Categorical data

Also called

- discrete data
- frequency data
- qualitative data
- data on nominal or ordinal scale

as opposed to

- quantitative data
- numerical data
- continous data
- data on interval or ratio scale

Marital status: 1 – single, 2 – married, 3 – divorced, 4 – widowed, 5 – cohabitating

(there is no natural ordering of the categories: nominal data)

Marital status: 1 – single, 2 – married, 3 – divorced, 4 – widowed, 5 – cohabitating

(there is no natural ordering of the categories: nominal data)

Alcohol consumption: 0 – never, 1 – occasionally (1-3 times a year), 2 – often (weekly), 3 – regularly (almost every day)

(categories have a natural ordering: ordinal data)

Marital status: 1 – single, 2 – married, 3 – divorced, 4 – widowed, 5 – cohabitating

(there is no natural ordering of the categories: nominal data)

Alcohol consumption: 0 – never, 1 – occasionally (1-3 times a year), 2 – often (weekly), 3 – regularly (almost every day)

(categories have a natural ordering: ordinal data)

Presence of a symptom: 0 – no, 1 – yes

(just two categories: binary or dichotomous data)

Marital status: 1 – single, 2 – married, 3 – divorced, 4 – widowed, 5 – cohabitating

(there is no natural ordering of the categories: nominal data)

Alcohol consumption: 0 – never, 1 – occasionally (1-3 times a year), 2 – often (weekly), 3 – regularly (almost every day)

(categories have a natural ordering: ordinal data)

Presence of a symptom: 0 – no, 1 – yes

(just two categories: binary or dichotomous data)

Categories are coded: codes can even be letters or text.

Don't calculate statistics such as average, median, SD, etc. from the codes, even if they are numbers!

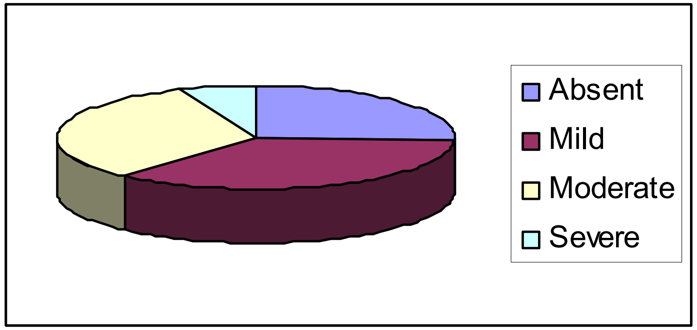
Analysis of a single categorical variable

- Frequency table
- Mode (=the most likely category)
- Barchart
- Pie chart

Severity of symptoms in a sample of 220 patients:

Severity	Absent	Mild	Moderate	Severe
Frequency	56	79	71	14
%	25.5	35.9	32.3	6.4

Mode (or modal category)



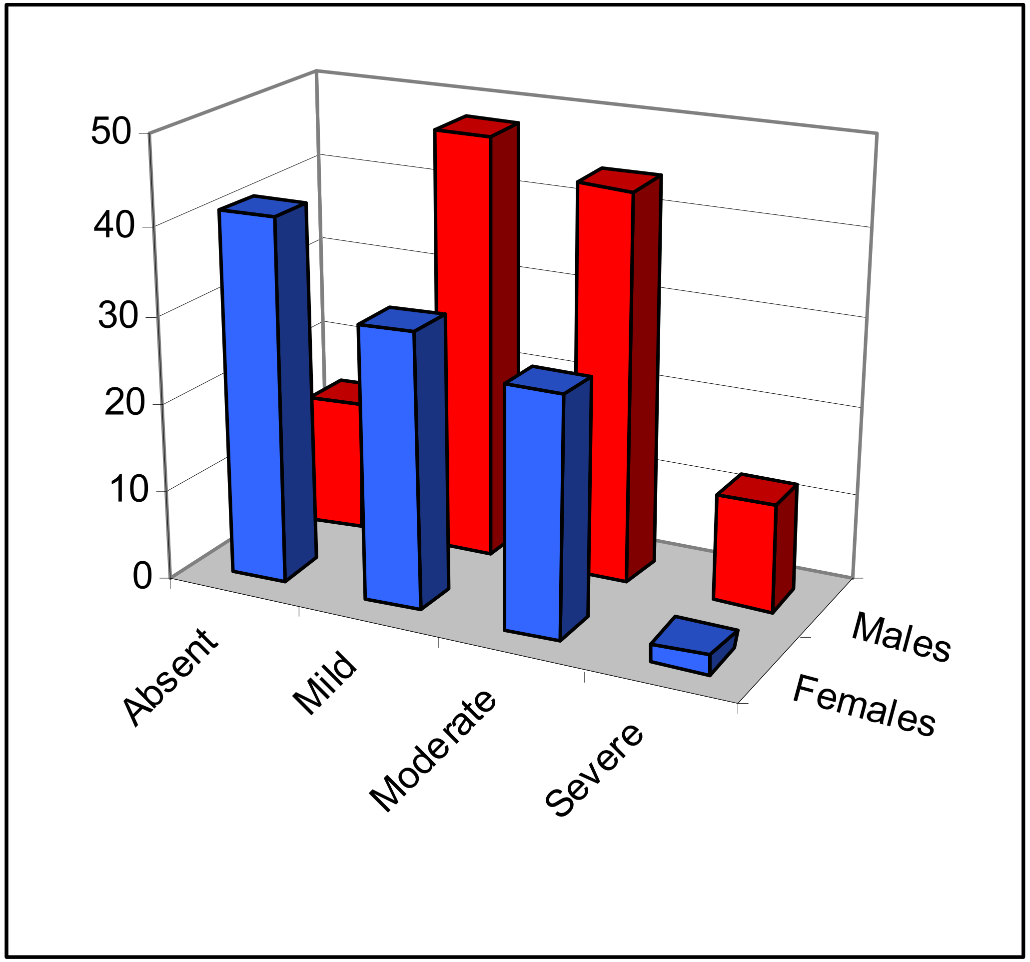
Analysis of the relationship between two categorical variables:

- Contingency table (=two-dimensional frequency table)
- Three-dimensional barchart
- Association measures
- Tests of independence

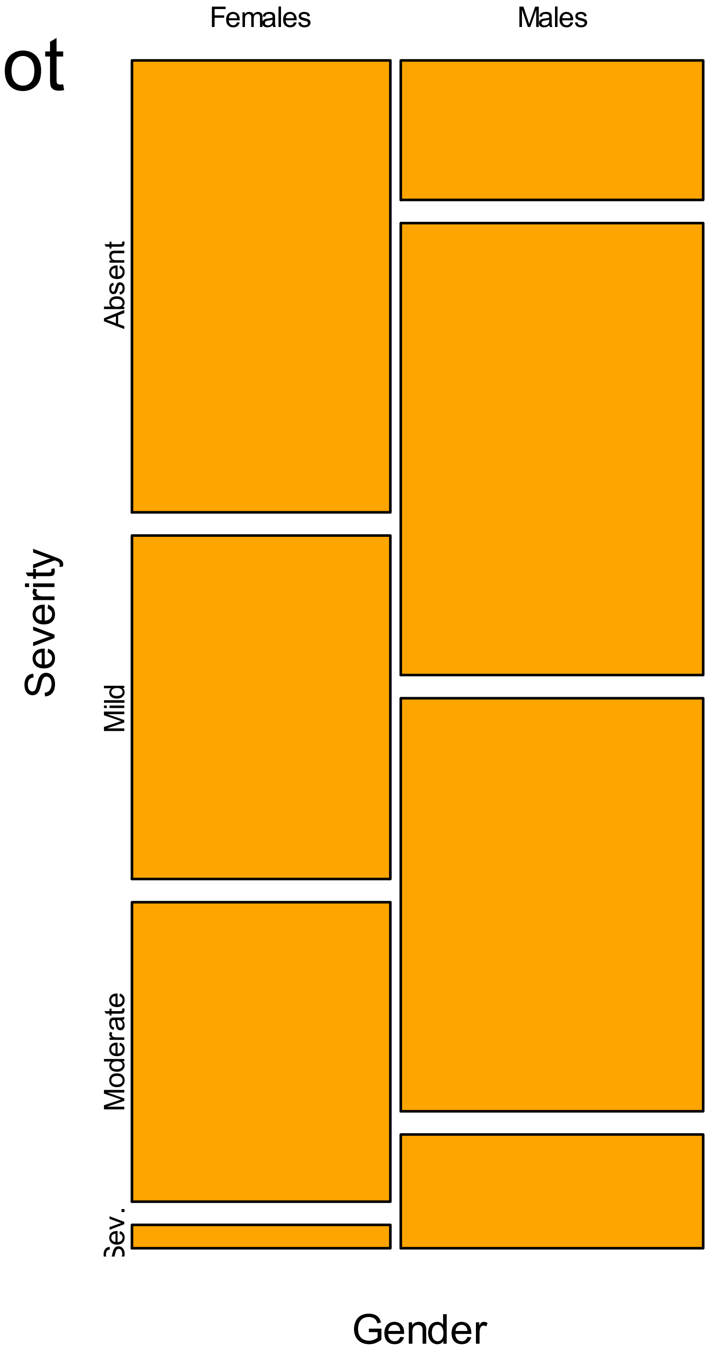
Severity of symptoms by gender:

	Absent	Mild	Moderate	Severe	Total
Females	41	31	27	2	101
Males	15	48	44	12	119
Total	56	79	71	14	220

3-D barchart



Mosaic plot



Association

Does *X* contain any information about *Y*?

Does gender in the above example provide any information about the severity of symptoms?

Association

Does *X* contain any information about *Y*?

Does gender in the above example provide any information about the severity of symptoms?

- If not, then we say *X* and *Y* are ***independent***

  If we are interested in *Y*, it is a waste of time and energy to measure *X* because it tells nothing about *Y*

Association

Does *X* contain any information about *Y*?

Does gender in the above example provide any information about the severity of symptoms?

- If not, then we say *X* and *Y* are ***independent***

  If we are interested in *Y*, it is a waste of time and energy to measure *X* because it tells nothing about *Y*

- If yes, then we say there is an ***association*** between *X* and *Y*

  Extreme case: *X* fully determines Y. If we measure *X*, we don't need to measure *Y* at all.

Association

Does *X* contain any information about *Y*?

Does gender in the above example provide any information about the severity of symptoms?

- If not, then we say *X* and *Y* are ***independent***

  If we are interested in *Y*, it is a waste of time and energy to measure *X* because it tells nothing about *Y*

- If yes, then we say there is an ***association*** between *X* and *Y*

  Extreme case: *X* fully determines Y. If we measure *X*, we don't need to measure *Y* at all.

I don't mean causally!

Please don't call it correlation!!!

Measures of association

…quantify how strong an association exists between *X* and *Y*.

The traditional setting (there are other variants as well):

0 — no association — *X* and *Y* are independent

1 — complete association — *X* fully determines *Y*

Measures of association

…quantify how strong an association exists between *X* and *Y*.

The traditional setting (there are other variants as well):

0 — no association: *X* and *Y* are independent

1 — complete association: *X* fully determines *Y*

Most frequently used measures of association:

- Cramer's V
- Goodman and Kruskal's lambda

Good measures of association are invariant to changing the codes and/or the order of categories!

Correlation

Correlation is a relationship of special kind, which is meaningful only **for variables with a natural ordering of their categories**.

It is a **monotonic relationship** between *X* and *Y*, which can be **positive or negative**.

Correlation

Correlation is a relationship of special kind, which is meaningful only **for variables with a natural ordering of their categories**.

It is a **monotonic relationship** between *X* and *Y*, which can be **positive or negative**.

- Positive correlation: the more (better, higher, etc.) the *X*, the more (better, higher, etc.) the *Y*.

Correlation

Correlation is a relationship of special kind, which is meaningful only **for variables with a natural ordering of their categories**.

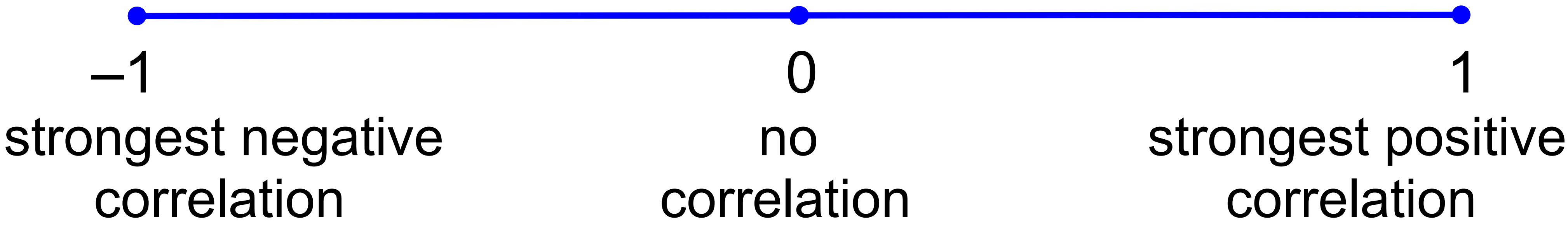
It is a **monotonic relationship** between *X* and *Y*, which can be **positive or negative**.

- Positive correlation: the more (better, higher, etc.) the *X*, the more (better, higher, etc.) the *Y*.
- Negative correlation: the more (better, higher, etc.) the *X*, the less (worse, lower, etc.) the *Y*.

Correlation coefficients

…quantify how strong a correlation exists between *X* and *Y*.

The traditional setting:



Correlation coefficients

…quantify how strong a correlation exists between *X* and *Y*.

The traditional setting:

–1	0	1
strongest negative correlation	no correlation	strongest positive correlation

Ordering of the subjects according to their *X* values is fully identical with their ordering according to their *Y* values

Correlation coefficients

…quantify how strong a correlation exists between *X* and *Y*.

The traditional setting:

–1 strongest negative correlation

0 no correlation

1 strongest positive correlation

Ordering of the subjects according to their *X* values is exactly the inverse of their ordering according to their *Y* values

Ordering of the subjects according to their *X* values is fully identical with their ordering according to their *Y* values

Correlation coefficients

…quantify how strong a correlation exists between *X* and *Y*.

The traditional setting:

–1	0	1
strongest negative correlation	no correlation	strongest positive correlation
Ordering of the subjects according to their *X* values is exactly the inverse of their ordering according to their *Y* values	Independence of *X* and *Y* implies zero correlation (but zero correlation does not imply independence)	Ordering of the subjects according to their *X* values is fully identical with their ordering according to their *Y* values

Correlation coefficients applicable to categorical data:

- Kendall's tau
- Spearman's rho

Don't use Pearson's correlation coefficient with categorical data! It treats the codes as if they were numbers and assumes a linear relationship!

Correlation coefficients applicable to categorical data:

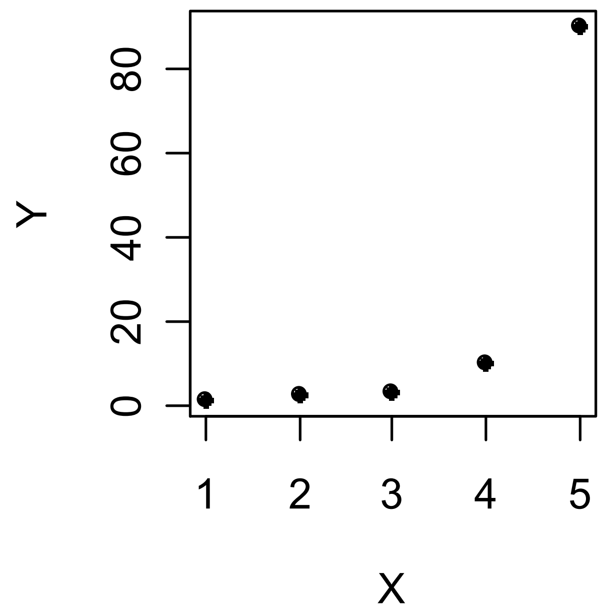
- Kendall's tau
- Spearman's rho

Don't use Pearson's correlation coefficient with categorical data! It treats the codes as if they were numbers and assumes a linear relationship!

Example:

Let us have the data *X*: 1, 2, 3, 4, 5, and *Y*: 1, 2, 3, 10, 90, exhibiting a perfect monotonic relationship, so we expect a correlation of 1 between *X* and *Y*.

However, Pearson's coefficient gives just 0.76 while both Kendall's and Spearman's coefficients result in the right value 1.



Tests of independence

H_0: X and Y are independent

H_1: X and Y are not independent

Tests of independence

H_0: X and Y are independent

H_1: X and Y are not independent

Available tests:

- Chi-squared test (also called Pearson's chi-squared test)
- Fisher's exact test

Small p-values ($p \leq 0.05$) indicate non-independence, i.e. presence of some kind of association between X and Y.

Be aware that the chi-squared test is valid only for large samples!

For small samples use Fisher's exact test instead!

Example:

Let us look at the contingency table gender by severity of symptoms, and test whether severity of symptoms is independent of gender!

Let us carry out both tests by the statistical software R!

- Pearson’s chi-squared test results in $p = 0.000011$
- Fisher’s exact test results in $p = 0.000008$

Example:

Let us look at the contingency table gender by severity of symptoms, and test whether severity of symptoms is independent of gender!

Let us carry out both tests by the statistical software R!

- Pearson's chi-squared test results in p = 0.000011
- Fisher's exact test results in p = 0.000008
- So both of them lead to rejection of independence
- With this sample size also the chi-squared test produces a good approximation of the p-value

Example:

Let us look at the contingency table gender by severity of symptoms, and test whether severity of symptoms is independent of gender!

Let us carry out both tests by the statistical software R!

- Pearson's chi-squared test results in $p = 0.000011$
- Fisher's exact test results in $p = 0.000008$
- So both of them lead to rejection of independence
- With this sample size also the chi-squared test produces a good approximation of the p-value

A free but absolute professional software with a full spectrum of the best analysis methods!

Joint analysis of several categorical variables

…can be made using loglinear models.

But I can't go into the details now, I just wanted to mention the name, so that you can remember to it, once you need.

Joint analysis of several categorical variables

…can be made using loglinear models.

But I can't go into the details now, I just wanted to mention the name, so that you can remember to it, once you need.

Methods for the **joint analysis of several categorical and continuous variables** (again just some names…)

- Analysis of variance, general linear models
- Logistic regression, generalized linear models
- Discriminant analysis

Distribution-free methods

Also called **non-parametric methods**, as opposed to parametric methods.

- Models
- Statistical tests
- Confidence intervals
- Correlation coefficients

Distribution-free methods

Also called **non-parametric methods**, as opposed to parametric methods.

- Models
- Statistical tests
- Confidence intervals
- Correlation coefficients

Parametric methods **assume that data follow a certain distribution** (e.g. normal, Poisson, etc.), and they are invalid if this assumption does not hold.

Distribution-free methods

Also called **non-parametric methods**, as opposed to parametric methods.

- Models
- Statistical tests
- Confidence intervals
- Correlation coefficients

Parametric methods **assume that data follow a certain distribution** (e.g. normal, Poisson, etc.), and they are invalid if this assumption does not hold.

Nonparametric methods are **valid for a wider class of distributions** (***assumptions are weaker, but there are some!***).

Distribution-free methods

Also called **non-parametric methods**, as opposed to parametric methods.

- Models
- Statistical tests
- Confidence intervals
- Correlation coefficients

That's why they are called distribution-free!

Parametric methods **assume that data follow a certain distribution** (e.g. normal, Poisson, etc.), and they are invalid if this assumption does not hold.

Nonparametric methods are **valid for a wider class of distributions** (***assumptions are weaker, but there are some!***).

Parametric methods:

- Student's t-test are valid only for ***normally distributed*** data
- Pearson's correlation coefficient is valid only for ***normally distributed data***
- ANOVA is valid only if data follow the ***normal distribution in each group***

Nonparametric methods:

- Wilcoxon's signed rank test is valid for any ***continuous and symmetric*** distribution
- Sign test is valid for any ***continuous*** distribution
- Spearman's rank correlation is valid for any data on ***ordinal and interval scale***

Most frequently used distribution-free methods:

- Sign test (one sample, paired samples)
- Mood’s median test (two or more samples)
- Wilcoxon signed rank test* (one sample, paired samples)
- Wilcoxon rank sum test*
  also called Mann-Whitney U-test (two samples)
- Kruskal-Wallis test* (several samples)
- Confidence interval for the median (one sample)
- Spearman’s rank correlation coefficient*
- Kendall’s tau (correlation coefficient)

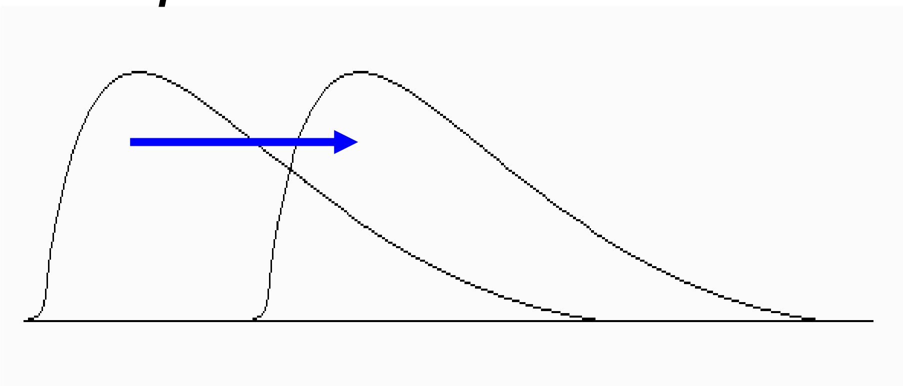
*rank-based methods

This is just a selection!

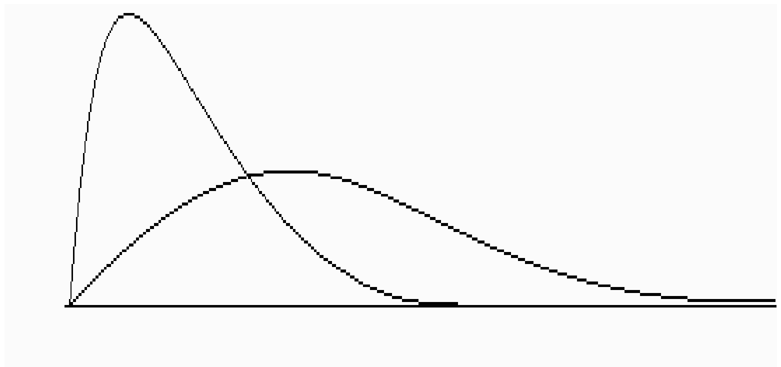
Be aware that **Wilcoxon rank sum test and Kruskal-Wallis test** in their original form are valid only if the variables to compare have **distributions of the same shape**!

Having the same shape means that the difference between the groups is simply a ***shift*** (which is in most cases irrealistic; then even the variances must be equal).

shapes are same here



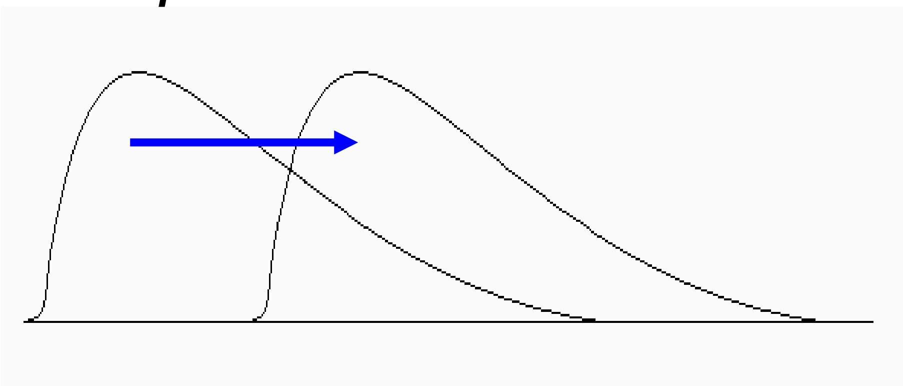
but not here



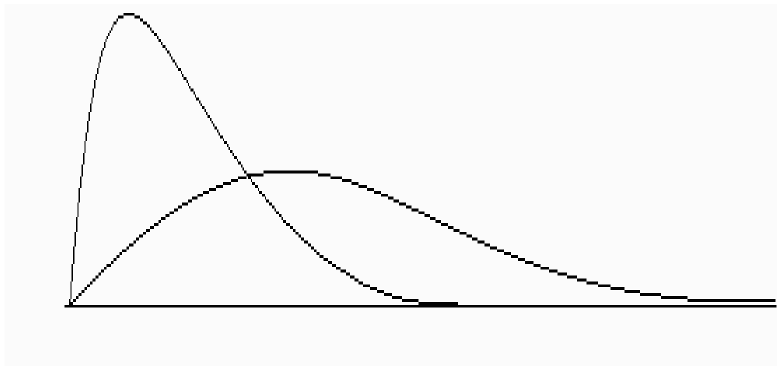
Be aware that **Wilcoxon rank sum test and Kruskal-Wallis test** in their original form are valid only if the variables to compare have **distributions of the same shape**!

Having the same shape means that the difference between the groups is simply a ***shift*** (which is in most cases irrealistic; then even the variances must be equal).

shapes are same here



but not here



Fortunately, there are newer versions of these tests, which don't require this rather restrictive assumption! Check the latest literature!