

Science is a process of systematic learning which proceeds by alternating between inductive and deductive methods of investigation.



The methods of induction and deduction are the connections between data and theory of a science. The deductive method proceeds in a logically consistent fashion to project what data should result from a particular theory. Induction is an informal process which tries to postulate some theory to reasonably explain the observed data.

Statistics to be a complete science must embody both inductive and deductive methods. The first topic of the course, exploratory data analysis, was an attempt to understand observed distributions of data and was therefore an inductive method. Without some theory of randomness however, the ability of EDA methods to induce precise explanations was limited. Therefore, we introduced the theory of probability and discussed its

applications to various hypothetical sample spaces. Probability theory is the basis of deductive methods of statistics.

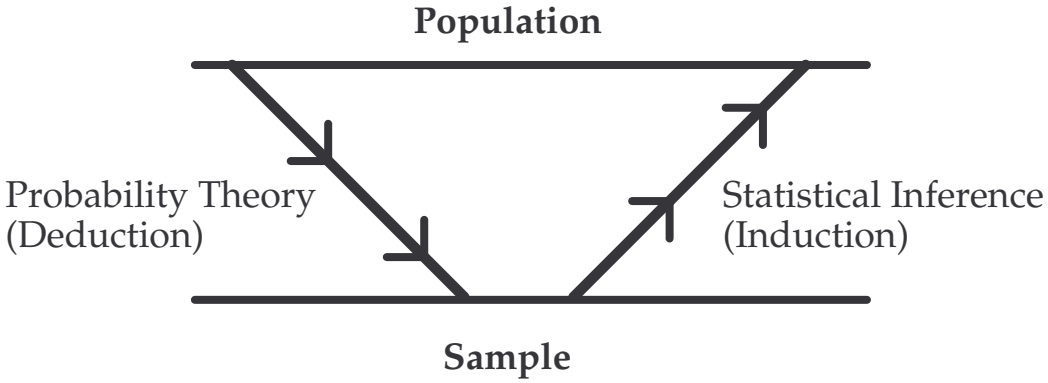
Probability theory proceeds by assuming a hypothetical sample space or population on which a probability measure is defined. Probability distributions of random variables defined on this sample space are then derived mathematically. The probability of any sample observation from the hypothetical population can then be determined.

Deductive Theory of Probability

Known Probability Distribution of Population

Sampling Distributions of Sample Statistics

If probability theory is the deductive method of statistics, then by implication, theory in statistical science must be represented by some well-defined population with a known probability distribution and data by the sample drawn from that population. Statistical inference then becomes the inductive methods for using sample data to make inferences about the probability distribution of the population from which the sample was drawn.

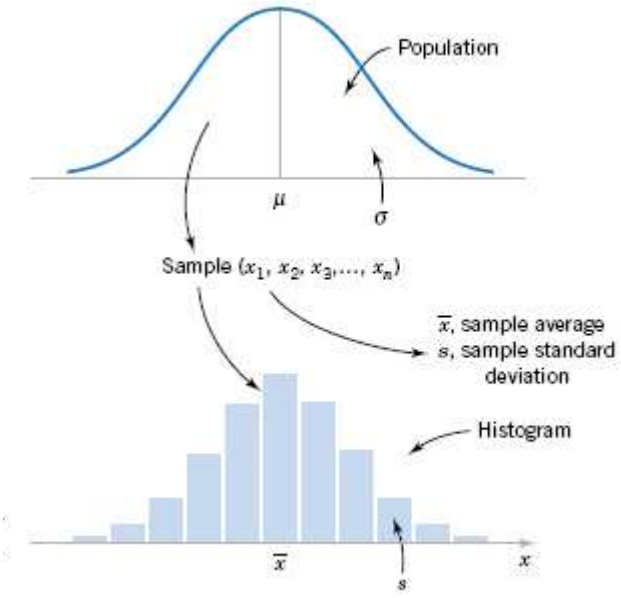


Statistical inference then is the inverse of probability theory. It is the process of making statements about an unknown population on the basis of a known sample from that population.

Viewing statistical science in this light may led us to ask why we have studying probability theory at all. Surely probability theory is answering the wrong question. The hypothetical problem of sampling from a known population never occurs in practice. In statistics, the population is always unknown and we not generate sample data in order to obtain information about the unknown population. In practice, is not statistical inference the only problem of statistical science?

7.2. Random Sampling

In most statistics problems, we work with a sample of observations selected from the population that we are interested in studying. Following Figure illustrates the relationship between the population and the sample.



We have informally discussed these concepts before; however, we now give the formal definitions of some of these terms.

Definition:
A **population** consists of the totality of the observations with which we are concerned.

In any particular problem, the population may be small, large but finite, or infinite. The number of observations in the population is called the **size** of the population. For example, the number of undefiled bottles produced on one day by a soft-drink company is a population of finite size. The observations obtained by measuring the carbon monoxide level every day is a

population of infinite size. We often use a **probability distribution** as a **model** for a population.

For example, a structural engineer might consider the population of tensile strengths of a chassis structural element to be normally distributed with mean _ and variance. We could refer to this as a **normal population** or a normally distributed population. In most situations, it is impossible or impractical to observe the entire population. For example, we could not test the tensile strength of all the chassis structural elements because it would be too time consuming and expensive. Furthermore, some (perhaps many) of these structural elements do not yet exist at the time a decision is to be made, so to a large extent, we must view the population as **conceptual.** Therefore, we depend on a subset of observations from the population to help make decisions about the population.

Definition:
A **sample** is a subset of observations selected from a population.

For statistical methods to be valid, the sample must be representative of the population. It is often tempting to select the observations that are most convenient as the sample or to exercise judgment in sample selection. These procedures can frequently introduce **bias** into the sample, and as a result the parameter of interest will be consistently underestimated (or overestimated) by such a sample. Furthermore, the behavior of a judgment sample cannot be statistically described. To avoid these difficulties, it is desirable to select a **random sample** as the result of some chance mechanism. Consequently, the selection of a sample is a random experiment and each observation in the sample is the observed value of a random variable. The observations in the population determine the probability distribution of the random variable. To define a random sample, let X be a random variable that represents the result of one selection of an observation from the population. Let $f(x)$ denotes the probability density function of X. Suppose that each observation in the sample is obtained independently, under unchanging conditions. That is, the observations for the sample are obtained by observing X independently under unchanging conditions, say, n times. Let X_i denote the random variable that represents the ith replicate. Then, $X_1, X_2 \ldots X_n$ is a random sample and the numerical values obtained are denoted as $x_1, x_2, \ldots, x_n$. The random variables in a random sample are independent with the same probability distribution $f(x)$ because of the identical conditions under which each observation is obtained. That is, the marginal probability density function of $X_1, X_2 \ldots X_n$ is

$$f(x_1),\ f(x_2),\ldots f(x_n)$$

respectively, and by independence the joint probability density function of the random sample is

$$f_{X1X2,\ldots Xn}(x_1, x_2, \ldots, x_n) = f(x_1)\, f(x_2)\ldots f(x_n).$$

Definition:
The random variables $X_1, X_2 \ldots X_n$ are a random sample of size n if (a) the X_i's are independent random variables, and (b) every X_i has the same probability distribution.

To illustrate this definition, suppose that we are investigating the effective service life of an electronic component used in a cardiac pacemaker and that component life is normally distributed. Then we would expect each of the observations on component life in a random

sample of n components to be independent random variables with exactly the same normal distribution. After the data are collected, the numerical values of the observed lifetimes are denoted as $x_1, x_2, \ldots, x_n$.

The primary purpose in taking a random sample is to obtain information about the unknown population parameters. Suppose, for example, that we wish to reach a conclusion about the proportion of people in the United States who prefer a particular brand of soft drink. Let p represent the unknown value of this proportion. It is impractical to question every individual in the population to determine the true value of p. In order to make an inference regarding the true proportion p, a more reasonable procedure would be to select a random sample (of an appropriate size) and use the observed proportion $\hat{p}$ of people in this sample favoring the brand of soft drink.

The sample proportion, $\hat{p}$ is computed by dividing the number of individuals in the sample who prefer the brand of soft drink by the total sample size n. Thus, $\hat{p}$ is a function of the observed values in the random sample. Since many random samples are possible from a population, the value of $\hat{p}$ will vary from sample to sample. That is, $\hat{p}$ is a random variable. Such a random variable is called a **statistic.**

Definition:
A **statistic** is any function of the observations in a random sample.

We have encountered statistics before. For example, if $X_1, X_2 \ldots X_n$ is a random sample of size n, the **sample mean** $\overline{X}$ the **sample variance** S^2, and the **sample standard deviation** S are statistics.

Although numerical summary statistics are very useful, **graphical displays** of sample data are a very powerful and extremely useful way to visually examine the data. In first lecture we presented a few of the techniques that are most relevant to engineering applications of probability and statistics.

7.3. Sampling Distribution

Let's assume that given random sample comes from normal distribution:

$\underline{X}=(X_1, \ldots, X_n)'$, $X_i \to N(\mu, \sigma^2)$

1. $\overline{X_n} = \dfrac{\sum_{i=1}^{n} X_i}{n} \to N\left(\mu, \dfrac{\sigma^2}{n}\right)$ … comes from central limit theorem for large number n

2. $Z_n = \dfrac{\overline{X_n} - \mu}{\sigma} \cdot \sqrt{n} \to N(0,1)$ … comes from a transformation of previous distribution

3. $\dfrac{S_n^2}{\sigma^2} \cdot (n-1) \to \chi^2(n-1)$ … was explained in χ^2 discussion

where $S_n^2 = \dfrac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2$; $\dfrac{S_n^2}{\sigma^2} \cdot (n-1) = \sum_{i=1}^{n} \dfrac{(X_i - \overline{X})^2}{\sigma^2} = \sum_{i=1}^{n} \left(\dfrac{X_i - \overline{X}}{\sigma} \right)^2$

4. $\frac{\overline{X}_n - \mu}{s} \cdot \sqrt{n} \to t_{n-1}$ … was derived in the discussion about using of Student's distribution since:

$$\frac{\frac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n}}{\sqrt{\frac{\frac{S^2}{\sigma^2} \cdot (n-1)}{n-1}}} = \frac{\overline{X}_n - \mu}{S} \cdot \sqrt{n}$$

Now assume two samples from the normal distribution
$\underline{X}=(X_1, \ldots, X_n)'$, $X_i \to N(\mu_1, \sigma_1^2)$, $\underline{Y}=(Y_1, \ldots, Y_m)'$, $Y_j \to N(\mu_2, \sigma_2^2)$. Then it holds:

5. $$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \to N(0,1)$$

$$\overline{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right)$$

$$\overline{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

6. $$\frac{\frac{\frac{S_x^2}{\sigma_1^2} \cdot (n-1)}{n-1}}{\frac{\frac{S_y^2}{\sigma_2^2} \cdot (m-1)}{m-1}} = \frac{\frac{S_x^2}{\sigma_1^2}}{\frac{S_y^2}{\sigma_2^2}} \to F_{n-1,m-1}$$ … explained in F – distribution

Now assume that the variances are the same and unknown: $\sigma_1^2 = \sigma_2^2$. Then can be shown that it holds:

7. $$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_x^2(n-1) + S_y^2(m-1)}} \cdot \sqrt{\frac{n.m.(n+m-2)}{n+m}} \to t_{n+m-2}$$

Summary of notions

The **random sample** is the special random vector whose elements are independent random variables with the same probability distribution.

If the random sample comes from the normal distribution of probability we can derivate other significant statistics with known distribution from given random sample, e.g. t-statistics $\frac{\overline{X}_n - \mu}{s} \cdot \sqrt{n} \to t_{n-1}$ or two-sample t-statistics:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_x^2(n-1) + S_y^2(m-1)}} \cdot \sqrt{\frac{n.m.(n+m-2)}{n+m}} \to t_{n+m-2} .$$

These other statistics will be later used for the construction of interval estimation or for hypothesis testing.

Questions

1. What is the statistical induction?
2. Characterize the term a random sample.

8. HYPOTHESIS TESTING

Study time: 80 minutes

Aim - you will be able to

- conclude by the pure significance test
- use basic sample and two sample tests
- conclude by paired test and tests for proportions

Explication

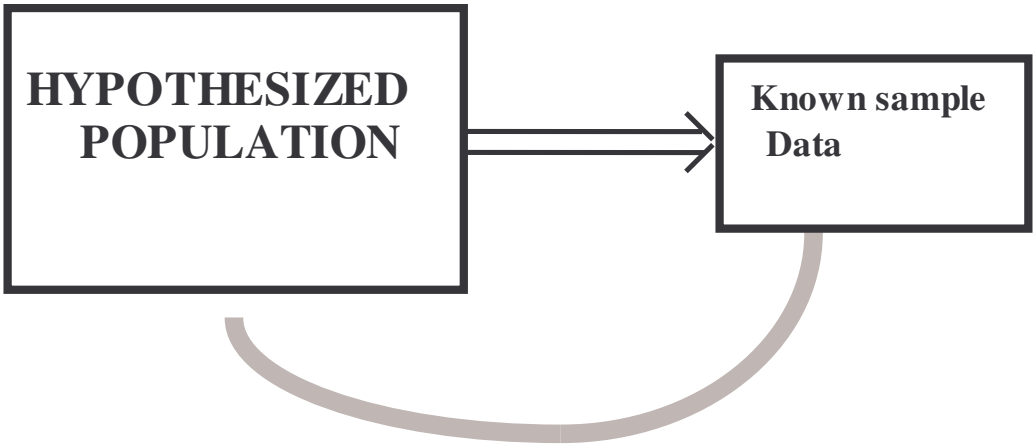
8.1. Introduction

In this chapter we will construct test and with their help we accept or reject some population hypothesis.
The most frequent case is a situation when we can describe a population by some probability distribution which depends on θ parameter. Based on the trial result we can for example want to accept or reject an opinion that θ has some concrete value θ_0. In other situation we can be interested in our hypothesis validity that given population comes from a concrete distribution. Procedures leading to similar decisions are called **significance tests**.
Statistic hypothesis - the assumptions about population whose trueness can be verified by statistic significance tests
Significance tests - procedures which decide if a verified hypothesis should be accepted or rejected based on random sample
Null hypothesis H_0 - verified hypothesis whose rejection is decided by a significance test
Alternative hypothesis H_A - hypothesis which is accepted when we reject null hypothesis

8.2. Pure Significance Tests

The pure significance test asks whether the sample result is extreme with respect to some hypothesized distribution.

If the sample data lies at an extremely high or extremely low percentile of the hypothesized distribution, then the hypothesis is in doubt.



Is data consistent with hypothesized population ?

The pure significance test consists of the following components:

1. Null Hypothesis: H_0
   – The null hypothesis expresses some belief about the nature of the population. It must be specified precisely enough to define a probability measure on the population.

2. Sample Statistic: T(X)
   – The sample statistic is a function of the sample data drawn from the population. The choice of sample statistic is determined by the characteristics of the population's probability distribution with which the null hypothesis is concerned.

3. Null Distribution: $F_0(x)$
   $F_0(x) = P(T(\underline{X}) < x \mid H_0)$
   - The null distribution is the probability distribution of the sample statistic when the null hypothesis is correct. The null hypothesis must be specified precisely enough to determine the null distribution.

4. To determine whether the observed sample statistic $t=x_{OBS}$ is extreme with respect to the null distribution, a statistic known as the p-value is computed. The p-value has 3 definitions depending on the context of the null hypothesis, but in all cases, the interpretation of the p-value is the same.

   Definition 1: $P_{VALUE} = F_0(x_{OBS})$

This definition is used when we are concerned that the distribution of the sample statistics may be less than the null distribution.

Definition 2: $P_{VALUE} = 1 - F_0(x_{OBS})$

This definition is used when we are concerned that the distribution of the sample statistics may be greater than the null distribution. Such is the case in our first example.

Definition 3: $P_{VALUE} = 2 \min [F_0 (x_{OBS}), 1 - F_0 (x_{OBS})]$

This definition is used when we are concerned that the distribution of the sample statistics may be either greater or less than the null distribution. Note that this definition is only applied when the null distribution is symmetric.

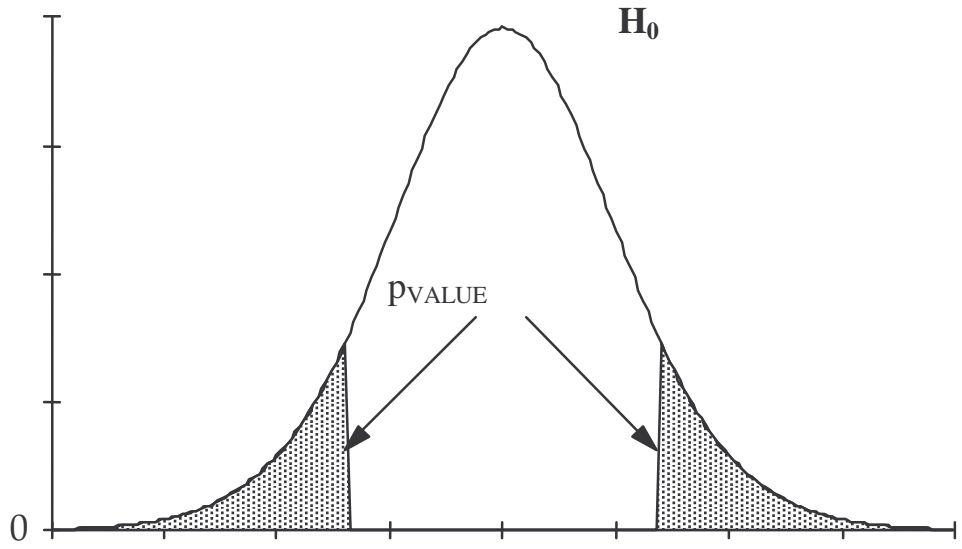


Figure: *Graphical presentation of* p_{VALUE} *for definition 3 by area below spline of density of the null distribution.*

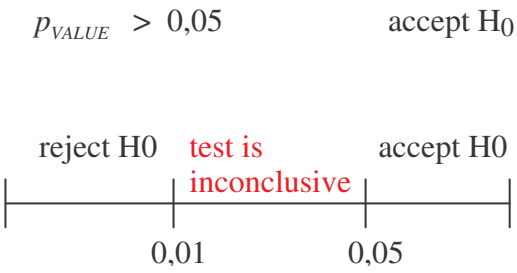
When the null hypothesis is correct, the distribution of the p-value under all three definitions is uniform. That is,

$$P(p_{VALUE}(\underline{X}) < p | H_0) = p$$

Therefore, the p-value has the same interpretation for all null hypotheses independent of the original null distribution. Clearly, smaller p-values are more extreme with respect to null distribution. Therefore, the smaller the p-value, the stronger the evidence of the sample statistic against the null hypothesis. But how small must the p-value be before the evidence is strong enough to reject the null hypothesis? Strictly speaking, this would again depend on the context in which the hypothesis is tested. However, since the weight of evidence against the null hypothesis increases continuously with decreasing p-value, it would be unreasonable to designate a single p-value cut-off point below which the null hypothesis is rejected and above which it is accepted. Rather we should expect an inconclusive region separating the accept and reject p-values.

5. Conclusion in terms of p_{VALUE}

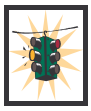
$p_{VALUE} < 0{,}01$	reject H_0
$0{,}01 < p_{VALUE} < 0{,}05$	test is inconclusive



8.3. Alternate hypothesis

From the definition of the *p-value*, it is clear that the pure significance test procedure for hypothesis testing requires not only a specific null hypothesis but also but notion of which alternative might be correct if the null hypothesis is rejected. The alternate hypothesis need not be specified as precisely as the null hypothesis. To select the appropriate definition of the *p-value*, it is only necessary to know the direction of the alternative with respect to the null. However, the alternate hypothesis will also influence the choice of sample statistic. Those values of the sample statistic which have a small *p-value* under the null hypothesis should tend to have a larger *p-value* for prospective alternatives and vice versa. (Large null *p-values* should have small alternate *p-values*).

8.4. Hypothesis Tests for mean and median



Solved example 1

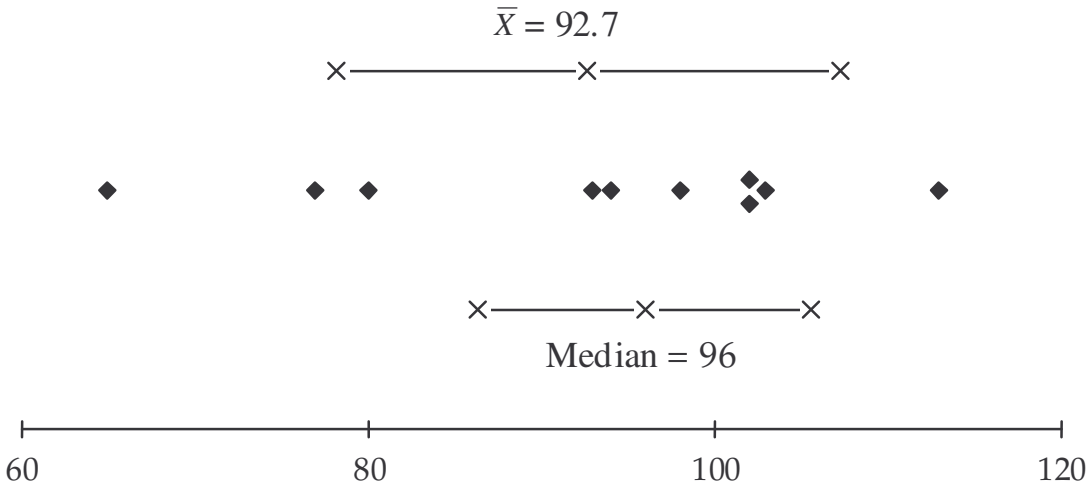
Consider the following ten IQ scores. IQ test scores are scaled to have a mean of 100 and a standard deviation of 15.

65	98	103	77	93
102	102	113	80	94

We wish to test the hypothesis that the mean is 100.

Solution:

We can illustrate this sample:



H_0: X (IQ) has N(100, 15); $\mu_0 = 100$; $\sigma = 15$

Under H_0 $\Rightarrow$ $\bar{X} \to N\left(100, \frac{15}{\sqrt{10}}\right)$

$\bar{X} = 92.7$; $s = 14.51$

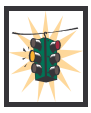
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{92.7 - 100}{15/\sqrt{10}} = -1.54$$

$p_{VALUE} = \Phi(-1.54) = 0{,}06178$

Accept H_0: $\mu_0 = 100$

Notes: a) When the sample size n is large, the Central Limit Theorem permits the use of this test when the original population is not normally distributed.

b) If σ is not known and the original population is not normally distributed, the sample standard deviation s may be substituted when the sample size is large.



Solved example 2

- we have same data as with example1

Student's test for mean of small samples

H_0: X je N(100, σ); $\mu_0 = 100$; σ is uknown

$\overline{X} = 92.7$; $s = 14.51$

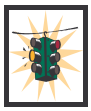
$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{92.7 - 100}{14.51/\sqrt{10}} = -1.59$$

$t \rightarrow t_{n-1} = t_9$

$p_{VALUE} = t_{n-1}(-1.59) = 0{,}073149$

Accept H_0: $\mu_0 = 100$.

Notes: a) When the sample size n becomes larger than about 30, the t distribution becomes very similar to the normal distribution.



Solved example 3

- we have same data as with example1

Sign test for median

An alternative to testing the hypothesis is that the mean equal to 100 is to test the median equal to 100. If the median is m_0, then the probability of any observation exceeding the median is 0.5. Therefore, the number of observations in a sample of n which exceed the hypothesized median will have a binomial distribution with parameters n and 0.5.

H_0: X (IQ) has median m0 = 100

Let Y = number of observations > m_0

$$Y \rightarrow B\left(n, \frac{1}{2}\right) = B(10, \frac{1}{2})$$

$$Y = 4$$

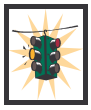
$$p_{value} = P(Y \leq 4)$$

$$= \sum_{k=0}^{4} \binom{10}{k} \frac{1}{2^{10}} = \frac{386}{1024} = 0{,}377$$

The test result shows no inconsistency between the data and the hypothesis.

Notes: a) The following test makes no assumption about the form of the original distribution and can therefore be applied to any distribution.

b) The sign test has lower power than the t test when the original distribution is normal or the z test when the Central Limit Theorem applies but is not affected by departures from these conditions and is not sensitive to outliers.



Solved example 4

- we have same data as with example1

Wilcoxon signed rank test for medians

A second alternative to tests for sample means based on the normal distribution is to replace the observed values by their ranks and calculate a test statistic from the ranks. To test whether the median is equal to some hypothesized value m_0, we first calculate the absolute difference of each observation from m_0. The absolute differences are then replaced by their ranks or the number of their position. The ranks are then signed -1 if the original observation is less than m_0 and +1 if the original observation is greater than m_0. If the hypothesis that the true median equals m_0 is true, then each rank or integer between 1 and n, the sample size has equal probability of being positive or negative. Therefore, the expected value of the mean of the signed ranks should be zero. Therefore calculating the mean and standard deviation of the signed ranks and forming the z-score as we do for the t test would produce a reasonable test statistic.

H_0: X (IQ) has median $m_0 = 100$

$y_i = |x_i - m_0|$
$r_i = \text{rank}(y_i)$
$r_i^* = \text{sgn}(x_i - m_0)\, r_i = \text{signed rank}(y_i)$

For the observations of IQ scores these results are as follows:

IQ score	Absolute difference y_i	Rank of absolute difference r_i	Signed rank r_i^*
93	7	6	-6
94	6	5	-5
77	23	9	-9
80	20	8	-8
103	3	4	4
113	13	7	7
98	2	2	-2
102	2	2	2
65	35	10	-10
102	2	2	2

For the three observations which have the same absolute difference from the hypothesized median, the average of the three ranks has been assigned.

The test statistic of the ranks is calculated as follows. First calculate the mean and standard deviation of the signed ranks.

$$\bar{r} = \frac{\sum_{i=1}^{n} r_i^*}{n} = -2.5; \quad s_r = \sqrt{\frac{\sum_{i=1}^{n} (r_i - \bar{r})^2}{n-1}} = 5.9675$$

Then calculate the z-score for the mean signed rank remembering that the expected value of the mean signed rank is zero under H_0.

$$w = \frac{\bar{r}}{s_r / \sqrt{n}} = -1.325$$

$$p_{VALUE} = W(-1.325) = \Phi(-1.325) = .09257$$

Accept H_0.

Notes:

a) Like the sign test, the Wilcoxon signed rank test makes no assumption about the form of the original distribution. If the original distribution is normal, the Wilcoxon test will have less power than the t test, but will be less sensitive to departures from the assumption of normality. As a general rule for small samples it is reasonable to compute both the usual t test and the Wilcoxon test. If the two tests give very different p-values, this would act as a warning that the original distribution may be seriously non-normal.

b) Because the ranks are fixed pre-determined values, the Wilcoxon statistic will not be sensitive to outliers.

c) Computationally simpler formulas for computing the Wilcoxon test statistic which exploit the fact that ranks are fixed values are given in some books.

8.5. Errors through testing

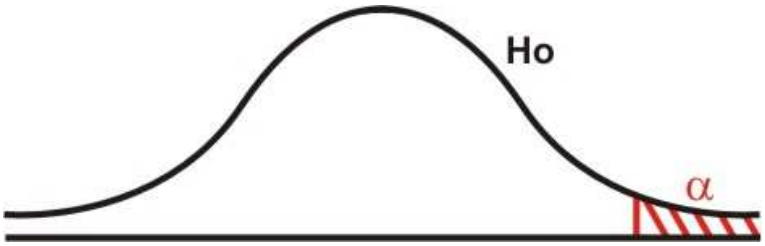
When we make a decision about competing hypotheses, there are two ways of being correct and two ways of making a mistake. This can be depicted by the following table.

		True situation	
		H_0	**H_A**
Decision	**H_0**	OK	Error II
	H_A	Error I	OK

If we make a decision leaning toward H_0 and H_0 is indeed true (true situation), then we did not make an error. If we make a decision leaning toward H_A and H_A is true, then again we did not make an error. These are the probabilities appearing in the upper left and lower right corners. The probabilities in the lower left and upper right are related to the errors made by not making the correct decisions. These probabilities are designated by the Greek letters α ("alpha") and

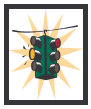
β ("beta").
α is called a Type I Error. It is the probability of falsely rejecting H_0. It is often referred to as the significance level and it represents the risk one is willing to take in rejecting falsely. The user or researcher has complete control over α. Typical (and subjective) values of α are 0.05 and 0 .01. If the consequence of a Type I Error is something in the nature of increased risk of death for a patient or increased risk in financial loses, then one would use a level of significance no greater than 0.01.
β is called a Type II Error. It is the probability of falsely accepting H_0. Unlike a Type I Error, it is difficult to quantify β. More will be said about this later. If the consequence of H_A is extremely attractive and if the results of a Type I Error are not catastrophic, it may be advisable to increase the risk of making a Type I Error and use a level of significance that is 0.05 or higher.
Admittedly it is difficult at this time to fully comprehend these concepts. Hopefully things will make more sense when we go more deeply into hypothesis testing.

$$P(Error\ I) = P(P_{VALUE} < \alpha \mid H_0) = \alpha$$



$$P(Error\ II) = P(P_{VALUE} > \alpha \mid H_A) = \beta$$

8.6. Two sample tests, paired sample tests and tests for proportions



Solved example 5

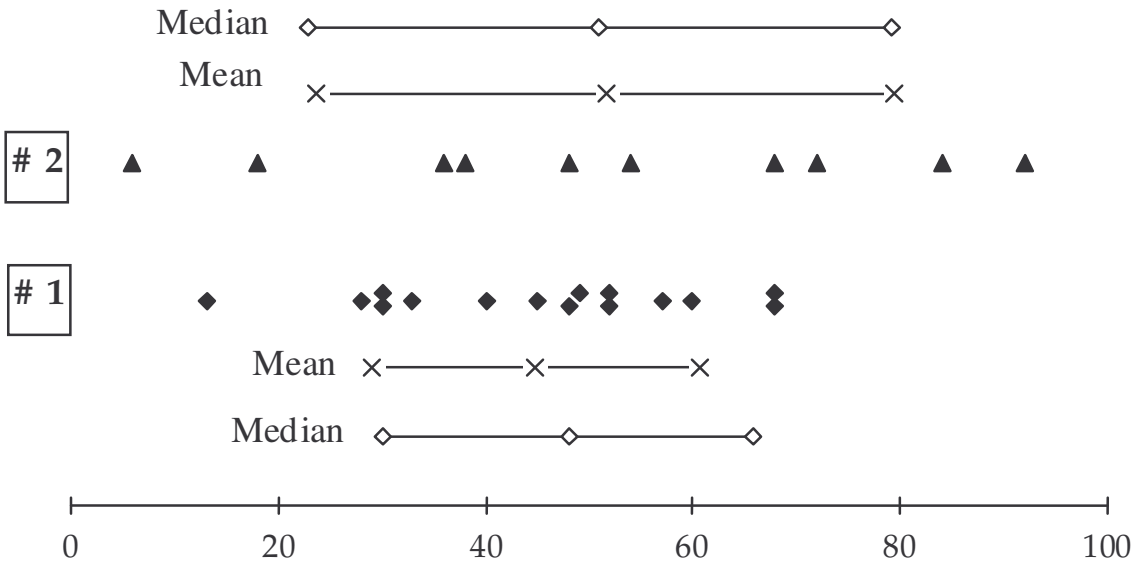
A situation which arises frequently in practice is the two sample test. Two samples have been obtained from difference sources and it is necessary to determine whether the two sources have the same mean or median. One source may be a control group and the other an experimental group. For example, to determine the effectiveness of a new teaching method, a controlled experiment may be conducted in which one group of students, the control group, is taught by traditional methods and a second group by the experimental method. The research question in this case is whether the students taught by the experimental method attained higher results.

Sample from population #1

60	49	52	68	68
45	57	52	13	40
33	30	28	30	48

Sample from population #2

38	18	68	84	72
48	36	92	6	54



The sample means and standard deviations are:

#1: $\overline{X} = 44.867;$ $s_1 = 15.77$

#2: $\overline{Y} = 51.6;$ $s_2 = 27.93$

1. We assume that both samples issue from normal distributions:

$X_i \to N(\mu_1, \sigma_1^2);\quad i = 1,...,n_1 \quad Y_j \to N(\mu_2, \sigma_2^2);\quad j = 1,...,n_2,$

Let:

$\sigma_1 = 15 \quad \sigma_2 = 25$

Test **H_0:** $\mu_1 = \mu_2$

The test statistics is

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \to N(0,1)$$

$$z = \frac{44.867 - 51.6}{\sqrt{\frac{15^2}{15} + \frac{25^2}{10}}} = -0.765$$

$$p_{VALUE} = \Phi(-0.765) = 0.222$$

2. ***Student's t test for difference of means***

The assumption of equality of variance in both populations requites the computation of a single estimate of standard deviation called the pooled sample standard deviation. The pooled standard deviation is the average of squared deviations of all observations from the sample mean of their respective populations. If x_i is the i^{th} sample observation from population #1, and y_j is the j^{th} sample observation from population #2, then the pooled standard deviation is

$$s_p = \sqrt{\frac{\sum_{i=1}^{n_1}\left(x_i - \bar{x}\right)^2 + \sum_{j=1}^{n_2}\left(y_j - \bar{y}\right)^2}{n_1 + n_2 - 2}}$$
$$= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Under the assumption of equal variance in both populations the estimated standard deviation of the difference of sample means will be

$$s_{\bar{x}-\bar{y}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Then the two sample t statistic is computed as:

$$t_{n_1+n_2-2} = \frac{\bar{x} - \bar{y}}{s_{\bar{x}-\bar{y}}} = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and will have a t distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Although the assumption of equal variances is questionable in light of the difference in sample standard deviations, the two sample t test applied to the IQ data above yields the following result.

$$s_p = \sqrt{\frac{14\left(15.77^2\right) + 9\left(27.93^2\right)}{23}} = 21.37$$

$$t_{23} = \frac{44.867 - 51.6}{21.37\sqrt{\frac{1}{15} + \frac{1}{10}}} = -0.772$$

$$p_{VALUE} = t_{n_1+n_2-2}(-0.772) = 0.224$$

3. ***Mann –Whitney or Wilcoxon rank test for difference of medians***

The two sample rank test is equivalent to ranking the total sample from the two populations and calculating the two sample t test using the ranks rather than the original observations. For the IQ data, this gives the following results.

Ranks for population #1

19	14	15.5	21	21
11	18	15.5	2	10
7	5.5	4	5.5	12.5

Ranks for population #2

9	3	21	24	23
12.5	8	25	1	17

The means and standard deviations of the ranks in each population are

$$\bar{r}_1 = 12.1; \qquad s_{r_1} = 6.29$$
$$\bar{r}_2 = 14.35; \qquad s_{r_2} = 8.89$$

The pooled sample standard deviation of ranks is

$$s_r = \sqrt{\frac{(n_1 - 1)s_{r_1}^2 + (n_2 - 1)s_{r_2}^2}{(n_1 + n_2 - 2)}}$$
$$= \sqrt{\frac{14\left(6.29^2\right) + 9\left(8.89^2\right)}{23}} = 7.42$$

The test statistic is

$$w = \frac{\bar{r}_1 - \bar{r}_2}{s_r\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = -0.743$$

$$p_{VALUE} = W(-0.743) = \Phi(-0.743) = 0.229$$

Paired sample tests

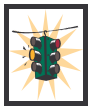
When testing the effect of some experimental condition or comparing the effects of two different conditions, the experimental design often applies either conditions or both experimental and control conditions to the same sampling units or in this case experimental units. The rationale for this design is that variation in experimental results due to differences in sampling units can be eliminated leaving only measurement variation to obscure the effects of the experimental conditions. However, to secure the benefits of reduction in variation offered by this design, the appropriate methods of data analysis and construction of test statistic must be applied.

Suppose two observations under different conditions are taken of n sampling units. For example heart rate before and after exercise. Let X_{i0} be the initial observed value for the i^{th} sampling unit and X_{i1} the subsequent observed value for the same sampling unit. Such a design is called a paired sample design. It is possible to analyze this data and test the hypothesis of no difference between the two experimental conditions using the two sample methods discussed above. However, this approach would failure to take advantage of the opportunity to eliminate variation due to differences in individual sampling units.

A statistically more efficient method to analyze this data is to take advantage of the paired nature of the data and create a single value for each sampling unit. In the simplest data model, this value would be the difference of the two observations for each sampling unit.

$$d_i = X_{i1} - X_{i0}$$

The value d_i is the result only of differences in experimental conditions and experimental error. The methods discussed in the section on one sample tests can then be used to test the hypotheses that the mean or median of d_i is zero which is equivalent to no difference between the two experimental conditions.



Solved example 6

Consider the following example of the heart rates of 12 patients at rest and after ten minutes of exercise.

Resting rate	Rate after Exercise	Difference of Rates	Signed Rank of Difference
42	52	10	3.5
173	175	0	1
113	147	34	11
115	83	-32	-10
69	123	54	12
101	119	20	6
94	69	-25	-7
93	123	30	8.5
112	82	-30	-8.5
67	57	-10	-3.5
104	100	-4	-2
76	89	13	5

The sign test is also applicable to this data. For this data, there are 5 negative signs out of 12 observations. If the true median were 0, the number of negative signs has a binomial distribution with parameters n = 12, and p = 0.5 and the probability if this event is:

$$p_{value} = P(Y \le 5) = \sum_{k=0}^{5} \binom{12}{k} \frac{1}{2^{12}} = 0.387$$

For such a small sample with unspecified population variance, we assume that the observations are normally distributed and apply the Student's t-test.

The mean and standard deviation of the paired differences are

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = 5$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}} = 26.86$$

and the *t*-statistic is

$$t_{11} = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{5}{26.86 / \sqrt{12}} = 0.645$$

p_{value} = from Definition 2 = $t_{11}(0.645) = 0.266$

Applying the Wilcoxon signed rank test to these data yields the following results. The mean and standard deviation of the paired differences are

$$\bar{r} = \frac{\sum_{i=1}^{n} r_i}{n} = 1.33$$

$$s_r = \sqrt{\frac{\sum_{i=1}^{n} (r_i - \bar{r})^2}{n-1}} = 7.55$$

and the W statistic is

$$W = \frac{\bar{r}}{s_r / \sqrt{n}} = \frac{1.33}{7.55 / \sqrt{12}} = 0.611$$

$$p_{value} = 1 - \Phi(0.611) = 0.271$$

Tests for proportions

When testing hypotheses about the proportion of a population having some attribute, the sample size, n, will be large enough in most cases to use the normal approximation to the distribution of the sample proportion. Under the null hypothesis that the population proportion is equal to some specified value,

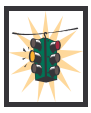
$$H_0: \quad p = p_0$$

The distribution of the sample proportion will be approximately normally distributed for large n.

$$\hat{p} \rightarrow N\left(p_0, \frac{p_0 (1 - p_0)}{n} \right)$$

and the p-value can be calculated from the z-score of the sample proportion.

$$p_{value} = \Phi\left(\frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0 (1 - p_0)}{n}}} \right)$$



Solved example 7

If the manufacturer's specifications for the defective rate of an item is not to exceed 3%, and 7 defective items are found in a sample of 95, then the p-value for testing the hypothesis that the sampled population meets the manufacturers specification is

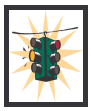
$$z = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0 (1 - p_0)}{n}}} = \frac{7/95 - 0.03}{\sqrt{\frac{(0.03)(0.97)}{95}}} = 2.5$$

$$p_{value} = 1 - \Phi(2.5) = 0.006$$

Reject H_0.

Two sample test for proportions

A two sample test for proportions arises when samples are taken from two populations and the null hypothesis to be tested is that the proportions in both populations are the same. If the samples from each population are large enough, the normal approximation can again to applied to the distribution of the difference of sample proportions. However, since the null hypothesis does not specify a single value for p in each population, the variance is estimated using the total proportion from the samples of both populations which is the maximum likelihood estimate of p under the null hypothesis of equal proportions in both populations.



Solved example 8

Let X_1 be the number of items in a sample of n_1 from population # 1 having the attribute and X_2 be the number of items in a sample of n_2 from population # 2 having the attribute. Then

$$\hat{p}_1 = \frac{X_1}{n_1}; \quad \hat{p}_2 = \frac{X_2}{n_2}; \quad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Then under the null hypothesis that the proportions in the two populations are equal,

H_0: $p1 = p2$

the distribution of the difference in sample proportions is:

$$\hat{p}_1 - \hat{p}_2 \rightarrow N\left(0, \hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

For example, suppose 12 defective items were found in a sample of 88 from one production run and 8 defective items in a sample 92 from a second run. Then,

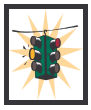
$$\hat{p}_1 = \frac{12}{88}; \quad \hat{p}_2 = \frac{8}{92}; \quad \hat{p} = \frac{12+8}{88+92}$$

Then the z statistic for testing the hypothesis that the defective rate in both runs is the same is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.136 - 0.087}{\sqrt{(0.11)(0.99)\left(\frac{1}{88} + \frac{1}{92}\right)}} = 1.054$$

$$p_{value} = 1 - \Phi(1.054) = 0.146$$

Accept H_0.



Solved example 9

We have two types of floppy disks - Sony and 3M. In any packet are 20 disks. There were found 24 defective disks into 40 Sony packets and there were found 14 defective disks in 30 3M packets. Does difference in the quality of Sony and 3M disks exist?

Solution:

$\hat{p}_1 = \frac{24}{40.20} = 0.030$ (proportion of defective Sony disks)

$\hat{p}_2 = \frac{14}{30.20} = 0.023$ (proportion of defective 3M disks)

$$\hat{p} = \frac{24+14}{(40+30).20} = 0.027$$

1. H_0: $p_1 = p_2$
   H_A: $p_1 > p_2$

2. We select test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}\cdot(1-\hat{p})\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \rightarrow N(0,1)$$

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}\cdot(1-\hat{p})\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.030 - 0.023}{\sqrt{0.027\cdot(1-0.027)(\frac{1}{800}+\frac{1}{600})}} = 0.80$$

3. p-value = 1- Φ (P_2) = 1 - Φ (0,80) = 0,21 Φ (0,80) = 0,79
   /
   (Z has a standard normal distribution)

4. p-value >>> 0.05 ⇒ accept H_0

We can't affirm that there exists statistical important difference in quality of Sony and 3M floppy disks.

Summary of notions

The pure significance test answer a question if given random sample $\underline{X}$ (its observed values) is or is not extreme in relation to some tested hypothesis about population. It consists of 5 steps. The last step concludes about acceptation or rejection H_0. The most often we use hypothesis tests for mean and median: **Student's test**, **Wilcoxon test** for median.
Through conclusion of the pure significance test we can commit errors. Cause we don't know a real situation. In case that we reject H_0 but it is true then we have **type 1 error**. If we accept H_A but H_0 holds in fact we have **type 2 error**.
The following tests are the most often used: **Student's test for difference of mean**, Wilcoxon rank test for difference of medians and **paired tests**. There are the most often used the **tests for proportions** in engineering practice

Question

1. How we get P_{VALUE}?
2. What is the alternate hypothesis?
3. Characterize two sample tests for proportions.

Problems

Example 1: Suppose we want to show that only children have an average higher cholesterol level than the national average. It is known that the mean cholesterol level for all Americans is 190. We test 100 only children and find that mean is 198 and standard deviation is 15. Do we have evidence to suggest that only children have an average higher cholesterol level than the national average?

{*Answer:* reject H_0 , therefore, we can conclude that only children have a higher cholesterol level on the average then the national average. }

Example 2: Nine dogs and ten cats were tested to determine if there is a difference in the average number of days that the animal can survive without food. The dogs averaged 11 days with a standard deviation of 2 days while the cats averaged 12 days with a standard deviation of 3 days. What can be concluded?

{*Answer:* We fail to reject the null hypothesis and conclude that there is not sufficient evidence to suggest that there is a difference between the mean starvation time for cats and dogs. }

9. POINT AND INTERVAL ESTIMATION

Study time: 40 minutes

Aim - you will be able to

- explain the properties of the point estimation
- construct interval estimations for mean, standard deviation and variance

Explication

9.1. Introduction

The estimation problem is distinguished from hypothesis testing. In hypothesis testing we had a preference towards null hypothesis and only rejected it in face of strong contrary evidence. In the case of estimation, all parameter values or potential hypotheses are equal and we want to choose as our estimates those values which are supported by or consistent with the data. An estimate by itself is just a number. Anyone can make an estimate. To be usable, the accuracy of the estimate must also be known. Therefore in addition to deriving estimates, we must also make some assessment of the error of estimation.

9.2. Interval Estimation

The objective of interval estimation is to find an interval of values which have a high likelihood or probability of containing the true parameter values. The strategy used is to find those values which would have a large p-value if they had been chosen as the null hypothesis, i. e. those parameter values which are not inconsistent with the data. In order to give a probability interpretation to the data, we usually choose a fixed p-value, either one-sided or two-sided depending on whether we want a one or two sided interval, and then include in our interval all parameter values whose p-value for the observed data exceeds the chosen minimum p-value, α. The probability of the sample having a p-value which exceeds the selected p-value is $1-\alpha$, and therefore the probability that the interval so constructed will include the true parameter value is also $1-\alpha$. We call the value $1-\alpha$ the confidence level of the interval.

Consider the example of sampling semi-conductor devices to determine the proportion of defective devices produced. In this case, suppose it is a new process and we wish to

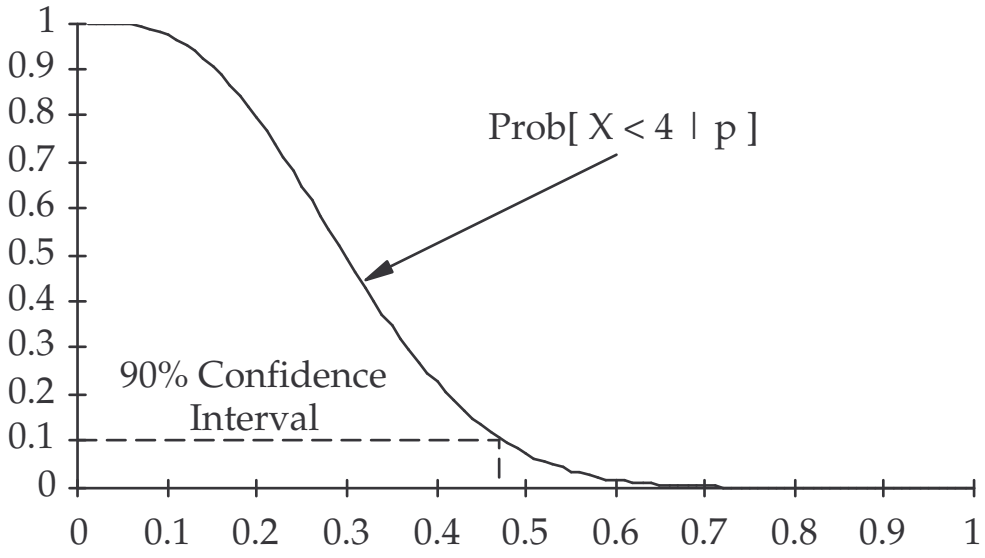
estimate a maximum value for the proportion of defectives. A sample of 12 devices is selected and tested. Three are found to be defective. Since we are interested in an upper bound, we ask how large the true proportion of defectives can be before our observed sample has a very small probability. For some proportion, p, of defective devices, the probability of obtaining less than 4 defectives in a sample of 12 is

$$\Pr ob\left[X < 4 \mid p \right] = \sum_{x=0}^{3} \binom{12}{x} p^x (1-p)^{12-x}$$

To obtain a (1-α) upper for p, we find the value of p such that the p-value is exactly α .

$$\alpha = \sum_{x=0}^{3} \binom{12}{x} p^x (1-p)^{12-x}$$

The following diagram illustrates the p-value as a function of the proportion of defectives, finds the value of p whose p-value is α = 0.1, and identifies the 90% upper confidence limit for p.



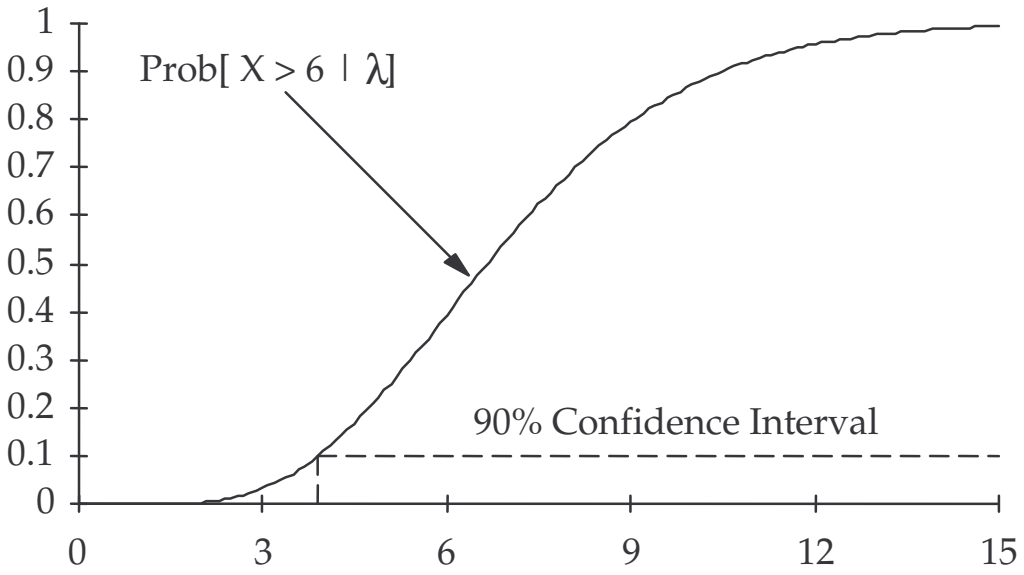
The probability that a population with 47% defectives will have less than 4 defectives in a sample of 12 is 10%. Thus, we are 90% confident that the true proportion of defectives is less than 47%. Another way of expressing this idea is to say that 90% of confidence intervals calculated by this methodology will include the true proportion of defectives.

Consider a second example. A firm which assembles PC's from basic components, loads the software, and tests the system before delivery is interested in estimated how long it takes a worker to complete preparation of a PC for delivery. They observe a worker for 4 hours, one half of his daily work period. In that time, the worker completes 7 PC's. If we assume that the time to complete a single PC is exponentially distributed then the number of PC's completed in 4 hours will have a Poisson distribution. The firm is interested in estimating an upper bound for the mean time to complete a PC or equivalently a lower bound

for the rate at which PC's are completed. Therefore we ask how low the rate λ can be before the probability of our sample result, more than 6, has a very small probability. For a given value of λ , the p-value of our sample is

$$\Pr ob\left[X > 6 \,\middle|\, \lambda \right] = \sum_{x=7}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!}.$$

The following diagram illustrates the p-value as a function of the completion rate λ, finds the value of p whose p-value is α=0.1, and identifies the 90% upper confidence limit for λ.



If the rate of preparing PC's for delivery is per 4 hour period, then the probability of completing 7 or more PC's in 4 hours is 10%. Therefore, we are 90% confident that the true rate of preparation is at least 3.9 computers per 4 hours, or slightly less than one computer per hour. Alternatively the mean time to complete each PC is no more than 61 minutes 32.3 seconds. This is obviously a conservative estimate since in our sample, computers were completed at the rate of 7 per 4 hours or 1.75 per hour with an equivalent mean preparation time of 34 minutes 17 seconds. To obtain a less conservative estimate at the same confidence level, a larger sample size is required.

This analysis depends on the assumption that the time to complete a PC is exponentially distributed. In practice this is unlikely to be a very good model because in theory according to the exponential distribution, the PC could be completed instantaneously. The Poisson process is a more appropriate model for events which occur randomly such as traffic accidents.

Now consider an example where we wish to estimate both an upper and lower bound for the parameter. In this case, we use the p-value for testing hypotheses against two-sided alternatives. The 1-α confidence interval is the set of all parameter values having a p-value greater than α.

Files transmitted via computer networks are often bundled into groups of files having similar network pathways. Into order to determine the optimal number of files to include in a single bundle, network engineers need some estimate of the distribution of file size. A sample of 15 files is taken with the following result. Sizes are in MB units

4.027	1.887	3.806	7.018	2.753
5.956	8.117	2.857	4.525	7.282
0.140	6.186	5.171	10.558	5.534

The summary statistics for these data are

Mean	5.055	Standard. Deviation.	2.646
Median	5.171	1.483*MAD	2.739
Shorth	3.806	to	7.018

For any hypothetical value of the true mean file size, we can compute the t-statistic for our observed sample.

$$t_{n-1}(\mu) = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

and its associated p-value. Since in this case, we want both upper and lower bounds for our estimate of μ, we use the two-sided definition of p-value.

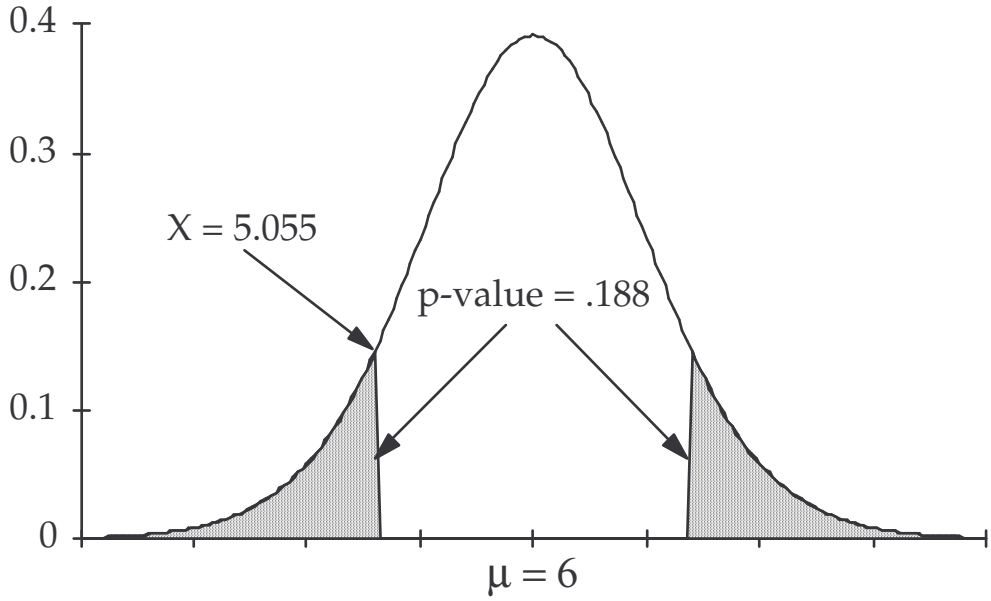
$$p(\bar{x}|\mu) = 2\min\left\{F_{n-1}\left[t(x|\mu)\right], 1 - F_{n-1}\left[t(x|\mu)\right]\right\}$$

For example for a hypothetical value of μ = 6, the observed t-value is

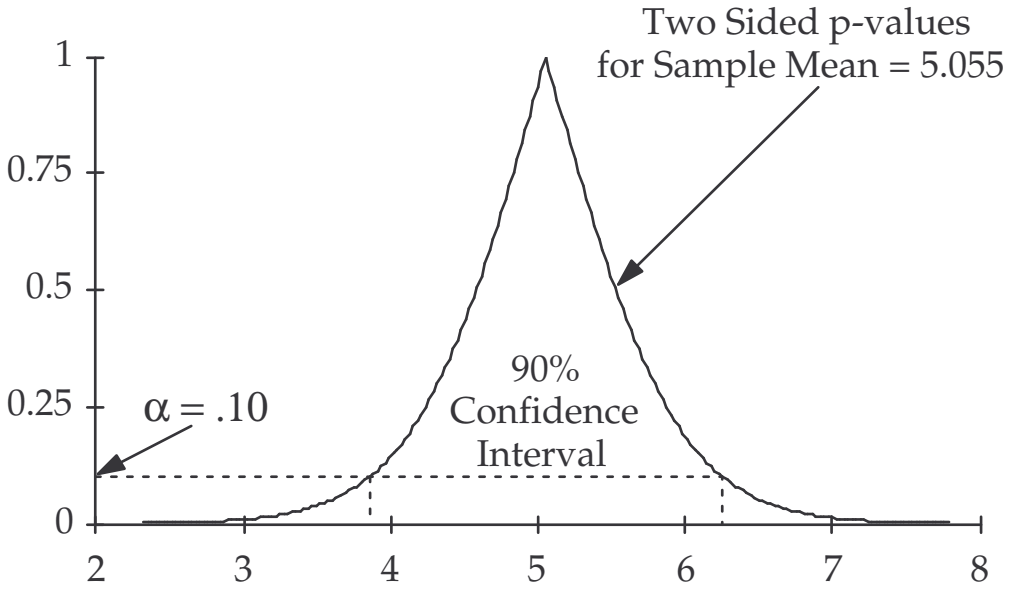
$$t_{14} = \frac{5.055 - 6}{2.646/\sqrt{15}} = -1.384$$

and the associated p-value is

$$p(\bar{x}|\mu) = \Pr ob\left[|t_{14}| \geq 1.384\right] = 0.188$$



Thus, we can determine the p-value associated with any value of μ. The (1-α) confidence interval will consist of all values of whose p-value is greater than α. The following chart shows the two-sided p-value at different values of μ. The p-value reaches its maximum value of 1 when μ = 5.055, the sample mean. If we include in our interval estimate all values of μ having a p-value of at least 10%, then we cane be 90% confident that the interval estimate contains the true mean value in the sense that 90% of interval estimates derived by this procedure contain the true value of the mean. We call such an interval a 90% confidence interval. In this example, the 90% confidence interval for mean file size



is the range (3.852, 6.258). Notice that μ = 6 with a p-value of 0.188 is included in the 90% confidence interval.

9.3. Construction of Confidence Intervals

There is a simple procedure for constructing confidence intervals for parameters whose test statistic has a symmetric distribution, such as the Student's t or the normal. This procedure eliminates the need to compute the p-value for every value of μ. To construct a (1-α) confidence interval, we need only determine the upper and lower bounds of the interval. The lower bound will be that value of μ less than the sample mean whose p-value is exactly α. Therefore the t-value of the sample mean with respect to the lower bound must be equal to the (1-α /2) percentage point of the Student's t distribution.

$$t_{n-1,\alpha/2} = \frac{\bar{x} - \mu_{Lower}}{s/\sqrt{n}}$$

Solving for μ_{Lower}, we have

$$\mu_{Lower} = \bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

Similarly μ_{Upper} must satisfy the equation

$$t_{n-1,\alpha/2} = \frac{\mu_{Upper} - \bar{x}}{s/\sqrt{n}}$$

Hence the (1 - α) confidence interval is given by

$$\left(\mu_{Lower}, \mu_{Upper}\right) \Leftrightarrow \bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

In our example of files sizes, the 90% confidence interval is

$$\bar{x} \pm t_{14,.05}\frac{s}{\sqrt{n}} \Leftrightarrow 5.055 \pm 1.761\frac{2.646}{\sqrt{15}} \Leftrightarrow 5.055 \pm 1.203$$

As previously determined graphically, this interval is (3.852, 6.258).

9.4. Sample Size Determination

As well as giving a range of reasonably good parameter values, an interval estimate also provides information about the quality of the estimated values. The quality of an estimate has two aspects,

1) Accuracy

2) Reliability

The accuracy of a interval estimate is equivalent to the length of the interval. The smaller the interval, the greater the accuracy. Reliability is given by the confidence level of

the interval. However, as for Type I and Type II errors of hypothesis tests, accuracy and reliability of a confidence interval are in conflict. For a fixed sample size, the confidence level can only be increased by increasing the length of the interval thereby reducing its accuracy. Increasing both the accuracy and reliability can only be achieved by increasing the sample size.

Determining the sample size required to construct an interval estimate having some fixed reliability and accuracy is a problem which arises commonly in practice. Suppose it is necessary to estimate a mean to an accuracy of Δ with a reliability of $(1-\alpha)$. That is, we require a $(1-\alpha)$ confidence interval of length not exceeding 2Δ. Then the sample size must be chosen large enough to satisfy the following inequality.

$$\Delta \geq t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

and therefore the sample size must exceed

$$n \geq \left[\frac{t_{n-1,\alpha/2}\, s}{\Delta} \right]^2$$

In practice it is common to substitute a conservative estimate for s and substitute the z quantile for the t quantile on the assumption that the required sample size will be large enough that the applicable t distribution will be approximately normal. If we wished to estimate file size to an accuracy of 256 KB, or .25 MB with 90% confidence, using a conservative estimate of 3 for s, we would require a sample size of

$$n \geq \left[\frac{z_{\alpha/2}\, s}{\Delta} \right]^2 = \left[\frac{1.645 * 3}{.25} \right]^2 = 389.67$$

9.5. Point Estimation

Randomness is difficult and unpopular. We are used to have specific answers to questions. An interval of estimates can be unsatisfying. We may ask "What is the single best point or value in the interval?" Such a single value would be a point interval. The single best value is clearly the value which has the highest p-value for the observed data. This point estimate is called the maximum likelihood estimate. But notice that as the p-value increases and the size of the confidence interval decreases, the confidence level decreases accordingly. In most cases, the maximum p-value will be 1, so the confidence level will be zero. That is, the point estimate will never be exactly correct. Therefore in this case it is extremely important to estimate the error of estimation.

While choosing the single point estimate to be the parameter value assigning the maximum p-value to the observed data is a logical consequence of the method of constructing confidence intervals, it may have indeterminate or non-unique solutions in certain cases, particularly for discrete random variables. Therefore, point estimates are determined by a method similar in spirit to maximizing p-values, the method of maximum likelihood. Rather than maximizing the p-value, maximum likelihood point estimates seek the parameter value which maximizes the probability mass or probability density of the observed sample data.

9.6. Maximum Likelihood Estimation

Because statistics measure general distributional properties such as location and scale, means, medians, and standard deviations can be applied to any distribution. But estimators are associated with parameters for a specific distribution. Developing satisfactory estimators for every individual distributional form would become a daunting task without some general procedure or approach. Fortunately, the idea of likelihood offers such an approach. Intuitively, if the conditional probability or likelihood of the observed data is greater for one parameter value than another, then the first parameter value is a preferred estimate of the population parameter. By extension, the best choice of estimate for the population parameter should be the parameter value whose likelihood is maximum.

$$\hat{\theta} = \max_{\theta} \Pr ob\left[x \middle| \theta\right] = \max_{\theta} f(x, \theta)$$

An estimator derived by this criterion is called a maximum likelihood estimator or MLE and is always denoted with a small cap over the parameter symbol as indicated.

Consider the case of sampling for an attribute. If X is the number of items in a sample of size n having the desired attribute, then X will have a binomial distribution with parameters n which is known from the sampling procedure and p which is unknown. The likelihood of p for the observed X is the conditional probability of X given p.

$$f(p \mid x) = \Pr ob[x \mid p] = \binom{n}{x} p^x (1-p)^{n-x}$$

We can find the maximum likelihood estimate of p by finding the point of the likelihood function having zero slope. That is, by solving the following equation.

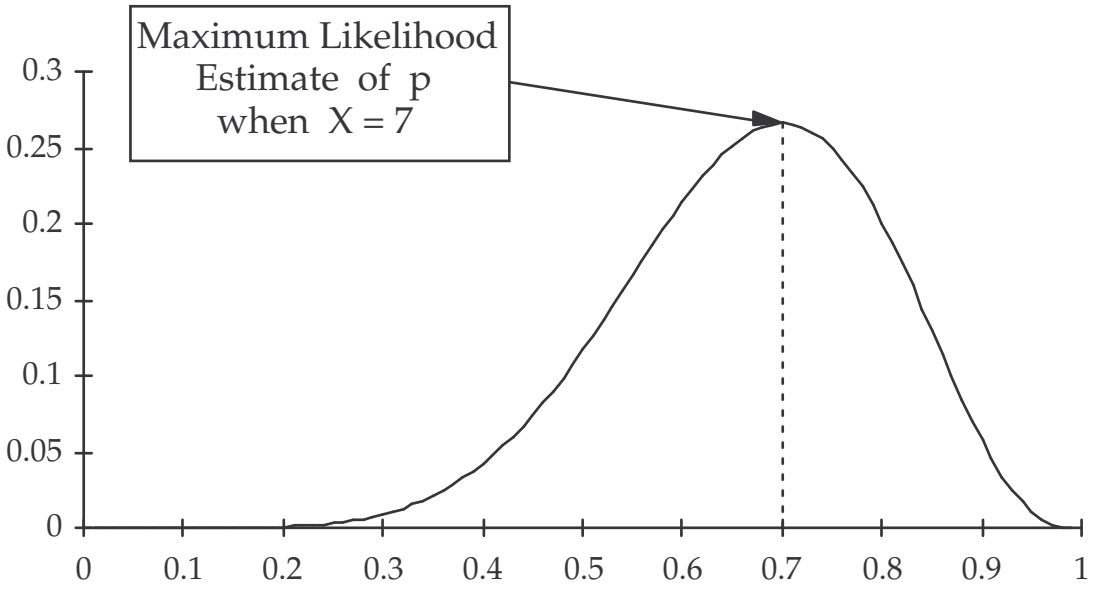
$$\frac{\partial f(p \mid x)}{\partial p} = \frac{\partial \binom{n}{x} p^x (1-p)^{n-x}}{\partial p}$$

$$= \binom{n}{x}\left[\mathrm{x}p^{x-1}(1-p)^{n-x} - (n-x)p^x(1-p)^{n-x-1}\right] = 0$$

The solution is the sample proportion. We know from the sampling distribution of this estimator that it is unbiased. It is also consistent, sufficient, and efficient among unbiased estimators.

$$\hat{p} = \frac{x}{n}$$

The maximum likelihood estimate for n = 10 and X = 7 is illustrated below.



The maximum likelihood estimator of λ for the Poisson distribution is derived in the same fashion.

$$f(\lambda \mid x) = \Pr ob[x \mid \lambda] = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

Solving for the value of λ where the slope of the likelihood function is zero

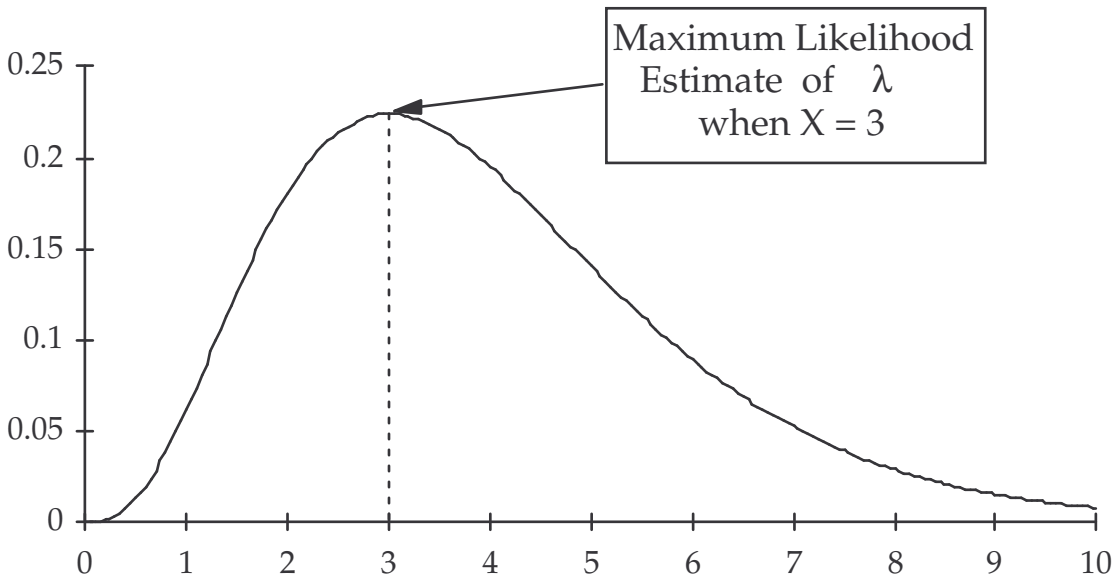
$$\frac{\partial f(\lambda \mid x)}{\partial \lambda} = \frac{\partial \dfrac{(\lambda t)^x e^{-\lambda t}}{x!}}{\partial \lambda}$$

$$= \frac{x(\lambda t)^{x-1} te^{-\lambda t} - (\lambda t)^x te^{-\lambda t}}{x!} = 0$$

we find

$$\hat{\lambda} = \frac{x}{t}.$$

The maximum likelihood estimate of λ when X = 3 and t = 1 is illustrated below.



The maximum likelihood estimator of λ is also unbiased, efficient, consistent, and sufficient for the Poisson distribution.

For the normal distribution, the maximum likelihood estimate of μ is obtained by minimizing the value in the exponent of the density.

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

The solution is the sample mean.

9.7. Estimation, Estimators, Properties of Estimator

What is an Estimate or an Estimator?

Formally, an estimator is a statistic defined on the domain of the sample data which is used as an estimate a population parameter. An estimate is the value of that statistic for a particular sample result. Since every statistic has a theoretical probability distribution for every hypothetical probability distribution of the population, it is possible to examine the properties of an estimator through its probability distributions. Since many estimators are based on likelihood functions, many of the properties of the likelihood function will also be exhibited in the probability distribution.

Because of the special requirements of an estimator, criteria particular to the problem of estimation have been proposed as means of evaluating the suitability of a statistic as an estimator, of comparing competing estimators, and of developing new and improved estimators. You will not that the following criteria can only be applied to statistics which are required to be close to some parameter, hence to estimators.

a) Consistency

Consistency is generally agreed to be an essential characteristic of an estimator. As the sample size increases, an estimator which is consistent will have smaller and smaller probability of deviating a specified distance from the parameter being estimated. In the limit of an infinitely large sample, the value of the estimator will be equal to the estimated parameter with probability one. An estimator which does not have this property would give more reliable results for smaller samples and therefore could not be using the information in the sample consistently.

b) Sufficiency

Every statistics is a reduction or summarization of the original sample data and as such discards some information contained in the original sample data. An estimator which is sufficient does not discard any information relevant to the parameter being estimated. This may seem at first to be a rather vague requirement but in fact sufficiency has a very specific probabilistic definition. Any statistic creates a partition of the sample space. All elements within any partition lead to the same value of the statistic. If the relative or conditional probabilities of the individual sample space elements are independent of the parameter being estimated, then the partition and the estimator which created it are sufficient. For example, the binomial random variable partitions the sample space of n Bernoulli trials into subsets all having the same number of successes. The probability of two elements having the same number of successes is always the same no matter what the value of p, the probability of success. Therefore no further information about p can be obtained by knowing which element in the partition actually occurred. Therefore, the sample proportion is a sufficient statistic or estimator of p. We always try to work with sufficient statistics.

c) Bias

Bias or unbiasedness concerns the expected or mean value of the statistic. The statistic should be close to the parameter being estimated and therefore its mean value should be near the parameter value. Bias is the difference between the mean of the estimator and the value of the parameter. Bias should be small. If bias is zero, we say the estimator is unbiased. Unbiased estimators are not always preferable to biased ones if they are not sufficient or have larger variance. An estimator need not be unbiased in order to be consistent.

d) Efficiency

As has be said several times and reinforced by the orientation of the preceding criteria, an estimator should be close to the parameter being estimated. Simply being unbiased will not insure that the estimator is close to the parameter. The variance of the sampling distribution of the estimator must be small as well. For two estimators, the estimator whose sum of variance and squared bias is smaller is said to be more efficient. If the two estimators are unbiased, then the estimator with smaller variance is more efficient. If an unbiased estimator has minimum possible variance for all unbiased estimators, then it is said to be efficient.

Summary of notions

From a methodological point of view we use two kinds of parameters estimations. It is a **point estimation** where distribution parameter is approximated by a number and so called **interval estimation** where this parameter is approximated by an interval where the parameter belongs with a high probability. **Unbiassed, consistent and efficient** estimations of parameters are the most important for a quality of point estimation.

In the case of interval estimation of a parameter we can search for **two-sided** or **one-sided** estimations.

Questions

1. What is the consistent estimation of parameter?
2. How we can describe 100.(1-α) % two-sided confidence interval for a θ-parameter?

Problems

Example 1: In random sample of chipsets there is a 10% not suitable for new quality demands. Find 95% confidence interval for a p-chipset proportion not suitable for a new norm if a sample size is:

a) n = 10
b) n= 25
c) n = 50
d) n = 200

{ *Answer:* a) –0.06<p<0.26, b) 0.00<p<0.20, c) 0.03<p<0.17, d) 0.07<p<0.13}

Example 2: Four students were random selected from the first parallel group. Their exam results were 64, 66, 89 and 77 points. Three students were random selected from the second parallel group. Their exam results were 56, 71 and 53 points. Find 95% confidence interval for difference between means values of exam results.

{ *Answer:* (-4;32)}

10. ANOVA – One Factor Analysis of Variance

Study time: 60 minutes

Aim - you will be able to

- explain structure of *F-ratio*
- conclude by the test named Analysis of Variance
- construct the ANOVA table
- realize the post hoc analysis

Explication

10.1. Introduction

We talked about one-sample and two-sample tests for mean in the previous lectures. The analysis of variance (ANOVA) is an extension of these tests. It enables compare any mean of independent random samples. The analysis of variance (in its parametric form) assumes normality of the distributions and homoscedasticity (identical variances). If these conditions are not executed then we must use nonparametric *Kruskall-Wallis test*. It is an analog of the one-factor sorting in the analysis of the variance. It doesn't assume distributions normality but its disadvantage is a smaller sensitivity.

10.2. Construction of the F-statistic

Let's have *k*-random samples that are independent on each other. These samples have the standard distribution with the same variation:

$(X_{11}, X_{12}, ..., X_{1n_1}) \rightarrow N(\mu_1, \sigma^2)$

$(X_{21}, X_{22}, ..., X_{2n_2}) \rightarrow N(\mu_2, \sigma^2)$

...

$(X_{k1}, X_{k2}, ..., X_{kn_k}) \rightarrow N(\mu_k, \sigma^2)$,

---

$$\sum_{i=1}^{k} n_i = N$$

Let n_i ... number of observations in i-th sample.

***Formulation of the problem*:**
The hypothesis of interest is H_0: $\mu_1 = \mu_2 = ... = \mu_k = \mu$
The alternate hypothesis is: H_A: At least two μ_i's are different.

We want determine on H_0 in terms of one test. Cause we try to find such test statistic that enable not only H_0 implementation but also it was sensitive on the H_0 validity.

Define the **total sum of squares** (or **total variability**) as

$$SS_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2 \text{ , where } \overline{X} \text{ is the mean of all observations.}$$

- the total sum of squares is our raw measure of variability in the data

This total sum of squares we can separate into 2 components:

$$SS_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2 \Rightarrow SS_{TOTAL} = SS_W + SS_B \text{ ,}$$

where
SS_W ... the within group variation (the sum of squares within groups) - is the raw variability within samples

$$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2$$

- the degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: N - k

S_i is a sample standard deviation of i-th random sample:

$$S_i = \sqrt{\frac{\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2}{n_i - 1}}$$

and

$$\overline{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} \quad \text{is a sample mean in } i\text{-th sample.}$$

SS_B ... the between group variation (the sum of squares between groups) - is the raw variability between samples:

$$SS_B = \sum_{i=1}^{k} n_i \cdot (\overline{X}_i - \overline{X})^2$$

The within group variance (mean square within groups): $S_W^2 = \frac{SS_W}{N-k}$

The between group variance (mean squares between groups): $S_B^2 = \frac{SS_B}{k-1}$

Properties of these variances:

1. $ES_W^2 = \frac{1}{N-k} E(\sum_{i=1}^{k} (n_i - 1) S_i^2) = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) E(S_i^2) = \sigma^2$

because $E(S_i^2) = \sigma^2$.

The within mean square is an unbiased estimate of the variance, independently of H_0.

2. $ES_B^2 = \sigma^2 + \frac{1}{k-1} \sum_{i=1}^{k} n_i (E\overline{X} - E\overline{X}_i)^2$

$ES_B^2 = \sigma^2 \Leftrightarrow$ when H_0 is true

Therefore the ratio of the two sums of squared divided by their degrees of freedom will have an F distribution under the hypothesis of equal population means.

$$F = \frac{SS_B / k-1}{SS_W / N-k} = \frac{S_B^2}{S_W^2}$$

Definition:
We call this F statistic as ***F-ratio***.

Why is useful use *F*-ratio as the test statistic?

We see that if H_0 is true then *F-ratio* is any random number close to 1 ... $F \approx 1$. If H_0 is faslse then this number is markedly bigger than 1 (see property 2). The statistic *F-ratio* is sensitive to validity of the hypothesis H_0. So we can use it during following testing as test statistic we have to determine its statistical behavior what means to determine its propability distribution.

We know that $\frac{S_w^2}{\sigma^2} \cdot (N-k) = \sum_{i=1}^{k} \frac{(n_i - 1) S_i^2}{\sigma^2} \rightarrow \chi^2(N-k)$,

because $\frac{(n_i - 1) S_i^2}{\sigma^2} \rightarrow \chi^2(n_i - 1)$, and further is known that sum of random variables $\chi^2(n_i - 1)$ is again a random variable of a same type with degrees of freedom number same as summarized variables.

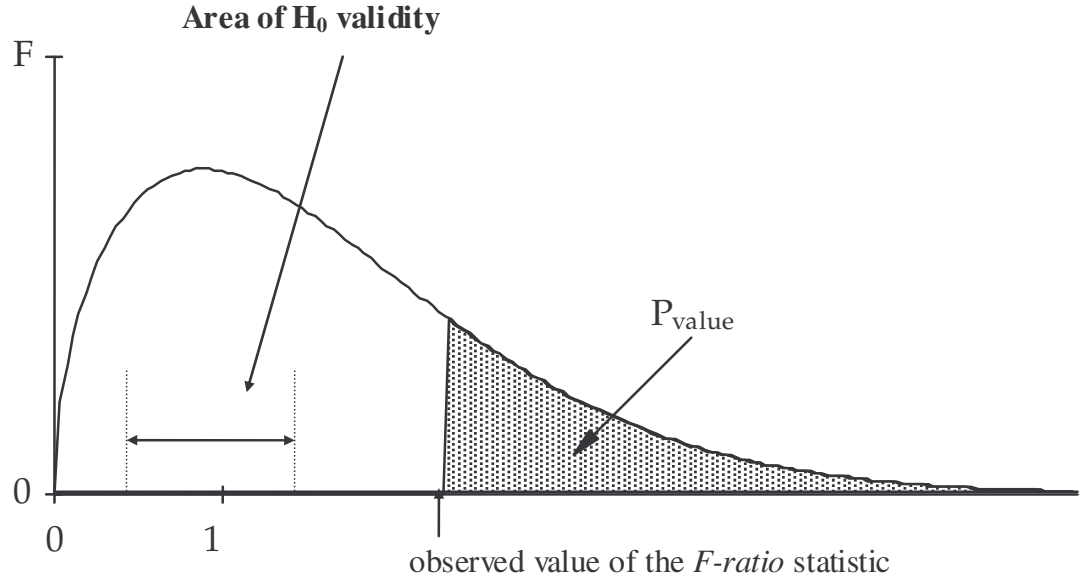
If H_0 is true then: $\frac{S_B^2}{\sigma^2} \cdot (k-1) = \frac{1}{\sigma^2} \sum_{i=1}^{k} n_i \cdot (\overline{X}_i - \overline{X})^2 = \sum_{i=1}^{k} (\frac{\overline{X}_i - \overline{X}}{\frac{\sigma}{\sqrt{n_i}}})^2 \rightarrow \chi^2(k-1)$

Then we know (Fisher-Snedecor distribution) that following ratio:

$$\frac{\frac{S_B^2}{\sigma^2} \cdot (k-1)}{\frac{k-1}{\frac{S_w^2}{\sigma^2} \cdot (N-k)}} = \frac{S_B^2}{S_W^2} = F_{k-1, N-k}$$

it must have *F distribution* with *(k-1)* and *(N-k)* degrees of freedom.
If we know a F-ratio statistical behavior we can use it for analysis and determination of previously stated problem in H_0. Following figure illustrated a usage of *F-ratio* to determine

a hypothesis H_0 validity.



10.3. ANOVA Table

We summarize the data in an ANOVA table:

Source	Sum of squares	Degrees of freedom	Mean squares	*F-ratios*	P-value
total	$SS_{TOTAL} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^2$	$N-1$			
between	$SS_B = \sum_{i=1}^{k} n_i \cdot (\overline{X}_i - \overline{X})^2$	$k-1$	$S_B^2 = \frac{SS_B}{k-1}$	$F = \frac{S_B^2}{S_W^2}$	see definition
within	$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$	$N-k$	$S_W^2 = \frac{SS_W}{N-k}$		

Analysis of variance table - ANOVA

The big values of F-ratio indicate small values of p_{value} what means rejection of H_0. The F-ratio value will be a big number if the within group variation is a negligible part of the total variability and equivalently if the between variation is a significant part of the total variability.

10.4. Solved example

We assume three data sets for illustration of *F-ratio* statistical behavior. In each data set, the sample means are the same but the variations within groups differs. When the within group variation is small than the *F-ratio* is large. When the within group variation is large than the *F-ratio* is small. The examples illustrate three cases: small within group variation; normal within group variation; and large within group variation.

Example 1:

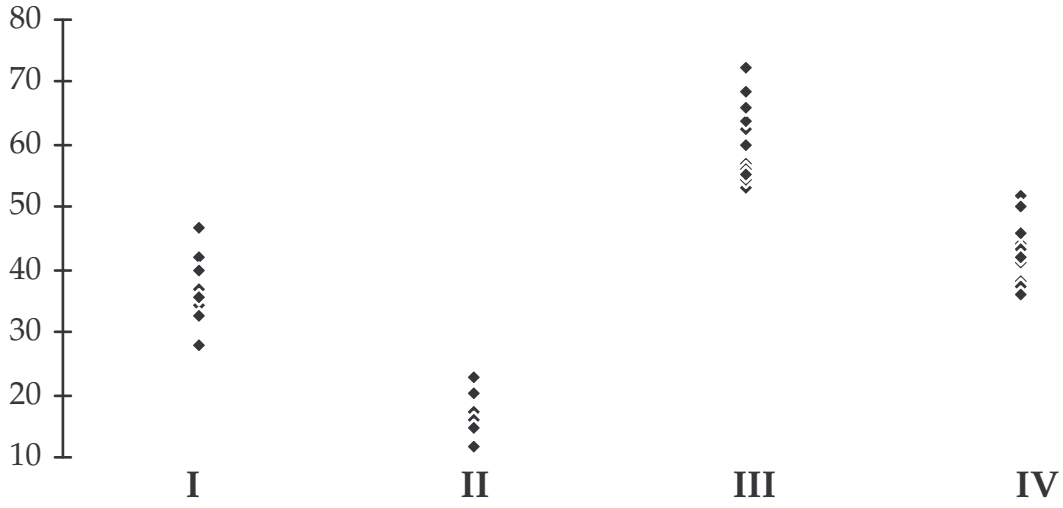
Small within group variation

Groups	I	II	III	IV
Data	42 34.5 32.5 40 46.5 28 37 35.5	17.5 12 16 15 20.5 23 15	68.5 72 53 64 57 56 54.5 62.5 63.5 60 66 55	38 44 52 50 43.5 41 42 46 37.5 36
Sample size	8	7	12	10
Group means	37	17	61	43
Group standard deviations	5.78	3.71	6.06	5.27

ANOVA Table

	Degrees of freedom	Sum of squares	Mean squares	*F-ratio*
total	36	9872.7027		
between	3	8902.7027	2967.57	100.96
within	33	970	29.39	

P-value = 0.0000



Example 2:

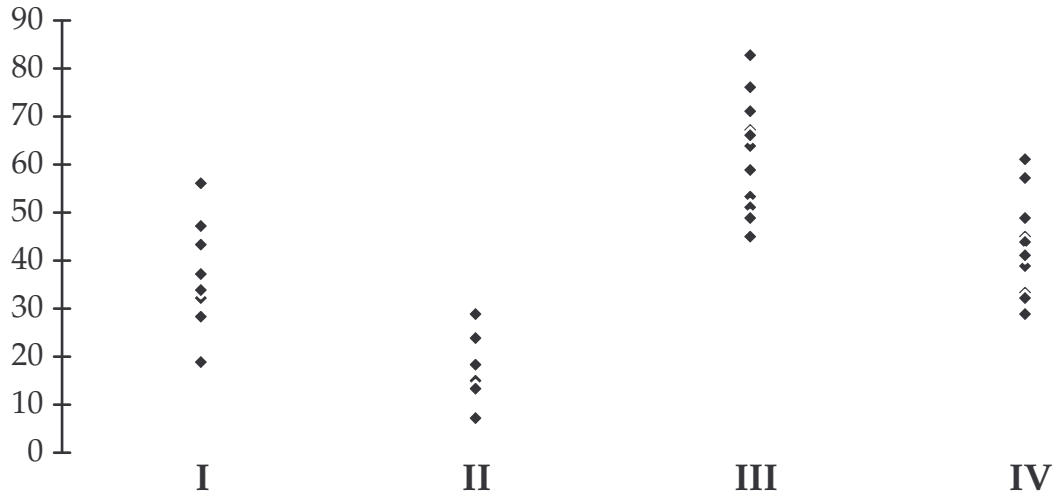
Normal within group variation

Groups	I	II	III	IV
Data	47 32 28 43 56 19 37 34	18 7 15 13 24 29 13	76 83 45 67 53 51 48 64 66 59 71 49	33 45 61 57 44 39 41 49 32 29
Sample size	8	7	12	10
Group means	37	17	61	43
Group standard deviations	11.56	7.42	12.12	10.53

ANOVA table

	Degrees of freedom	Sum of squares	Mean squares	*F-ratio*
total	36	12782.7027		
between	3	8902.7027	2967.57	25.24
within	33	3880	117.58	

P-value = 0.0000



Example 3:

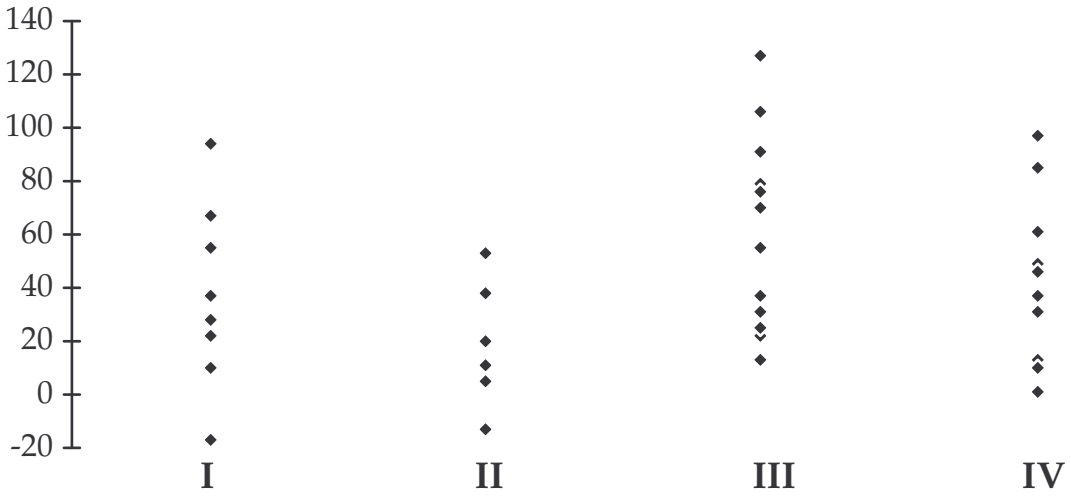
Large within group variation

Groups	I	II	III	IV
Data	67 22 10 55 94 -17 37 28	20 -13 11 5 38 53 5	106 127 13 79 37 31 22 70 76 55 91 25	13 49 97 85 46 31 37 61 10 1
Sample size	8	7	12	10
Group means	37	17	61	43
Group standard deviations	34.69	22.25	36.36	31.59

ANOVA table

	Degrees of freedom	Sum of squares	Mean squares	*F-ratio*
total	36	43822.7027		
between	3	8902.7027	2967.57	2.804
within	33	34920	1058.18	

P-value = 0.0549



10.5. Post Hoc analysis

A large *F-ratio* indicates only that some differences exist among the group means, but not where those differences occur. If the *F-ratio* is large, our analysis would be incomplete without identifying which group means differ. This process is called **post hoc** analysis, and consists of comparing the means of all pairs of samples to determine if there is a difference of means.

Several methods are available for post hoc multiple comparisons. We will discuss the simplest method here, least significant differences. The Least Significant Difference or **LSD-method** consists of applying the two-sample *t* test to every pair of sample means. However, we make one adjustment and use the square root of mean square within rather than the pooled standard deviation from the two samples as our estimate of population standard deviation. Thus for any pair of sample means, we compute LSD as,

$$(LSD)_{i,j} = \frac{\overline{X}_i - \overline{X}_j}{S_W \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \rightarrow t_{N-k}$$

where $S_W = \sqrt{S_W^2} = \sqrt{\frac{SS_W}{N-k}}$.

This statistic has Student distribution with *N-k* degrees of freedom.

The LSD method is illustrated for the three examples given previously.

Example 1: Small within group variation

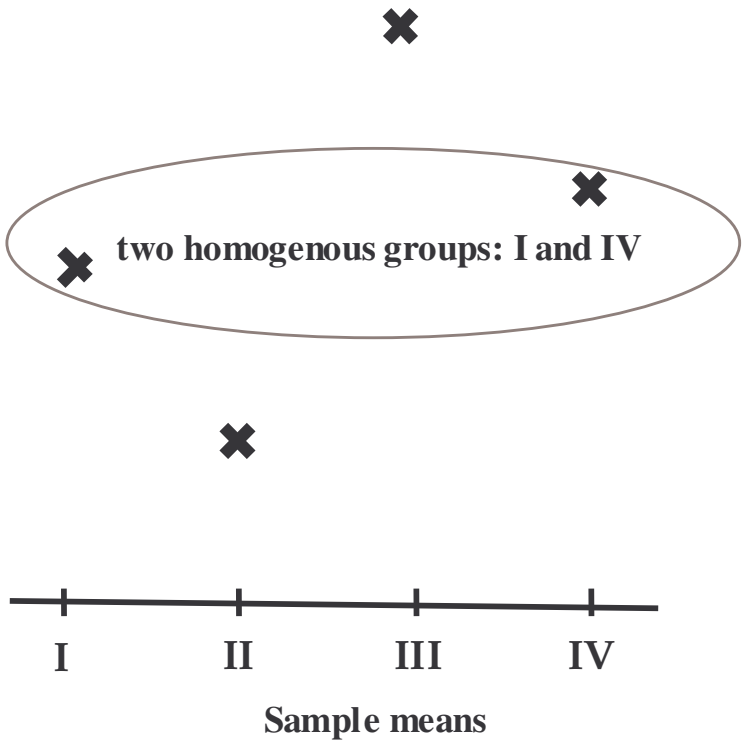
We determine $(LSD)_{i,j}$ for all pairs of given four groups and the obtained values we inscribe in the following table:

Sample sizes		8 I	7 II	12 III	10 IV
8	I	0	-7.128	9.698	2.333
7	II	7.128	0	17.064	9.731
12	III	-9.698	-17.064	0	-7.754
10	IV	-2.333	-9.731	7.7541	0

In this case, there is very strong evidence of differences between all groups except I and IV where the evidence is not as strong.

Example 2: **Normal within group variation**

Sample sizes		8 I	7 II	12 III	10 IV
8	I	0	-3.564	4.849	1.167
7	II	3.564	0	8.532	4.8656
12	III	-4.849	-8.532	0	-3.877
10	IV	-1.167	-4.866	3.877	0



In this case, even though the sample means are the same, there is no evidence of differences between the means of Groups I and IV. Therefore there are essentially three Groups: II; III; and I and IV together.

Example 3: **Large within group variation**

Since the *F-ratio* for this example is very small, we would normally conclude that there is no evidence against the null hypothesis of equal group means and not proceed further. Any two-sample *t* test which produces small p-values should be regarded as spurious. However, for illustration, we have produced the table of least significant differences.

Sample sizes		8 I	7 II	12 III	10 IV
8	I	0	-1.188	1.616	0.389
7	II	1.188	0	2.844	1.622
12	III	-1.616	-2.844	0	-1.292
10	IV	-0.389	-1.622	1.292	0

In this example, the only least significant difference which has a small **p-value** is between groups II and III. However, because the overall *F-ratio* was too small, this difference would be disregarded and we would conclude that no differences exist between the means of any of the groups.

Note:
There exists other tests then LSD method which allows similar multiple comparisons what means a post hoc analysis. Also there were developed more flexible methods which are accessible thru the more advanced software (e.g. Duncan test, Tukey test for significant differences, Scheffe test and Bonferoni test). These tests are based on similar decision strategy and that's on setting of a critical difference requested for determination if two sample means from several groups are different. In many cases these tests are much more effective than LSD method.

10.6. Kruskal-Wallis test

The *F-ratio* test statistic used in the standard analysis of variance is known to be very sensitive to the assumption that the original observations are normally distributed. Because the test statistic is based on squared deviations from the mean, it can be badly distorted by outliers. For two-sample analysis, the Wilcoxon/Mann Whitney rank test was introduced as a nonparametric alternative which is less sensitive to outliers than the *t* test. For multiple samples, the Kruskal-Wallis rank test can be used for the same purpose. Like the Wilcoxon/Mann Whitney test, the Kruskal-Wallis test substitutes the ranks of the original data values and performs an analysis of variance on the ranks. For the large deviation data of the previous example the ranks for each group are listed in the following table.

Groups	I	II	III	IV
	28	11	36	9.5
Ranks	12.5	2	37	23
of original	6.5	8	9.5	35
data	25.5	4.5	31	32
	34	21	19	22
	1	24	16.5	16.5
	19	4.5	12.5	19
	15		29	27
			30	6.5
			25.5	3
			33	
			14	
Sample size	8	7	12	10
Mean rank	17.6875	10.7143	24.4167	19.35
Standard deviation	11.1674	8.5919	9.6668	10.6538

The test statistic is a modification of calculating the *F-ratio* for the ranks. In this example, the test statistic and p-value are:

K-W test statistics = 7.24325 p-value = 0.0645

The p-value for the Kruskal-Wallis test is slightly higher than for the *F-test*, but the conclusions are the same in both cases. The null hypothesis of equal group means is not rejected.

Summary of notions

Analysis of variance (ANOVA) is an extension of the two-sample tests for means and it enables compare any mean of independent random samples. ***F-ratio*** is the test statistics in analysis of variance. *F-ratio* statistics is sensitive to validity of the hypothesis H_0, which is formulate as an equality of the samples means. Particular interresults (that we execute during analysis of variance) are recorded into **ANOVA table**. The second step (in ANOVA) is **post hoc** analysis, and consists of comparing the means of all pairs of groups of purpose to choose homogenous groups. **LSD-statistics** is a criterion for assignment to homogenous groups.
Described procedure ANOVA is sensitive to the assumption that the original observations are normally distributed. If this condition is not executed then we must use nonparametric ***Kruskall-Wallis*** *rank test*.

Questions

1. Describe construction and statistical behavior of the *F-ratio* statistics.
2. What is the usual output from analysis of variance?
3. What is post hoc analysis?

Problems

Example 1: We made a research of dependency of earning and achieved education. In the table there are earnings in thousand CZK at 7 randomly selected men at each level of education. (B - basic, H - high, U - university).

	B	H	U
1	10.9	8.9	11.2
2	9.8	10.3	9.7
3	6.4	7.5	15.8
4	4.3	6.9	8.9
5	7.5	14.1	12.2
6	12.3	9.3	17.5
7	5.1	12.5	10.1

Do a simple sorting and determine if education does influence earning.

{*Answer:* p-value = 0.057}

Example 2: From a large set of homes we randomly selected 5 single homes, 8 couple, 10 three-member, 10 four-member and 7 five-member homes. We watched their month spending for food and drinks for one family member (in CZK). Confirm by analysis of variance if a month spending for food and drinks depends on a number of family members.

	Spending for one family member (in CZK)				
Number of family members	**1**	**2**	**3**	**4**	**5**
	3.440	2.350	2.529	2.137	2.062
	4.044	3.031	2.325	2.201	2.239
	4.014	2.143	2.731	2.786	2.448
	3.776	2.236	2.313	2.132	2.137
	3.672	2.800	2.303	2.223	2.032
		2.901	2.565	2.433	2.101
		2.656	2.777	2.224	2.121
		2.878	2.899	2.763	
			2.755	2.232	
			3.254	2.661	

{*Answer:* Use suitable software package.}

11. SIMPLE LINEAR REGRESSION

Study time: 60 minutes

Aim - you will be able to

- explain a general linear model notion
- explain a linear regression model principle
- use regression analysis results
- verify a regression model by determination index

Explication

11.1. Introduction

Mathematical formulation of statistical models

Symbolically, the basic additive formulation of statistical models can be expressed as

$$\underline{Y} = f(X) + \zeta(\varepsilon)$$

where Y is the observed value, $f(X)$ is the systematic component and $\zeta(\varepsilon)$ is the random component. This schematic model explicitly identifies three type of variables.

Y – Response, Criterion, dependent Variable (observed value of primary interest)
X– Predictor, Stimulus, Independent Variable (those factors to which the value of the systematic component may be attributed)
ε - random error

Only Y and X are observable. Random error is always unobservable.
$\zeta(\varepsilon)$ is always estimated as the residua difference between the estimated systematic component and the observed response, Y.

$$\overline{\zeta(\varepsilon)} = Y - \overline{f(X)}$$

Therefore the estimated split of the observed response into its systematic and random components is as much a consequence of the choice of models, f and ζ, and the method of estimation as it is of the observed stimulus and response, X and Y.

11.2. General linear model

The general linear statistical model is a special simple case of the schematic statistical models discussed above. The so-called linear statistical model stipulates that the systematic component is a linear combination of the systematic factors or variables, and the random component is the identity function of random error.

- **random component:** $\zeta(\varepsilon) = \varepsilon$
- **systematic component**: $f(X) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$

Why use a Linear Systematic Function?

Linear systematic components have three fundamental properties which are desirable for statistical models – *simplicity, estimability* and *stability*.

Linear functions represent or give algebraic expression to the simplest kind of relationship. linear functions postulate either:

- stimulus and response tend to increase and decrease together
- response decreases as stimulus increases

For the simple linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

if $\beta_1 < 0$; the relation is negative => Y decreases as X increases
if $\beta_1 > 0$; the relation is positive => Y and X increase together

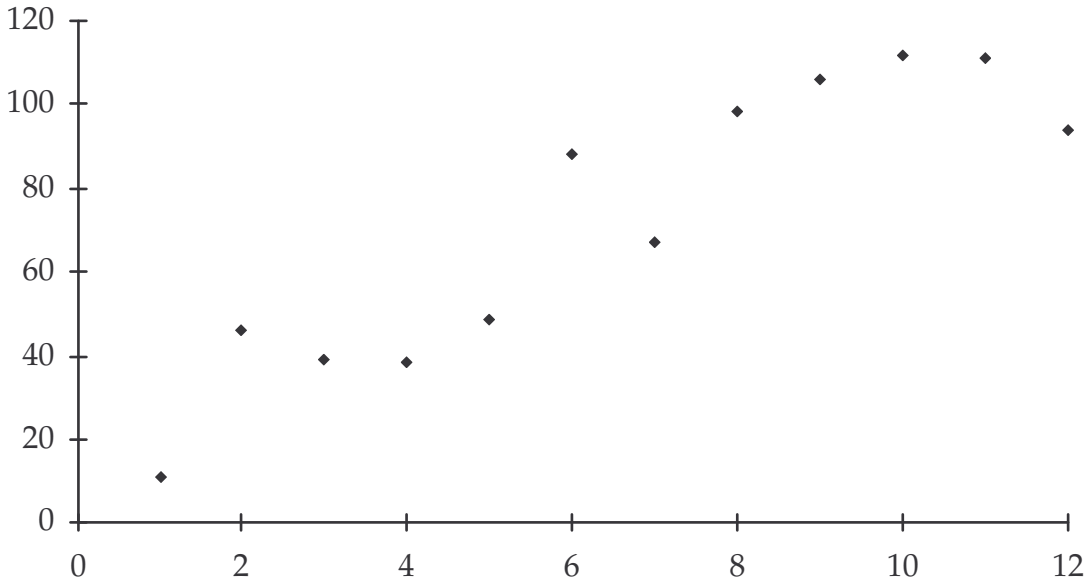
Assumptions about the random component

In decreasing order of impact on results and interpretation, the following three assumptions about the behavior of the random component of a linear statistical model are widely adopted.

1. *Independence* – the random errors ε_i and ε_j are independent for all pairs of observations i and j
2. *Equal Variance* – the random errors ε_i all have the same variance σ^2 for all observations
3. *Normality* – the random errors ε_i are normally distributed

11.3. Estimation of parameters for the simple linear regression model

The following scatter plot illustrates the type of data which is typically described by a simple linear model.



From the formulation of the general linear model, the special case of the simple linear model in which the systematic component is a linear function of a single variable, that is a straight line, may be expressed as:

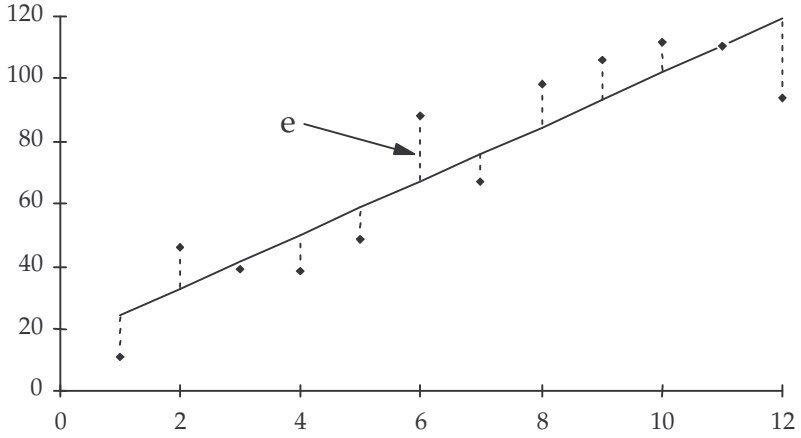
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\varepsilon_i \rightarrow N\left(0, \sigma^2\right)$$

and all ε_i are mutually independent.

For any estimates of the parameters, β_0 and β_1, say b_0 and b_1, the residual errors of estimation are:

$$e_i = Y_i - b_0 - b_1 X_i$$

as illustrated below.



The least squares parameter estimates are those values of b_0 and b_1 which minimize the sum of squared residual errors.

$$S(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

To find the parameter estimates which minimize the sum of squared residuals, we compute the derivatives with respect to b_0 and b_1 and equate them to zero.

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = \sum_{i=1}^{n} e_i = 0$$
$$\frac{\partial S(b_0, b_1)}{\partial b_0} = \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = \sum_{i=1}^{n} x_i e_i = 0$$

The solutions to the above equations are the least squares parameter estimates. Notice that the first equation insures that the residuals for the least squares estimates of β_0 and β_1 always sum to zero.
Because the least squares estimates are also maximum likelihood estimates under the assumption of normally distributed errors, they are usually denoted by the symbols β_0 and β_1. The solutions to the least squares equations are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = r_{xy} \frac{s_y}{s_x}$$

The intercept parameter, $\hat{\beta}_0$, merely places the vertical position of the line at the point where the residual errors sum to zero. The operative parameter is the slope estimate, $\hat{\beta}_1$, which has a particularly simple form in terms of the correlation and relative standard deviations of the response *Y* and the explanatory variable *X*.

$$\hat{\beta}_1 = \overbrace{r_{xy}}^{\substack{\text{Relation} \\ \text{Between} \\ X \text{ and } Y}} \quad \overbrace{\frac{s_y}{s_x}}^{\substack{\text{Scale} \\ \text{Factor}}}$$

The residual sum of squares for the simple regression model is

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = (1 - r_{xy}^2) \sum_{i=1}^{n} (y_i - \bar{y})^2 = (1 - r_{xy}^2)(n-1) s_y^2$$

which like the least squares slope estimate, $\hat{\beta}_1$, has a simple expression in terms of the correlation between *X* and *Y* and the variance of *Y*.

The residual sum of squares for a regression model measures how well the model fits the data. A smaller residual sum of squares indicates a better fit. Because a higher squared correlation between *X* and *Y* is associated with a smaller residual sum of squares as a proportion of the variance of *Y*, the squared correlation between *X* and *Y* is usually used as a measure of the goodness of fit of the regression model. When $r_{xy} = \pm 1$, the sample observations of *X* and *Y* all lie on a straight line and the residual sum of squares is zero. When $r_{xy} = 0$, *X* and *Y* are independent and the residual sum of squares will equal the sum of squared deviations of *Y* about its mean.

If the residual sum of squares measures the size of the random component of the regression model, then the remainder, the difference between the original sum of squared deviations of *Y* about its mean and the residual sum of squares of *Y* about the regression line must represent the systematic component of the model. To better understand what this systematic component measures, let the point on the regression line or predicted value of *Y* for the i^{th} observation of *X* be

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i .$$

Firstly, note that the least squares estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ insure that the mean of the predicted value of *Y* will always equal the mean of the original observations of *Y*. That is,

$$\frac{\sum_{i=1}^{n} \hat{y}_i}{n} = \frac{\sum_{i=1}^{n} \bar{y} - \hat{\beta}_1 (x_i - \bar{x})}{n} = \bar{y} .$$

Then as was the case in the analysis of variance, the total sum of squared deviations of *Y* from its mean

$$SS_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

can be partitioned into the sum of squared residual errors,

$$SS_{Error} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

and the sum of squared deviations of the predicted values of *Y* from their mean.

$$SS_{\mathrm{Re}\,gression} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

We see that

$$SS_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= SS_{\mathrm{Re}\,gression} + SS_{Error}$$

The sum of squares due to regression is often called the explained variation and conversely the sum of squared residual errors, the unexplained variation. The partitioning of the total variation of *Y* into these two components is due to the fact that the least squares estimates must satisfy the condition,

$$\sum_{i=1}^{n} x_i e_i = 0$$

That is, the residual errors must be orthogonal to the predictor variable.

The partitioning of the total sum of squared deviations of the response, *Y*, about its mean into the systematic component, explained variation, and the random component, sum of squared residuals is frequently presented as an Analysis of Variance table. The F-test computed by this ANOVA Table tests the null hypothesis that the systematic component of the model is zero.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F-ratio
Total	*n*-1	$(n-1)s_y^2$		
Regression	1	$r_{xy}^2(n-1)s_y^2$	$r_{xy}^2(n-1)s_y^2$	$\frac{(n-2)r_{xy}^2}{(1-r_{xy}^2)}$
Error	*n*-2	$(1-r_{xy}^2)(n-1)s_y^2$	$\frac{(1-r_{xy}^2)(n-1)s_y^2}{(n-2)}$	

Thus, the F-test for testing the significance of the regression model depends only on the correlation between response and explanatory variables and on the sample size. In practice, the null hypothesis of no regression effect is almost always rejected, but even if rejected does not imply that the regression model will provide satisfactory predictions.

As in the case of analysis of variance for factorial models, the estimate of the error variance, σ^2, is the mean squared error.

$$\hat{\sigma}^2 = SS_{Error}/n-2 = (1-r_{xy}^2)\left(\frac{n-1}{n-2}\right)s_y^2$$

This estimated error variance for the regression line is also called the conditional variance of *Y* given *X*, that is, the variance of *Y* remaining after the effect of *X* has been removed.

$$s_{y|x}^2 = \hat{\sigma}^2 = (1-r_{xy}^2)\left(\frac{n-1}{n-2}\right)s_y^2$$

A second consequence of least squares estimates of β_0 and β_1 is that the least squares line will always pass through the point of means $(\overline{X}, \overline{Y})$. In fact the z-value of the prediction for *Y* is simply the correlation between *X* and *Y* times the corresponding z-value for *X*. That is,

$$\hat{y}_i = \overline{y} - \left(r\frac{s_y}{s_x}\right)\overline{x} + \left(r\frac{s_y}{s_x}\right)x_i$$

$$\left(\hat{y}_i - \bar{y}\right) = \left(r\frac{s_y}{s_x}\right)\left(x_i - \bar{x}\right)$$

$$\left(\frac{\hat{y}_i - \bar{y}}{s_y}\right) = r\left(\frac{x_i - \bar{x}}{s_x}\right).$$

Clearly when $x_i = \bar{x}$, then $\hat{y}_i = \bar{y}$.

11.4. Distribution of least squares parameter estimates

If the predictor or explanatory variable *X* is assumed to be a fixed constant rather than a random variable, then both $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the normally distributed criterion or response variable, *Y*, and hence are normally distributed themselves. The mean and variance of the slope parameter estimate are

$$E\left[\hat{\beta}_1\right] = \beta_1$$
$$V\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\sigma^2}{(n-1)s_x^2}$$

These results can readily be established by noting that the least squares estimate of β_1 may be expressed as the following linear combination of the observations of *Y*.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)}{(n-1)s_x^2}\,y_i$$

Then the expected value of the slope parameter estimate is

$$E\left(\hat{\beta}_1\right) = \sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)}{(n-1)s_x^2}\,\mathrm{E}\left(y_i\right) = \sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)\left(\beta_0 + \beta_1 x_i\right)}{(n-1)s_x^2}$$

and the variance of the slope estimate is

$$V\left(\hat{\beta}_1\right) = \sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)}{(n-1)s_x^2}\,\mathrm{V}\left(y_i\right) = \sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)^2}{(n-1)^2 s_x^4}\,\sigma^2.$$

The significance of these results is that the least squares estimate of the slope parameter is unbiased and its variance becomes smaller as the sample size increases. In addition, the variance of the estimate becomes smaller when the variance or range of *X* becomes larger.

By substitution of the mean squared error estimate of σ^2 into the expression of the variance of the slope parameter estimate, the following estimated variance of the slope parameter is obtained

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{SS_{Error}}{(n-1)(n-2)s_x^2} = \frac{\left(1-r_{xy}^2\right)s_y^2}{(n-2)s_x^2}$$

Because $\hat{\beta}_1$ is unbiased, substitution into the least squares determining equation for $\hat{\beta}_0$ readily shows that the least squares intercept estimate, $\hat{\beta}_0$, is also unbiased.

$$E\left[\hat{\beta}_0\right] = \beta_0$$

The variance of $\hat{\beta}_0$ is obtained by again noting that from the least squares determining equation, $\hat{\beta}_0$ can be expressed as the following linear combination of the observations of *Y*.

$$\hat{\beta}_0 = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{\bar{x}\left(x_i-\bar{x}\right)}{(n-1)s_x^2}\right]y_i$$

By squaring and summing constant terms in this linear combination, the variance is found to be

$$V\left[\hat{\beta}_0\right] = \frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} = \sigma^2\left[\frac{1}{n} + \frac{n\,\bar{x}^2}{(n-1)s_x^2}\right]$$

This expression for the variance of the least squares estimate of the intercept consists of two parts, the reciprocal of the sample size, *n*, which is the usual factor for the variance of a mean, and the ratio of the square of the mean of *X* to its variance. As for $\hat{\beta}_1$, an estimate of the variance of $\hat{\beta}_0$ can be obtained by substituting the mean squared error estimate of σ^2.

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{\mathrm{SS_{Error}}}{\mathrm{n-2}}\left[\frac{1}{n} + \frac{n\,\bar{x}^2}{(n-1)s_x^2}\right] = \frac{\left(1-\mathrm{r_{xy}^2}\right)(n-1)s_y^2}{\mathrm{n-2}}\left[\frac{1}{n} + \frac{n\,\bar{x}^2}{(n-1)s_x^2}\right]$$

Since the predicted value, $\hat{y}_i$, is also a linear combination of the least squares parameter estimates, it too will be normally distributed.

$$\hat{y}_i = \sum_{k=1}^{n}\left[\frac{1}{n} + \frac{\left(x_i-\bar{x}\right)\left(x_k-\bar{x}\right)}{\left(n-1\right)s_x^2}\right]y_k$$

The expected value of $\hat{y}_i$ is obtained by direct substitution into the linear prediction equation.

$$E[\hat{y}_i] = E[\hat{\beta}_0] + E[\hat{\beta}_1] x_i = \beta_0 + \beta_1 x_i$$

As for $\hat{\beta}_0$ and $\hat{\beta}_1$, the variance of $\hat{y}_i$ is derived by squaring and summing the constant terms in the expression of $\hat{y}_i$ as a linear combination of the observations of *Y*.

$$V[\hat{y}_i] = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2} \right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2} \right)$$

Again notice that the expression for the variance of the predicted value for the i^{th} observation of *Y* consists of two components. The first component, the reciprocal of the sample size, is the usual factor for the variance of a mean. The second component is a normalized squared distance of x_i, the i^{th} observation of the explanatory variable, from its mean. Thus the variances of predictions of *Y* for values of x_i near its mean will be close to the variance of an ordinary sample mean. But for values of x_i far from its mean, the variances of the predictions will increase linearly with the squared normalized distance from the mean.

$$V[\hat{y}_i] = \sigma^2 \{ \overbrace{\frac{1}{n}}^{\text{Variance of Mean of } Y} + \overbrace{\frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}^{\text{Normed Distance From Mean of } X} \}$$

The foregoing expression for the variance of $\hat{y}_i$ is the variance of the estimate of the regression line, which is the conditional mean of *Y* given *X*. But the variance of a prediction for single observation of *Y* at *X* will be much greater. This prediction error will be the original variance of *Y*, σ^2, plus the variance of error due to estimation of the regression line. Therefore, the estimated variance of a single observation or prediction at X_i is

$$\hat{\sigma}^2_{\hat{y}_i} = \hat{\sigma}^2 \{ \overbrace{1}^{\text{Single Observation}} + \overbrace{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}^{\text{Regression Line}} \}$$

Both estimates of the regression line and predictions from the regression line will be more accurate for values of *X* near the mean.

11.5. Inference for the regression line

It is often of interest to test hypotheses about the parameters of the regression model or to construct confidence intervals for various quantities associated with the model. There may be theoretically prescribed values for the parameters. Confidence intervals for predictions from the regression model are frequently required. Inferential procedures follow in a natural way from the fact that least squares parameter estimates and hence the estimated regression line is all linear combination of the response variable, *Y*, and like *Y* will be normally distributed. In addition, the estimated variances of these parameters are derived from the sum of squared deviations of *Y* and hence will have a χ^2 distribution. Therefore, the following statistics all have Student's t distributions with (*n*-2) degrees of freedom.

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma} \Big/ \sqrt{(n-1)}s_x} \Rightarrow t_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\left[\frac{1}{n} + \frac{n\,\bar{x}^2}{(n-1)s_x^2}\right]}} \Rightarrow t_{n-2}$$

$$\frac{\hat{y}_i - \beta_0 - \beta_1 x_i}{\hat{\sigma}_{\hat{y}_i}} = \frac{y_i - \beta_0 - \beta_1 x_i}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}} \Rightarrow t_{n-2}$$

where β_0 and β_1 are the true or hypothesized values of the regression parameters. The most commonly tested null hypothesis is that the slope and intercept equal zero. This test is the t-test produced by most regression software. For the slope parameter, this test has a particularly interesting interpretation.

$$\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{r_{xy}\dfrac{s_y}{s_x}}{\sqrt{\dfrac{(1-r_{xy}^2)}{(n-2)}}\dfrac{s_y}{s_x}} = \frac{r_{xy}}{\sqrt{\dfrac{(1-r_{xy}^2)}{(n-2)}}}$$

which is simply the square root of the F-test for the regression model derived earlier. Thus, testing whether the systematic component is zero is equivalent to testing whether the slope of the regression line is zero. If $\beta_1 = 0$, then the regression line will be horizontal at the mean of *Y*, that is, the mean of *Y* will be predicted at every value of *X* and *X* will have no effect on predictions of *Y*.

Confidence intervals for the intercept, slope, regression line, and predictions from the regression line are calculated in the usual manner.

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \hat{\sigma} \Big/ \sqrt{(n-1)}s_x$$

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\left[\frac{1}{n} + \frac{n\,\bar{x}^2}{(n-1)s_x^2}\right]}$$

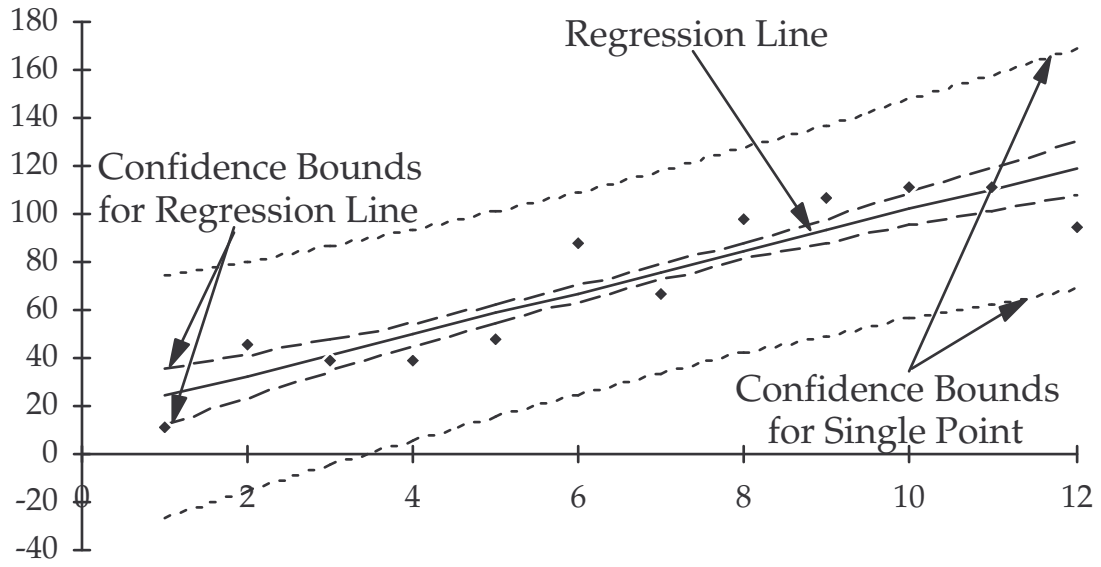
The confidence interval for the regression line is

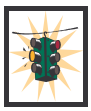
$$\hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}$$

And the confidence interval for predictions from the regression line is

$$\hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_i-\bar{x})^2}{(n-1)s_x^2}}$$

The following chart displays 95% confidence intervals for both the regression line and individual predictions from the regression line. Notice that the confidence bounds for the regression line are very narrow and include very few of the original data points. This is because the correlation between predictor and criterion variables is high and the fit of the regression line is good. On the other hand, all original observations are included within the confidence bounds for single points.





Solved example

The company repairs the desktop calculators and cashes. The data from 18 repairs are written in the table. Each repair has 2 important data. The former is a number of repaired calculators (X) and the latter is a total repair time (Y).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
x	7	6	5	1	5	4	7	3	4	2	8	5	2	5	7	1	4	5
Y	97	86	78	10	75	62	101	39	53	33	118	65	25	71	105	17	49	68

a) Find parameter estimates of the regression line.
b) Draw data and regression function.
c) Use t-tests for the values of all parameters of regression function.

Solution

– we can use STATGRAPHIC software:

Linear regression – Repair time vs. Number

Regression Analysis - Linear model: Y = b_0 + b_1 *x

Dependent variable: Repair Time
Independent variable: Number

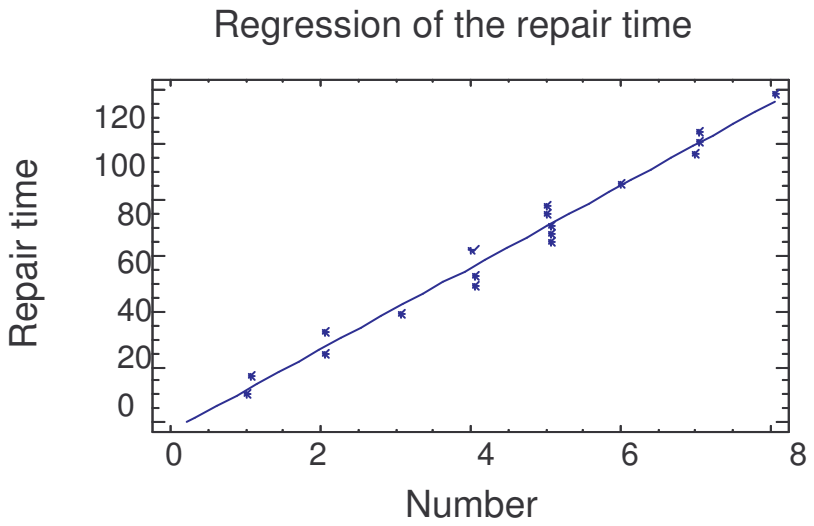
Parameter	Estimate	Standard Error	T Statistic	P-Value
b_0 - Intercept	-2,32215	2,56435	-0,905549	0,3786
b_1 - Slope	14,7383	0,519257	28,3834	0,0000

---

b_0 = *Intercept*, b_1 = *Slope*, the results of these values may be found in the second column. The following function introduces an equation for the estimate of predicted value:

Repair Time = -2,32215 + 14,7383 . Number

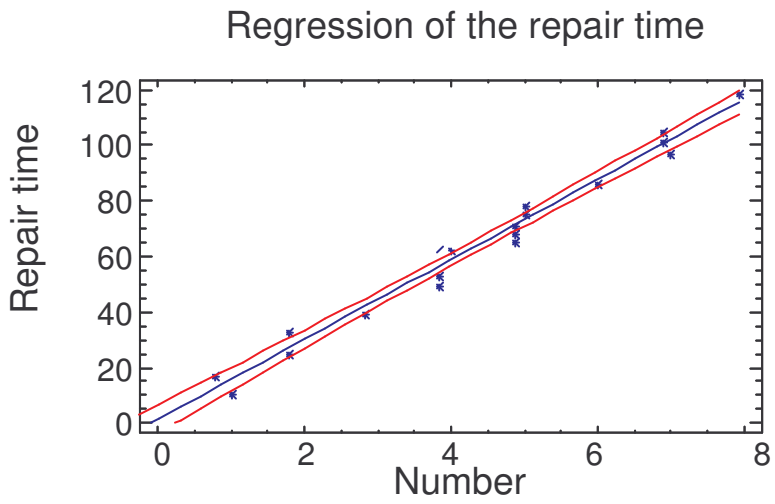
The observed values of the t-tests are shown in the fourth column (T Statistic) and corresponding p-values are displayed in the last one. It is obvious that hypothesis H_0: $\beta_0=0$ will not be rejected considering the important value in p-value column. Based on this, we can say that regression line passes through the beginning what is a logical conclusion, considering the data nature. The second of particular test says that Slope is a value that significantly differs from zero since we have rejected H_0 hypothesis H_0: $\beta_1=0$.



d) Let's find the 95% confidential interval for the repair time in dependence on the number of calculators.

e) Let's find point and interval estimation for an expected repair time for 5 calculators.

Solution



For value $x=5$:

$$\hat{Y}(x) = b_0 + b_1 x = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x}) \cdot (x - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \cdot Y_i = 71.3691$$

$$\mathrm{E}(\mathrm{Y}|\,\mathrm{x}) = \beta_0 + \beta_1 x \in \langle \hat{Y}(x) - S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}}(n-2), \hat{Y}(x) + S_{\hat{Y}} \cdot t_{1-\frac{\alpha}{2}}(n-2) \rangle = \langle 69.063,\ 73.6752 \rangle$$

f) Consider the quality of examined model of linear regression for the repair time in dependence on a number of calculators using a coefficient of determination

Solution

$$SS_Y = SS_{\hat{Y}} + SS_R$$

Source		Sum of Squares
Regression	$SS_{\hat{Y}}$	16182,6
Error	SS_R	321,396
Total	SS_Y	16504,0

$$I^2 = \frac{SS_{\hat{Y}}}{SS_Y} = 98.0526\ \%$$

Summary of notion

Regression model is a special case of general linear model. The basic assumptions are independence, homoscedasticity and normality.

Dependent variable is the variable of a regression model that is random and we try explaining its behavior and describing by mathematical curve.

Independent (explanatory) variables are the variables in the regression model whose behavior explains the behavior of the dependent variable.

Linear regression model with one explanatory variable is a basic model and it is based on the **Least-Squares Method**. By this method we can determine model parameters. The sum of squared deviations of the real values from modeled values is called the residual sum of squares.

We can obtain interval estimation for the expected value of the dependent variable. These interval bounds form **confidential interval** of the regression line.

Question

1. Describe and explain equation of linear regression.
2. What means p-value in the ANOVA table for linear regression?
3. What property describes a coefficient of determination?

Problems

Example 1: During control measurements of industrial components size we randomly chosen 8 components showing mostly positive divergences from normal values in the length and height*:*

length divergence [mm]	3	4	4	5	8	10	6	3
height divergence [mm]	4	6	5	6	7	13	9	4

Let's find the linear regression model of dependency between the length divergence and height divergence.

{*Answer:* Use a suitable software package.}

Example 2: In the years 1931-1961, water flow in profile of Šance and Morávka water reservoirs were measured. Averages per year (m^3/s) are given by the following table:

year	Šance	Morávka
1931	4,130	2,476
1932	2,386	1,352
1933	2,576	1,238
1934	2,466	1,725
1935	3,576	1,820
1936	2,822	1,913
1937	3,863	2,354
1938	3,706	2,268
1939	3,710	2,534
1940	4,049	2,308
1941	4,466	2,517
1942	2,584	1,726
1943	2,318	1,631
1944	3,721	2,028
1945	3,290	2,423

year	Šance	Morávka
1946	2,608	1,374
1947	2,045	1,194
1948	3,543	1,799
1949	4,055	2,402
1950	2,224	1,019
1951	2,740	1,552
1952	3,792	1,929
1953	3,087	1,488
1954	1,677	0,803
1955	2,862	1,878
1956	3,802	1,241
1957	2,509	1,165
1958	3,656	1,872
1959	2,447	1,381
1960	2,717	1,679

Let's assume that in one of following years, the average value of whole year water flow of Morávka reservoir is missing. In this year, the average water flow for Šance reservoir was 2,910 m^3/s. Based on linear regression, try to determine the average water flow in Morávka reservoir.

{*Answer:* Use a suitable software package.}

12. KEYS TO SOLUTION

12.1. Lecture 1

1. Exploratory data analysis is often a first step in revealing information hidden in large amount variables and their variants.
2. The base kinds of variables are a quantitative variable (nominal, ordinal) and a qualitative variable (discrete, continuous).
3. The frequency table is concerning with absolute and relative frequencies (for a qualitative variable).
4. The outliers are the variable values which significantly differ from other values.
5. b)
6. qualitative variable – bar chart, pie chart
   quantitative variable – box plot, stem and leaf plot
7. b) 14 thousand, d) cca (9;19)

12.2. Lecture 2

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2. $P(A \cap B) = P(A) \cdot P(B)$
3. Two events are independent if intersection probability of these two events is equal to a product of individual event probabilities.

12.3. Lecture 3+4

1. $F(x) = \sum_{x_i < x} P(X = x_i)$
2. $F(x) = \int_{-\infty}^{x} f(t)dt$ for $-\infty < x < \infty$
3. 50% quantile is called a median

   A mode is a value in which the discrete RV comes with the biggest probability.
4. The conditional distribution is the distribution of one variable at a fixed value of the other jointly distributed random variable.
5. $X_1 \ldots X_n$ are mutually independent $\Leftrightarrow$ $F(x_1,\ldots,x_n) = F_1(x_1) \ldots\ldots F_n(x_n)$.
6. The correlation coefficient measures the strength of the relation between two random variables.

12.4. Lecture 5

1. discrete distribution – binomial, geometric, negative binomial
   continuous distribution – poisson, exponential, Weibull, Gamma
2. A sequence of Bernoulli trials is defined as a sequence of random events which are mutually independent and which have only two possible outcomes and the probability

of event occur p is constant in any trial. On the basis of these trials expectations we can define the following random variables: binomial, geometric and negative binomial.
mean of the binomial random variable: EX=np

3. A Gamma distribution describes a time to k-th event occurrence in a Poisson process
4. β=2

12.5. Lecture 6

1. X ... RV with $N(\mu,\sigma^2)$ => $Z = \frac{X-\mu}{\sigma}$ ... N(0,1)
2. Chebyshev's inequality puts an upper bound on the probability that an observation should be far from its mean.
3. The law of large number is a theorem about convergence of means in the sequence of the random variables.
4. Chí-square distribution is a distribution derived from sum of squared standard normal random variables.

12.6. Lecture 7

1. Inferential statistics or statistical induction comprises the use of statistics to make inferences concerning some unknown aspect of a population.
2. A random sample is a set of items that have been drawn from a population in such a way that each time an item was selected, every item in the population had an equal opportunity to appear in the sample.

12.7. Lecture 8

1. The p-value calculation depends on defined null hypothesis:

a) $H0: \mu < \mu_0$ => p-value=F(x_{obs})

b) $H0: \mu > \mu_0$ => p-value=1-F(x_{obs})

c) $H0: \mu = \mu_0$ => p-value=2{F(x_{obs}), 1-F(x_{obs})}

2. It is a hypothesis that is accepted in case the rejection of null hypothesis.
3. $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\cdot(1-\hat{p})}\cdot\left(\frac{1}{n_1}-\frac{1}{n_2}\right)}$ ... N(0,1) , where $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$

12.8. Lecture 9

1. An $\hat{\theta}$ estimation is consistent if

   a) $\hat{\theta}$ is asymptotically unbiassed, $E\hat{\theta} \to \theta$

   b) $\lim_{n \to \infty} D\hat{\theta} = 0$

2. $P(T_D(\overline{X}) \le \theta \le T_H(\overline{X})) \ge 1-\alpha$

12.9. Lecture 10

1. $F = \frac{S_B^2}{S_W^2}$ ... F-distribution with *(k*-1*)* and *(N-k)* degrees of freedom
2. ANOVA table
3. The post hoc analysis is a second step of ANOVA and consists of comparing the means of all pairs of groups of purpose to choose homogenous groups.

12.10. Lecture 11

1. y=a+b*x, where y is a dependent variable and x is an independent variable. The values a (intercept) and b (slope) are estimates of regression line parameters
2. x and y are independent variables, in the case that p-value > 0,05
3. Coefficient of determination predicate about suitability of used model