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.

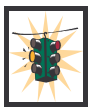
Explication

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.

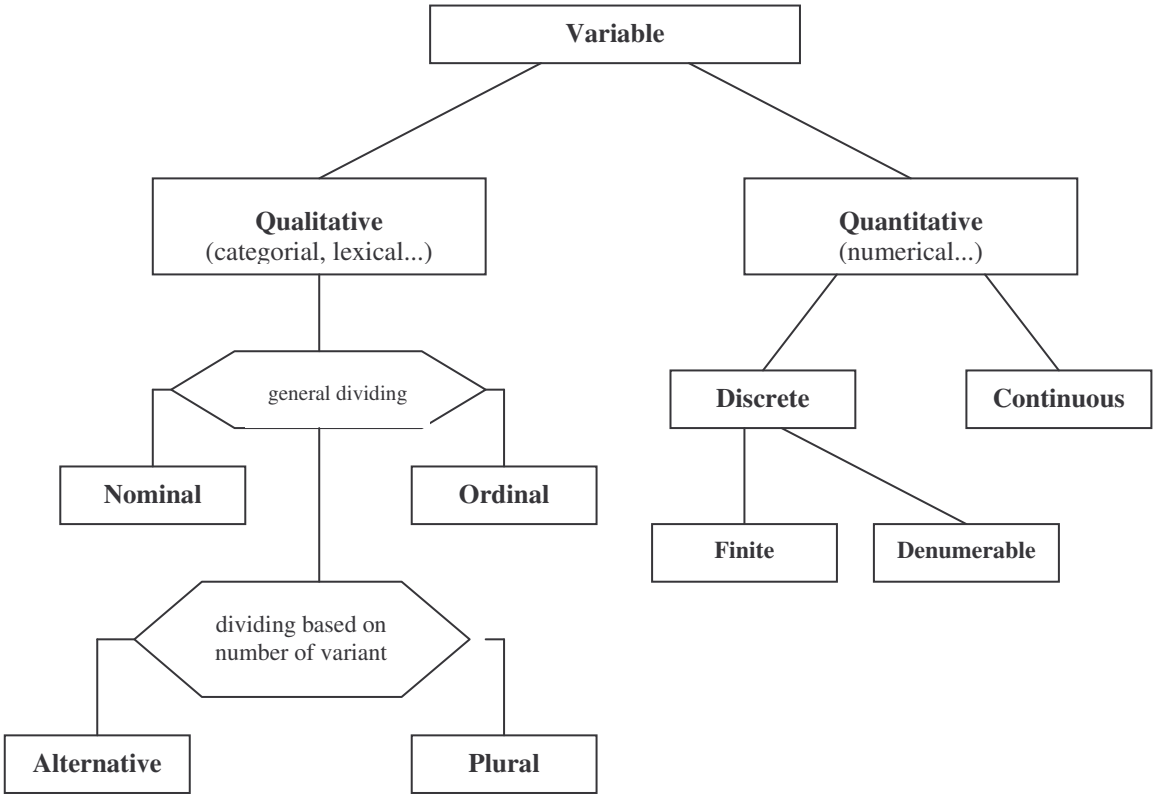
There are several ways how to make this selection. To avoid the omittance of some elements of population we choose so called **random sample** in which each element of population has a same chance to be selected.

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.

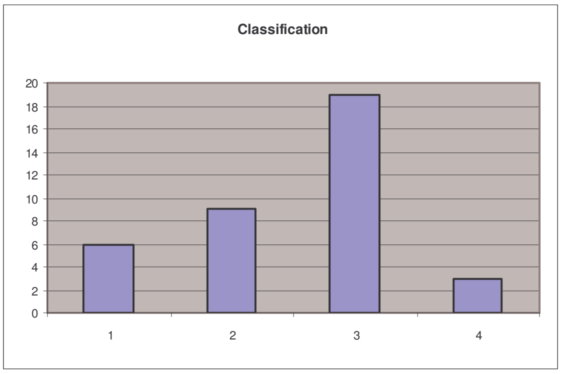
1.1.2 Graphical presentation qualitative variables

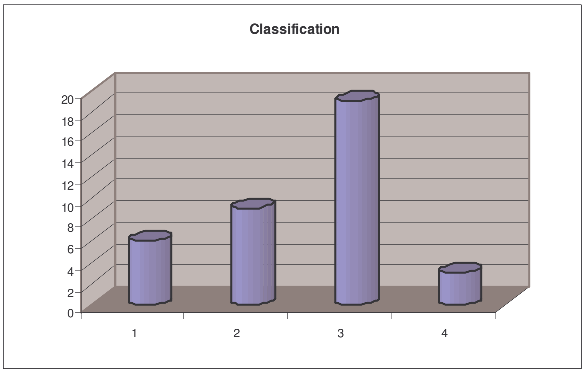
The statistics often use **graphs** for better plasticity of variables analysis. They are these two types for nominal variable:

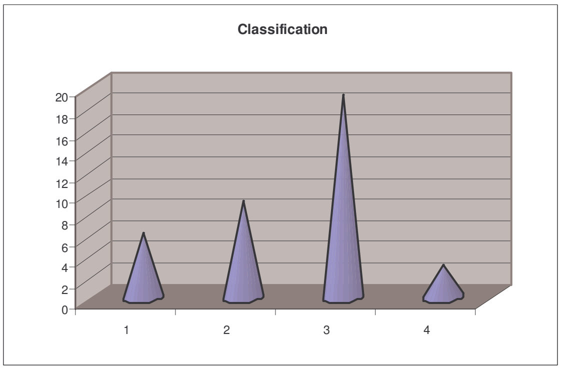
- **Histogram** (bar chart)
- **Pie chart**

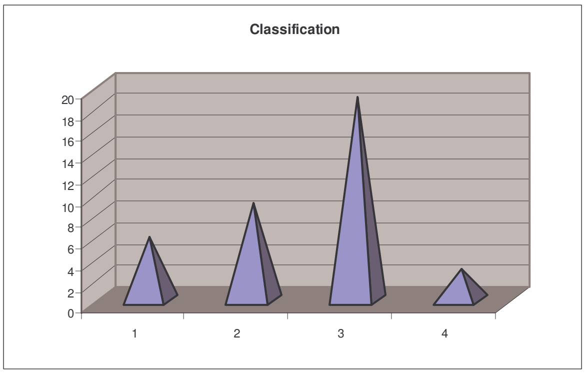
Histogram is a classical graph whereof we take variants of the variable on one axis and variable frequencies on the second one. Individual values of the frequency are then displayed as bars (boxes or vectors, squared logs, cones ...)

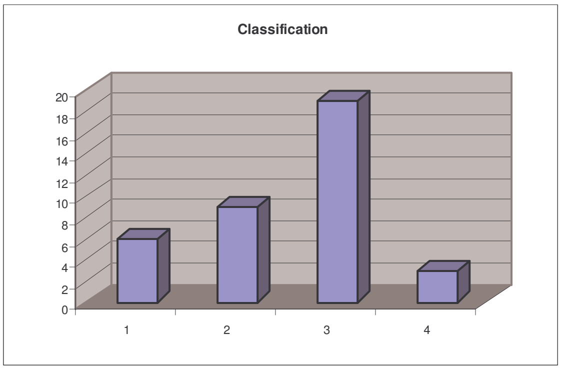
Examples:



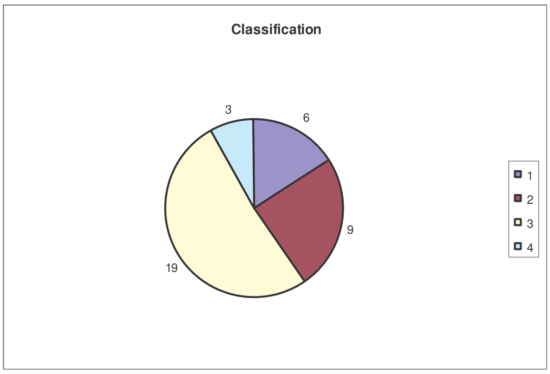


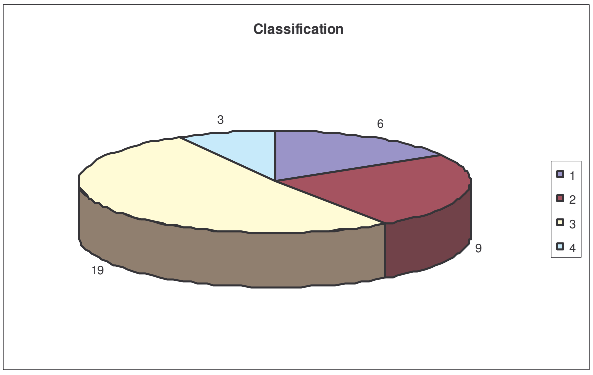


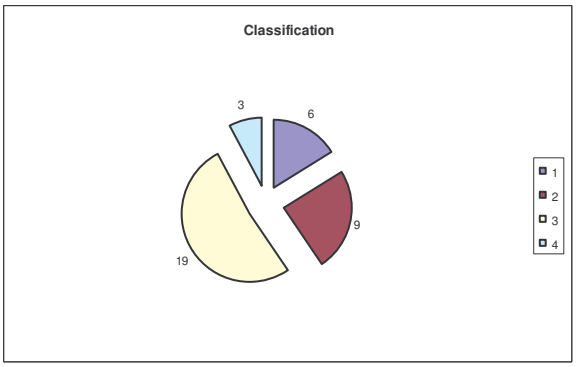


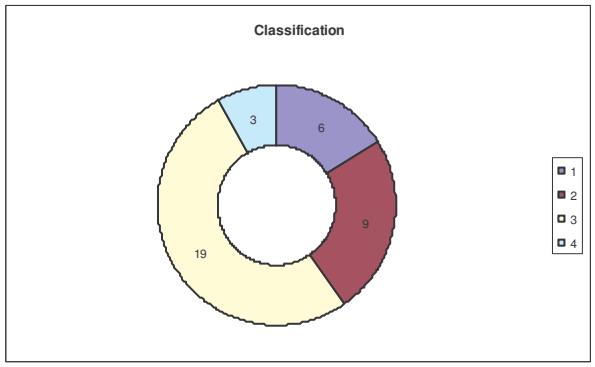


Pie graph represents relative frequencies of the individual variants of the variable. Individual relative frequencies are proportionally represent as a sector of a circle (when we change a circle to an ellipse we obtain three-dimensional effect).



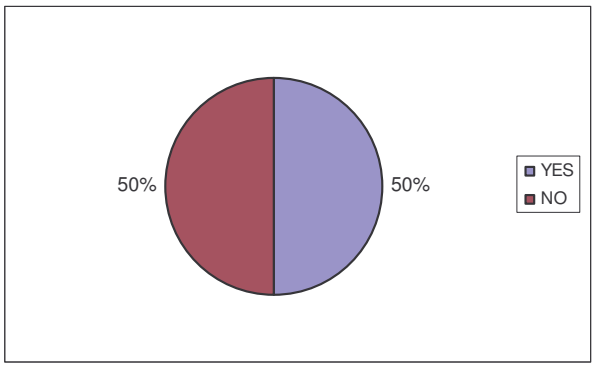




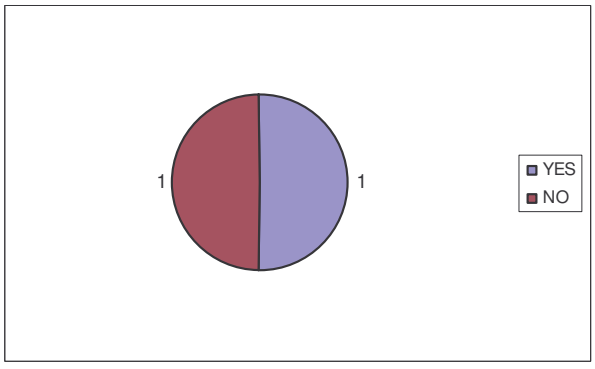


ATTENTION!!! There we must take care of graph description in the pie charts. Marking individual sectors by relative frequencies without adding their absolute frequencies is not sufficient.

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

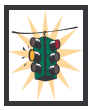


These are interesting results, aren't they? But they are true. Now we modify the chart the way it was recommended:



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

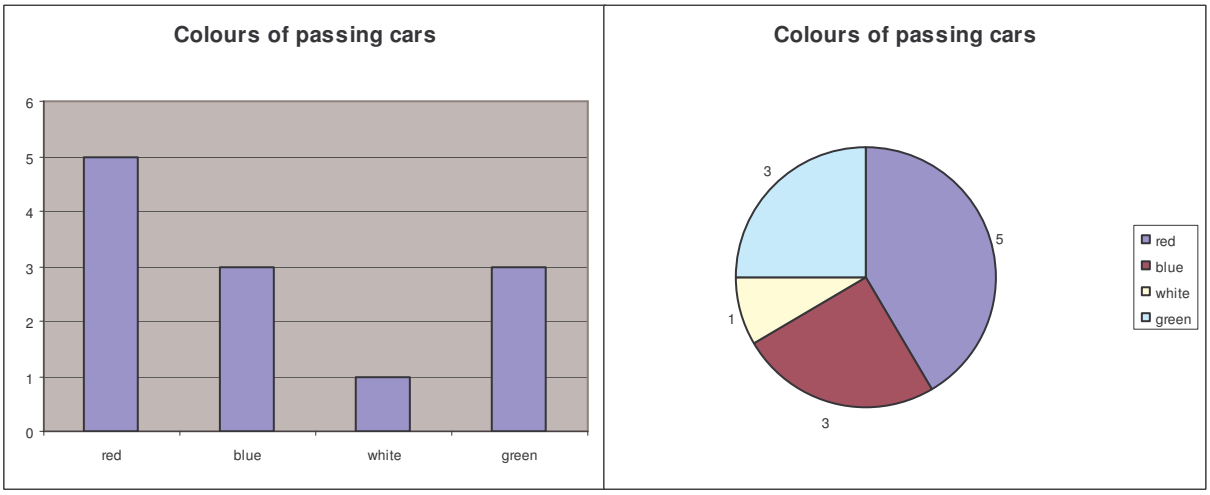
Solution:
It's obvious that it's a qualitative (lexical) variable and considering the fact that there is no point in ordering or comparing colors of cars we can say it's a nominal variable.

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:

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

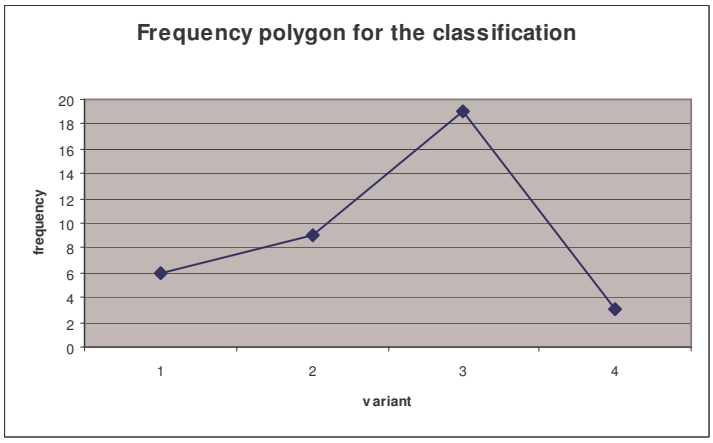
As well as at nominal variable we can present statical characteristics using frequency table at ordinal variable. It contains comparing with frequency table of nominal variable also values of cumulative and cumulative relative frequencies.

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	$\sum_{i=1}^{k} n_i = n$	-----	$\sum_{i=1}^{k} p_i = 1$	-----

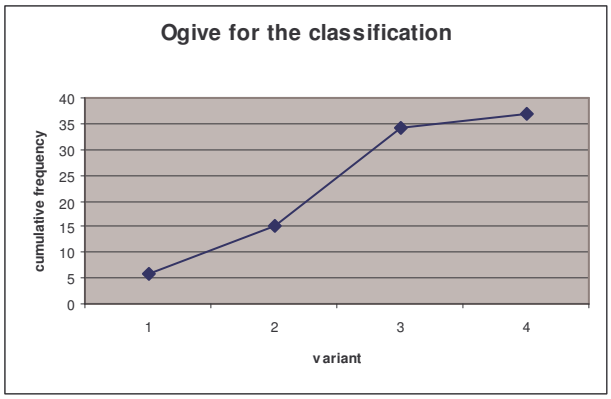
1.1.4 Graphical presentation ordinal variables

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.

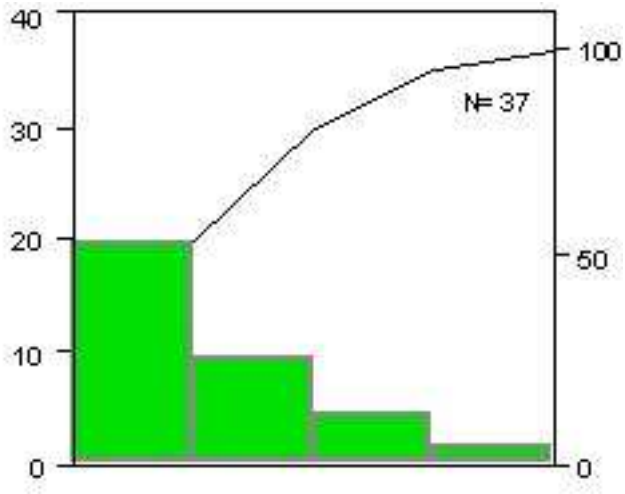


Ogive (cumulative frequency polygon)
- it's a frequency polygon of the cumulative frequency or the relative cumulative frequency. The vertical axis is the cumulative frequency or relative cumulative frequency. The horizontal axis represents possible variants. The graph always starts at zero at the lowest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

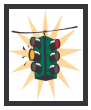


Pareto graph

- 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.

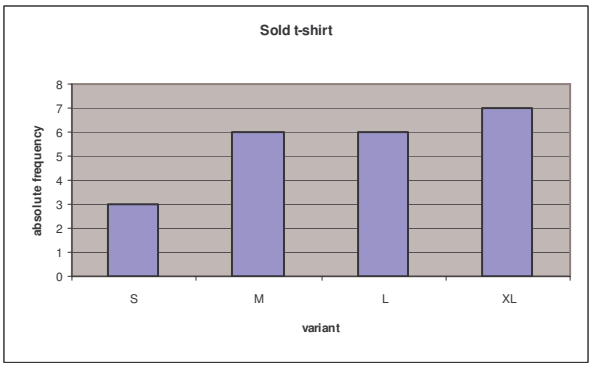
FREQUENCY TABLE				
T-shirt size	**Absolute frequency**	**Cumulative frequency**	**Relative frequency**	**Relative cumulative frequency**
	n_i	m_i	p_i	F_i
S	3	3	$3/22 = 0{,}14$	$3/22 = 0{,}14$
M	6	$3+6=9$	$6/22 = 0{,}27$	$9/22 = 0{,}41$
L	6	$9+6=15$	$6/22 = 0{,}27$	$15/22 = 0{,}68$
XL	7	$15+7=22$	$7/22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

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

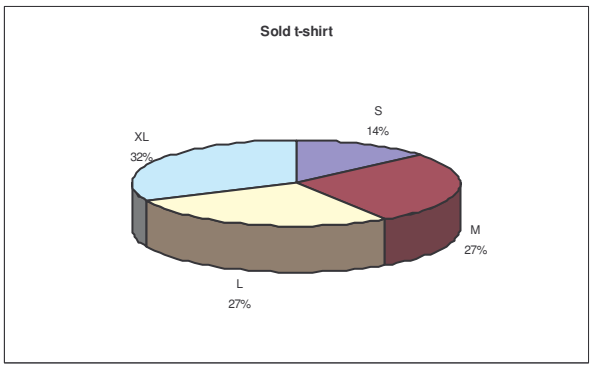
Graphical output will be histogram, pie graph and cumulative frequency polygon (we don't create pareto graph because we haven't got a technical data).

Graphical output:

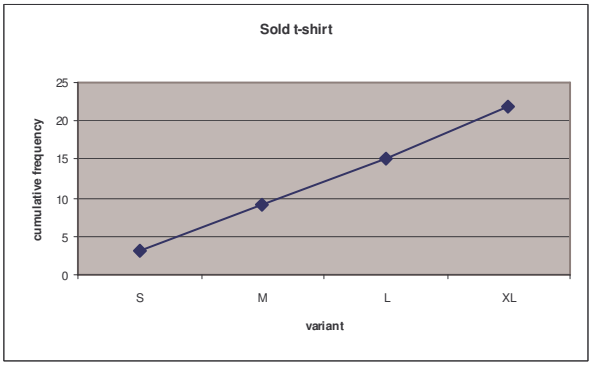
histogram



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).

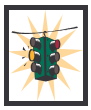
  But by the **continuous variable** we think of mode $\hat{x}$ as value around which is the most concentration of variable values.

  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**.

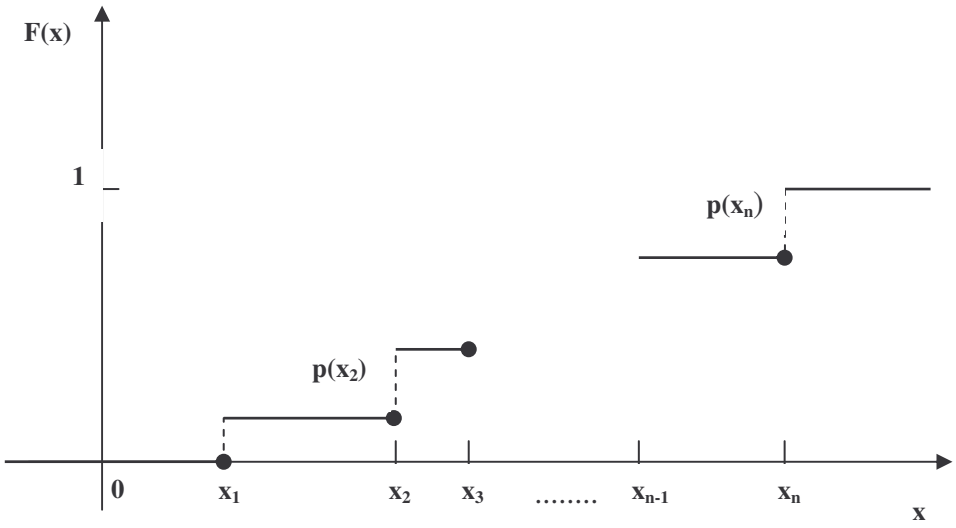
- **Empirical distribution function F(x) for the quantitative variable**

  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.

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

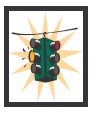


- **MAD**

  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**

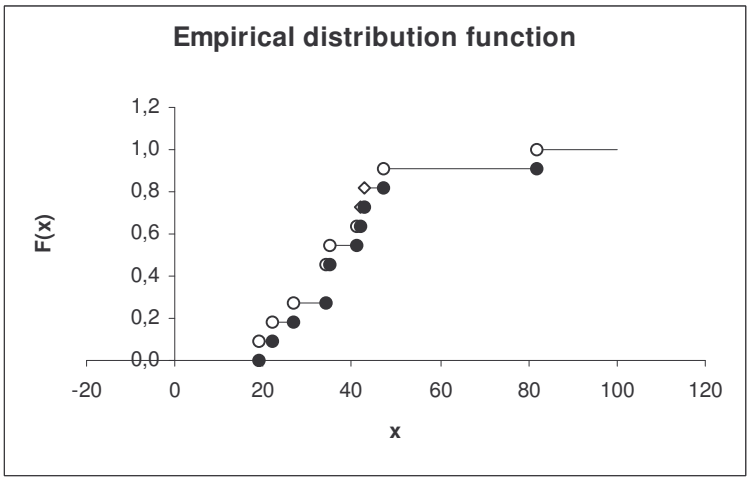
  - it is the most frequently measure of variability

  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

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

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

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

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

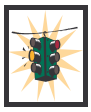
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

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) [°C]	
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:

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

Mean for the new technology:

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

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

- now we determine measures of variability

The old technology:

Sample variance:

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

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:

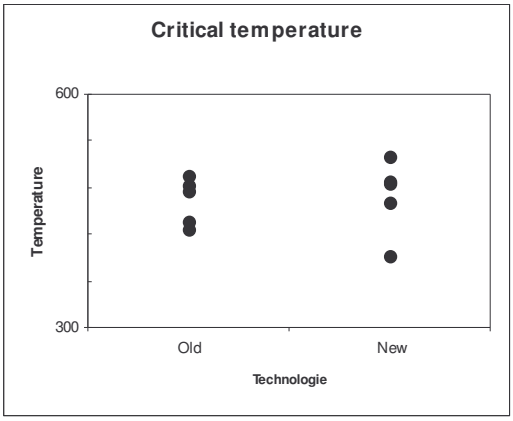
Sample variance:

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

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:

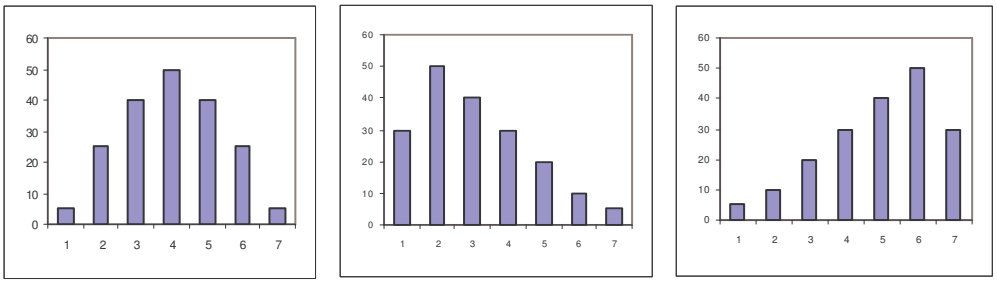
$$\alpha = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

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:

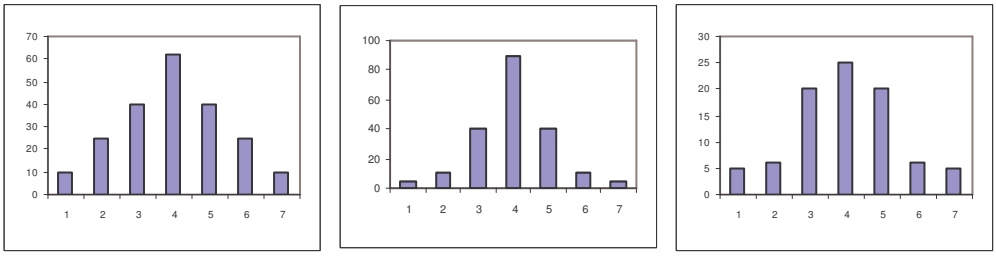
$$\beta = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

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

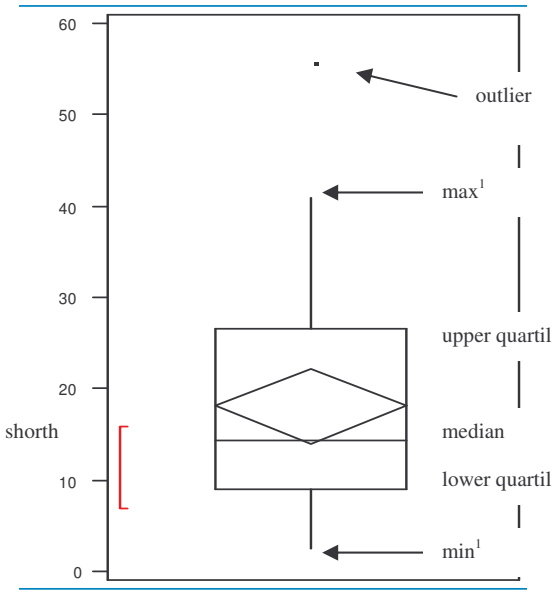
Now we have defined all numerical characteristics for description of the quantitative variable. We have show left how we can graphically represent quantitative variable.

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

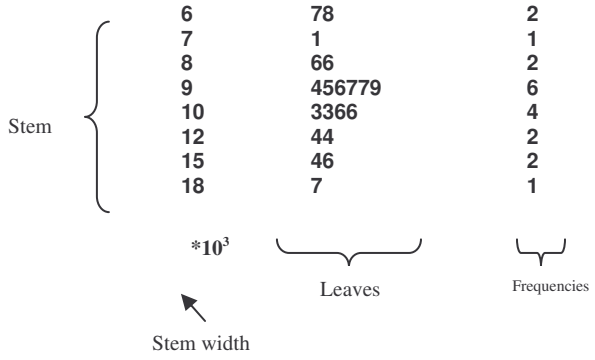
We saw it that simplicity is an advantage of box plot. But information's about concrete values of variable are missing us sometimes. We would digestedly inscribe the numeric values. To it we use stem and leaf plot.

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.

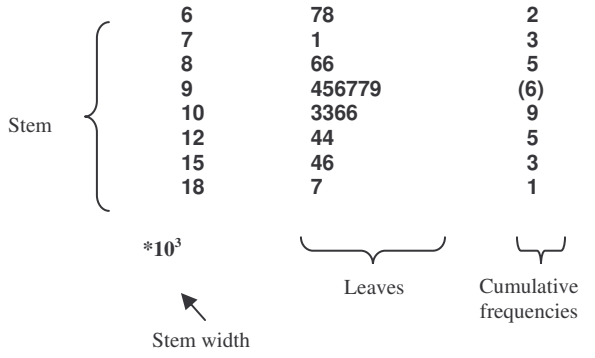


The second column in the graph -**leaves** - is numbers which represent "important" place. These numbers we write in appropriate rows.

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.

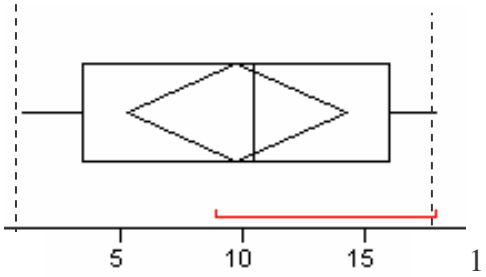
3. Which statistical characteristics can be contained in frequency table (for what type of variable)?

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.