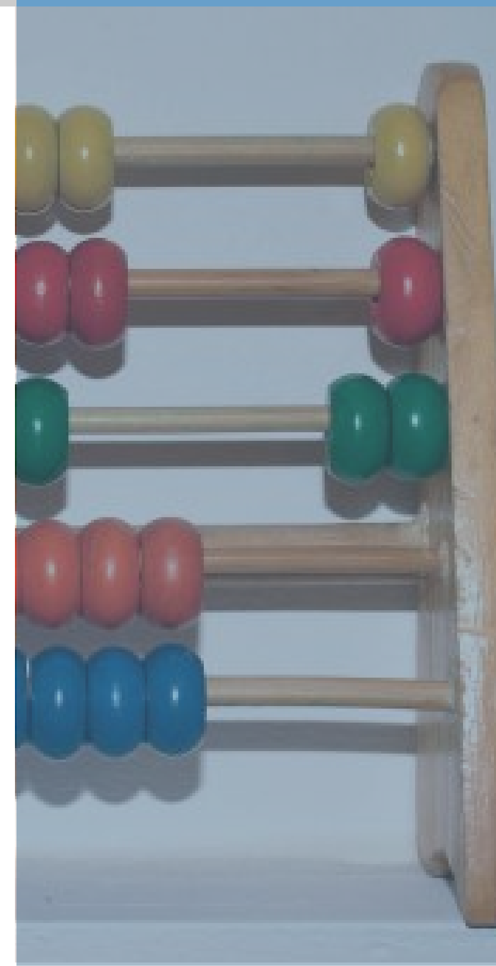
05_nonparametric_tests

Hypothese testing, Kolmogorov-Smirnov, Mann-Whitney-U

Inductive statistics or statistical inference

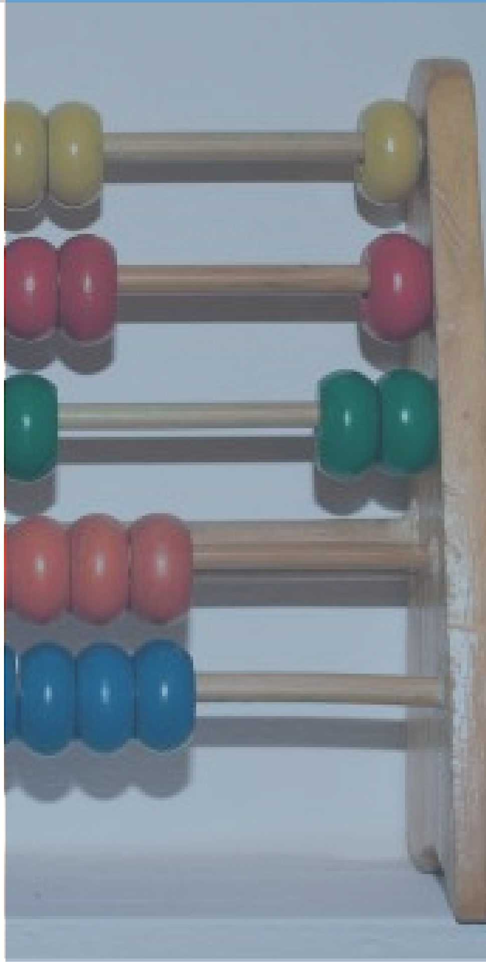
Is used to draw conclusions about (unknown) parameters of the population on basis of a sample
The results are always statistical ;-)
i.e. all statements are true with a certain probability but could be also false with a certain probability

The basis of statistical inference is probability theory (stochastic)



Christian-Albrechts-Universität zu Kiel

Population and sample [1]



Repetition:

Population
Amount of all items of relevance for an analysis.

Sample
Selection of items on basis of certain criteria (e.g. representativity) which will be analysed instead of the population

The difference should always be kept in mind
In archaeology only sampling is possible! The population can never be investigated!

Population and sample [1]

Features of the population: parameters

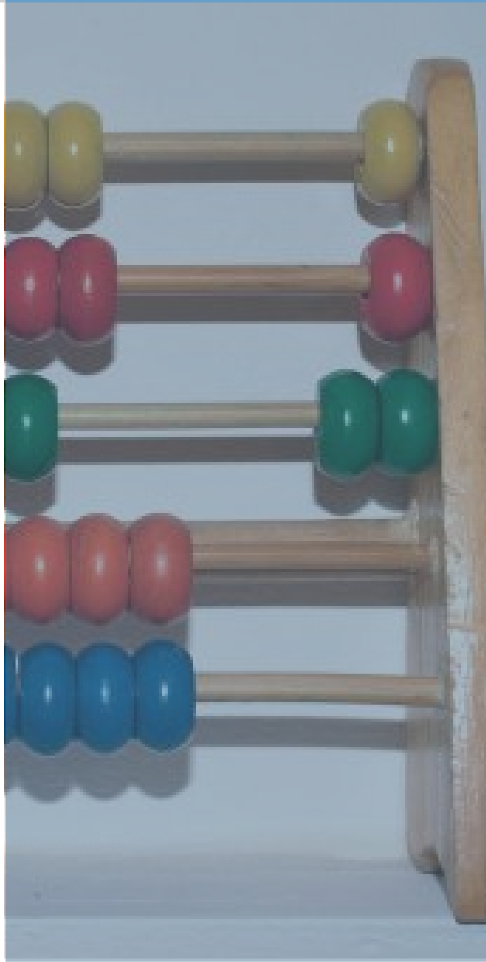
Parameters always exist, they have a certain value, but they are unknown and often (most of the time) also uncheckable.

Example:

$$\mu : \textit{mean of the population}$$
$$\bar{x} : \textit{mean of the sample}$$

$$\sigma : \textit{standard deviation of the population}$$
$$s : \textit{standard deviation of the sample}$$

In statistical tests only features of the sample could be checked. The quality of the statement of a test therefore depends on the choice of the sample (representativity)!



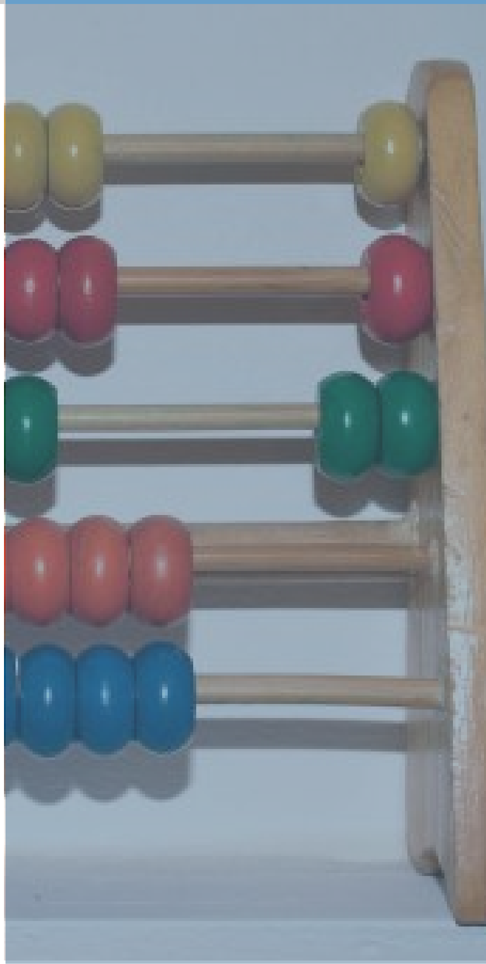
Statistical hypothesis testing

Validation of an assumption about the population
A assumption (hypothesis) about the population is made and than its probability is checked against the sample.

Usual questions:

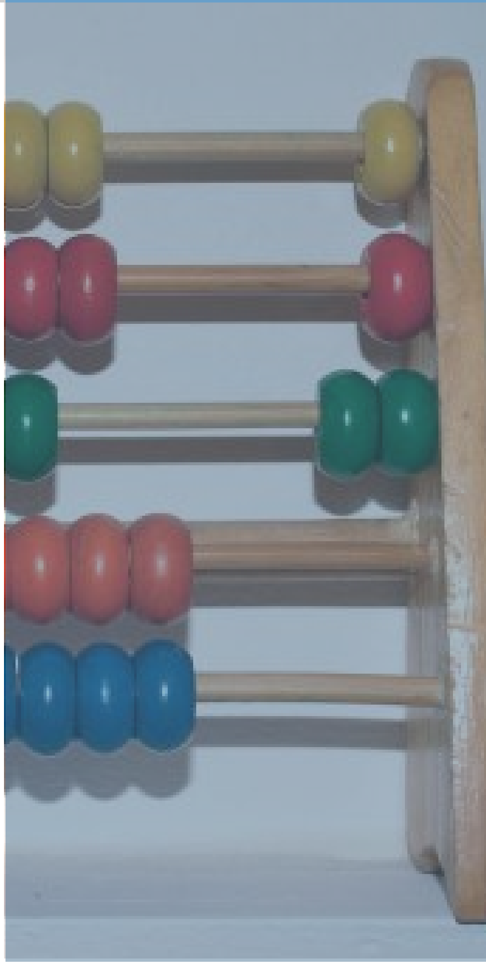
How probable is it that two or more samples descend from the different/the same population?
(eg. Is the custom of grave goods for man and women so different that two different social groups are visible?)

How probable is it that a given sample descend from a population with certain parameters?
(Is the amount of grave goods random or is a pattern visible?)



Christian-Albrechts-Universität zu Kiel

Null hypothesis [1]



Validation through falsification

In statistical tests most of the times not the statement is tested which one expects to be true but one tries to disprove the statement which one expects to be wrong: the null hypothesis.
This hypothesis states mostly, that a association do **not** exists or that there is **no** differences between the samples and the distribution of the observations is by chance.

Example: Is the composition of grave goods different between male and female deceased?

$$H_0: The\, composition\, is\, the\, same$$
$$H_1: The\, composition\, is\, different$$

reason:

1. It is (logical) easier to prove, that a statement is wrong (falsify) then to prove that a statement is true (verify).
2. Most of the times it is easier to formulate a null hypothesis (How exactly is the composition different?). It doesn't make a assumption about how the character of a association/difference exactly is.

Null hypothesis [2]

„Workflow“ of a statistical test

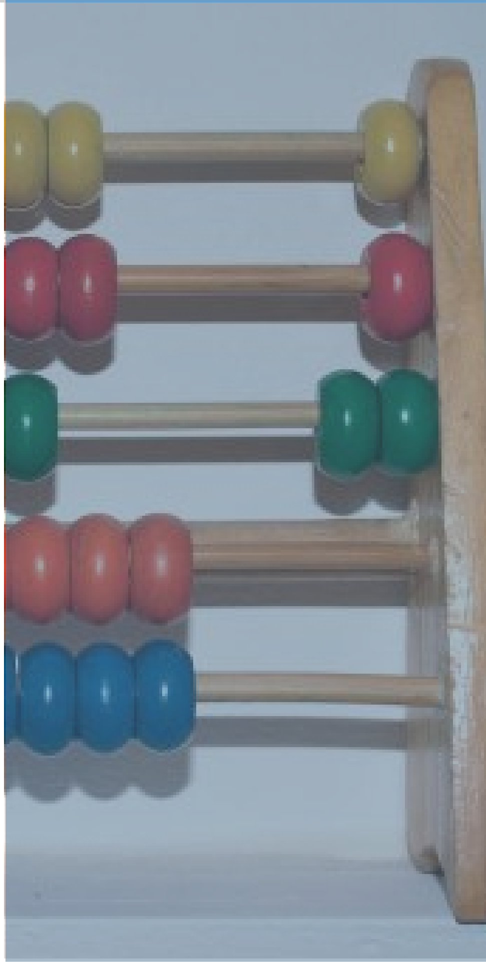
Construction of a alternative hypothesis:
The composition of the grave goods is different between male and female deceased.

Construction of the null hypothesis:
The composition of the grave goods is the same in male and female burials.

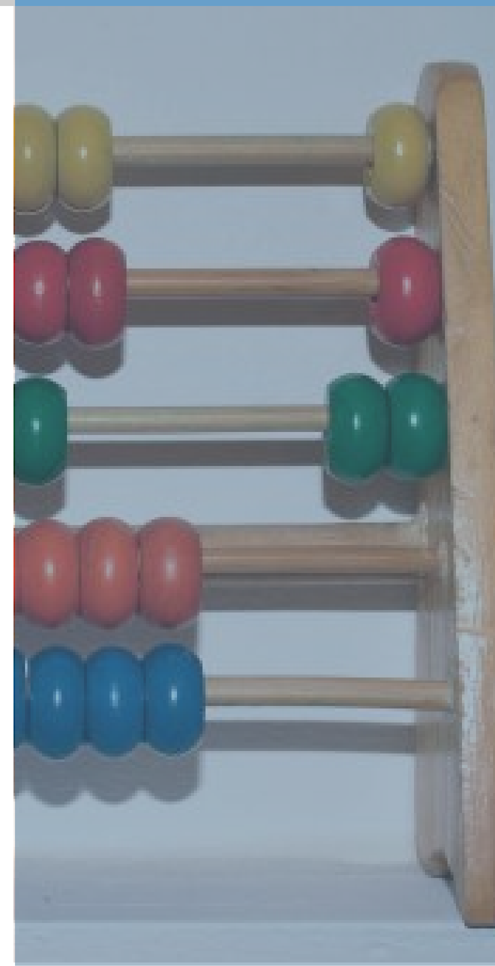
Test of the null hypothesis

If the result of the test is significant:
Rejection of the null hypothesis, choice of the alternativ hypothesis.
The composition of the grave goods is different between male and female deceased.

If the result of the test is not significant:
The null hypothesis could not be rejected.
We can not say if the composition of the grave goods is different between male and female deceased or not!



One-tailed/Two-tailed hypothesis



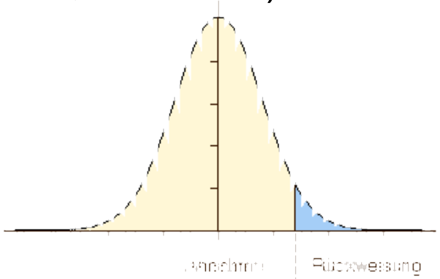
one-tailed oder two-tailed

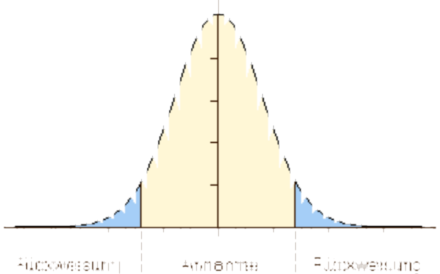
Dependend on the question there could be a different number of alternative hypothesis.

Example:
Is the number of grave goods in female burials higher than in male?
One-tailed hypothesis, possible answers are yes or no.

Is the number of grave goods in female burials different from male?
Two-tailed hypothesis, possible answers smaller-equal-greater.

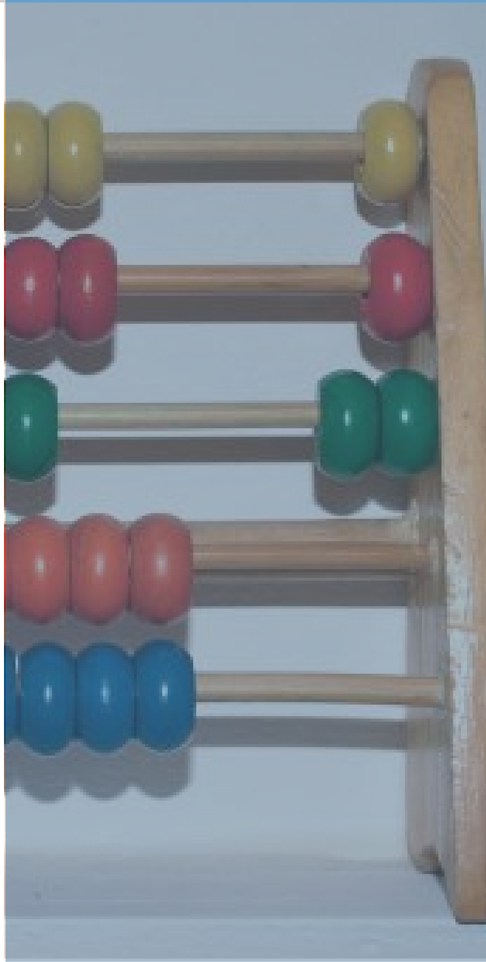
That's why in statistical tests the result is often two significances (one-tailed, two-tailed).





source: http://www.statistics4u.info/fundstat_germ/cc_test_one_two_sided.html

Stat. Significance



How true is true?

Statistical significance is effectively a measurement how probable a error is.

On basis of the significance the null hypothesis will be rejected and the alternative hypothesis will be choosen … or not.

There are classic boundary values for significance (significance levels):

0.05: significant, with 95% probability the decision is right.

0.01: very significant, with 99% probability the decision is right.

0.001: highly significant, with 99,9% probability the decision is right.

Often named with p-value or α.

α- und β-error [1]

If statistics go wrong...

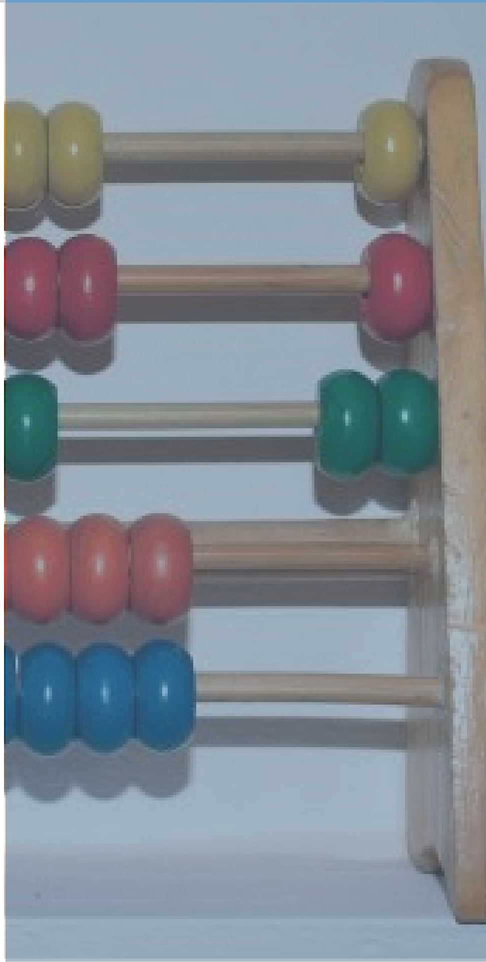
There are two kinds of possible errors:

The null hypothesis was rejected although it is true
Type I error, false positive, α-error
The result of a pregnancy test is false positive if it shows a pregnancy although there is none.

The null hypothesis was not rejected although it is wrong
Type II error, false negative, β-error
The result of a pregnancy test is false negative if it shows no pregnancy although there is one.

	True condition: H0 (There is no difference)	True condition: H1 (There is a difference)
By the use of a statistical test the decision was made for: H0	Correct decision	Type II error
By the use of a statistical test the decision was made for: H1	Type I error	Correct decision

source: wikipedia



α- und β-error [2]

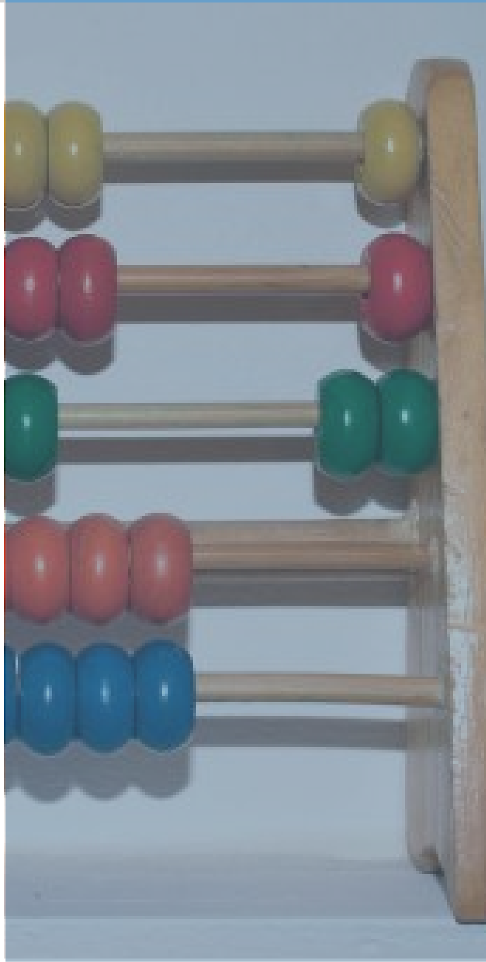
Tests and errors

Statistical tests should avoid both types of errors
balancing act (not to strict/not strict enought)

General Type I Errors are more serious than Type II Errors
This type leads to wrong assuptions because with it the alternative hypothesis seems to be proven, in case of a Type I Error nothing is proven

Power of a test
A test has more power if he avoids Type II Errors without risking more Type I errors.

A more powerful test helps to clarify issues better



Nonparametric tests

Parametric vs. nonparametric

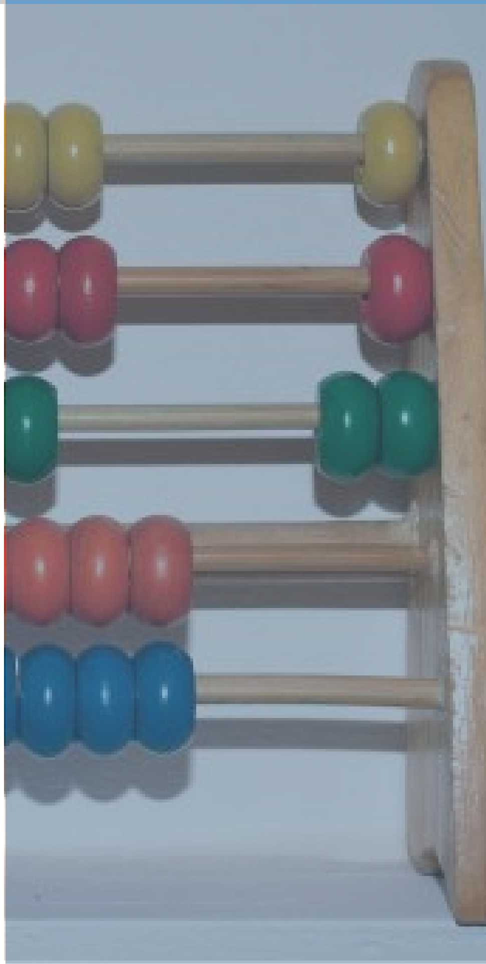
Parametric: The distribution of the values have to be in a certain form (e.g. normal distribution); assumptions about the distribution of the population are needed

non-parametric: no assumptions about the distribution of the sample and the population are needed

Nonparametric tests, advantages and disadvantages:

Advantage: Also appropriate if no statements about the distribution are possible or the distribution fits no for parametric tests.
Also smaller samples are possible.

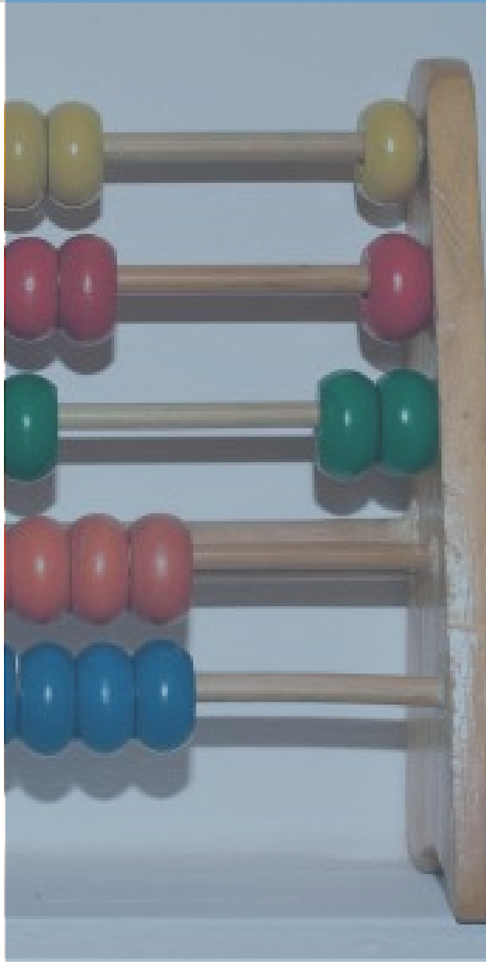
Disadvantages: Tests have general a lesser power.



+

Test

Kolmogorov-Smirnov-Test [1]



Test for difference of two distributions

requirements: at least one ordinal scaled Variable (one sample case) and 1 nominal scaled grouping variable (two sample case)

Procedure one sample case: the culmulative procentual frequency of the sample is compared with a standard distribution (often normal distribution)

Procedure two sample case: the culmulative procentual frequencies of the samples is compared

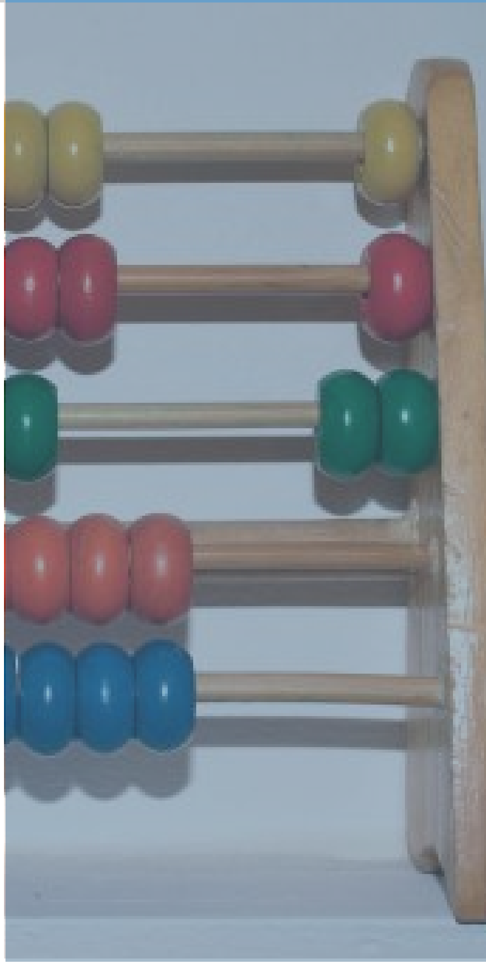
Kolmogorov-Smirnov-Test [2]

Example (after Shennan)

Female bronze age burials in a grave yard by age

Age at the moment of death	Wealth category	
	rich	poor
Infans I	6	23
Infans II	8	21
Juvenilis	11	25
Adultus	29	36
Maturus	19	27
Senilis	3	4
total	**76**	**136**

Question: Differ the live conditions of poor and rich buried people that much so that different life ages were reached?



Kolmogorov-Smirnov-Test [3]

requirements

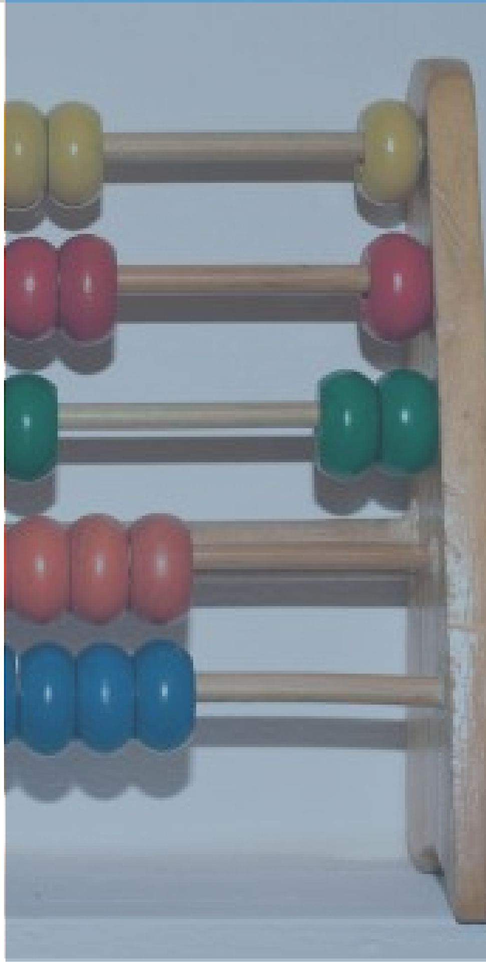
H_0: There is no difference between rich and poor graves according to age of death.

H_1: There is a difference between rich and poor graves according to age of death.

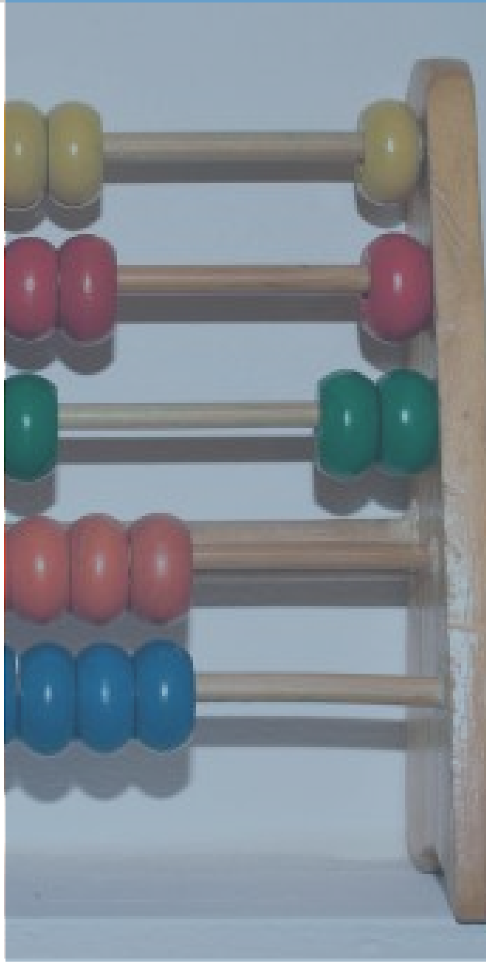
Two-tailed test.

Level of significance: 0.05

variables:
1. ordinal scaled age classes
2. (at least) nominale (ordinale) scaled wealth classes



Kolmogorov-Smirnov-Test [4]



Procedure: Calculation of the procentual frequency

Divide every cell of a column by the sum of the column

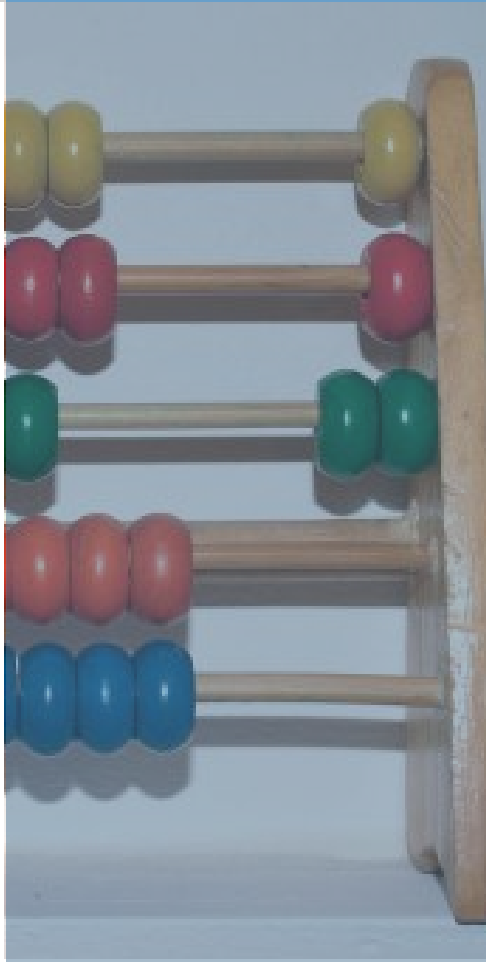
Age at the moment of death	Wealth category			
	rich		poor	
Infans I	6	0.079	23	0.169
Infans II	8	0.105	21	0.154
Juvenilis	11	0.145	25	0.184
Adultus	29	0.382	36	0.265
Maturus	19	0.250	27	0.199
Senilis	3	0.039	4	0.029
total	**76**	1.000	**136**	1.000

Kolmogorov-Smirnov-Test [5]

Procedure: Calculate the culmulative procentual frequency

Add to every procentual frequency the values of procentual frequencies of the lower ordinal scaled values

Age at the moment of death	Wealth category					
	rich			poor		
Infans I	6	0.079	0.079	23	0.169	0.169
Infans II	8	0.105	0.184	21	0.154	0.323
Juvenilis	11	0.145	0.329	25	0.184	0.507
Adultus	29	0.382	0.711	36	0.265	0.772
Maturus	19	0.250	0.961	27	0.199	0.971
Senilis	3	0.039	1.000	4	0.029	1.000
total	**76**	1.000		**136**	1.000	



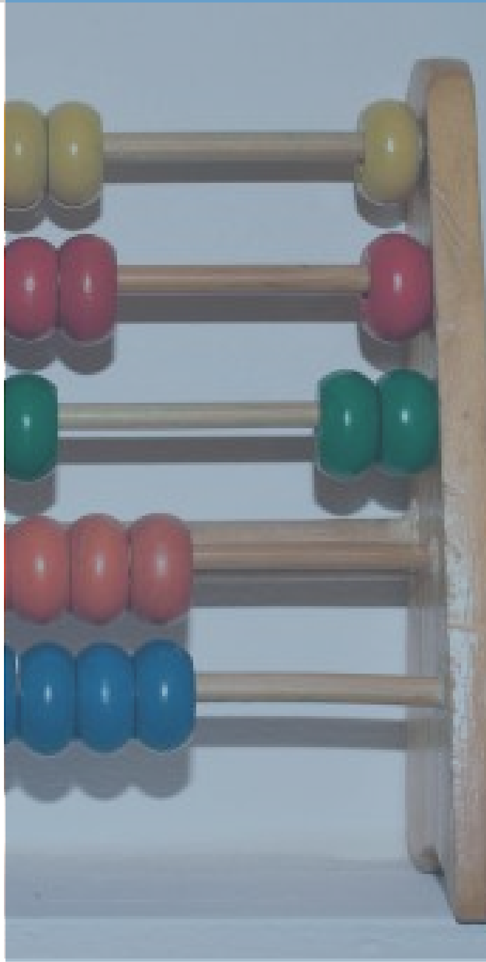
Kolmogorov-Smirnov-Test [6]

Procedure: Calculate the differences of the culmulative procentual frequencies

Substract the culmulative procentual frequencies from each other, make that value absolute (without sign)

Age at the moment of death	Wealth category		difference
	rich	poor	
Infans I	0.079	0.169	0.090
Infans II	0.184	0.323	0.139
Juvenilis	0.329	0.507	0.178
Adultus	0.711	0.772	0.061
Maturus	0.961	0.971	0.010
Senilis	1.000	1.000	0.000

Largest difference



Kolmogorov-Smirnov-Test [7]

Compare the maximum difference with a boundary value which is calculated from the total number of cases

Total number rich: 76
Total number poor: 136
Difference max (D_{max}): 0.178
Level of significance: 0.05

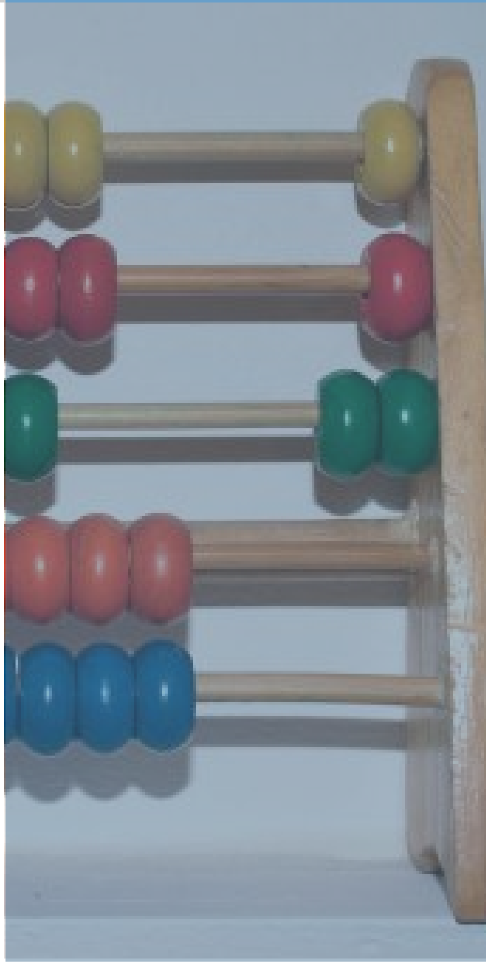
$$formula:$$
$$boundary\ value\ KS-Test = factor\ f * \sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

Factor f:
Level of significance 0.05: 1.36
Level of significance 0.01: 1.63
Level of significance 0.001: 1.95

That's why: $$boundary\ value\ KS-Test = 1.36 * \sqrt{\frac{76+136}{76\,136}} = 0.195$$

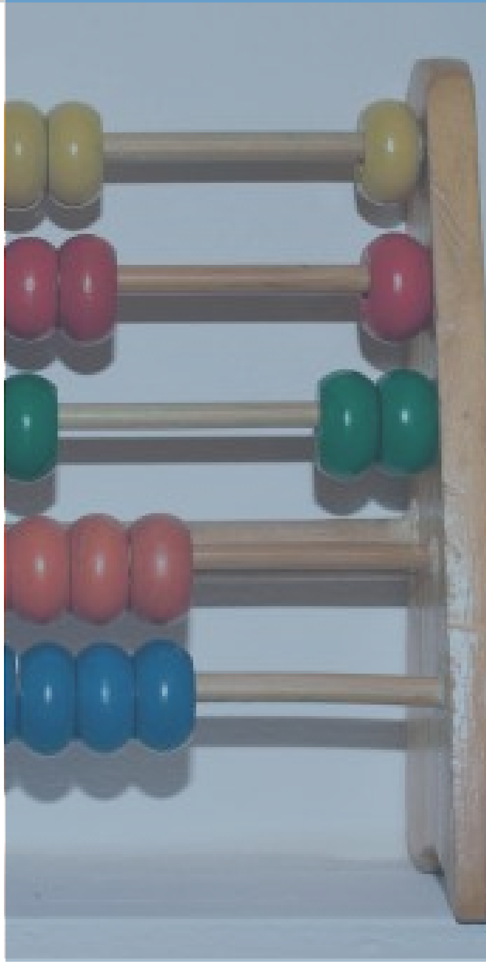
Dmax < boundary value, difference is not significant

But: That doesn't mean that the distributions are equal, only that they do not differ significant.



Christian-Albrechts-Universität zu Kiel

Kolmogorov-Smirnov-Test [8]



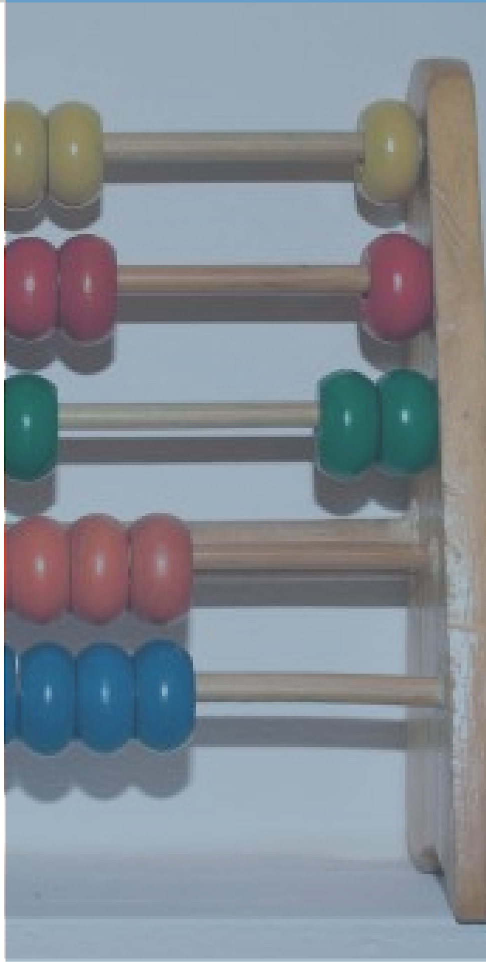
KS-Test in R

```
> graeberbrz<-read.csv2("graeberbrz.csv",row.names=1)
> table(graeberbrz)
     reichtum
alter arm reich
    1   6    23
    2   8    21
    3  11    25
    4  29    36
    5  19    27
    6   3     4
> alter<-graeberbrz$alter
> reichtum<-graeberbrz$reichtum
> ks.test(alter[reichtum=="arm"],alter[reichtum=="reich"])

      Two-sample Kolmogorov-Smirnov test

data:  alter[reichtum == "arm"] and alter[reichtum == "reich"]
D = 0.1784, p-value = 0.08977
alternative hypothesis: two-sided

Warning message:
In ks.test(alter[reichtum == "arm"], alter[reichtum == "reich"]) :
  kann bei Bindungen nicht die korrekten p-Werte berechnen
```

Kolmogorov-Smirnov-Test Exercise

Cups from relative closed finds from late neolithic inventories (Müller 2001)
Analyse with the Kolmogorov-Smirnov-Test if the heigths of cups with and without corner points differ significant on a 0.05-level.

File: mueller2001.csv



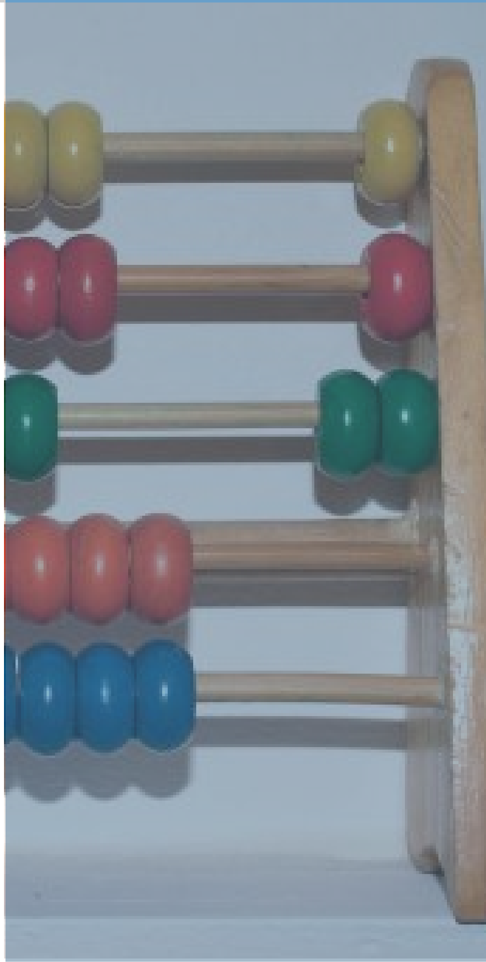
Kolmogorov-Smirnov-Test Lösung

Cups from relative closed finds from late neolithic inventories (Müller 2001)

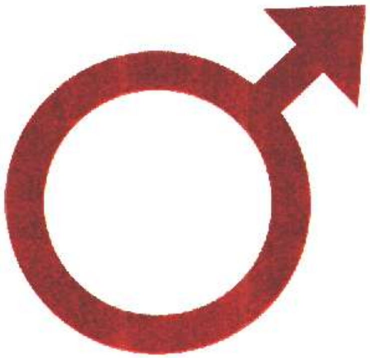
Analyse with the Kolmogorov-Smirnov-Test if the heigths of cups with and without corner points differ significant on a 0.05-level.

File: mueller2001.csv

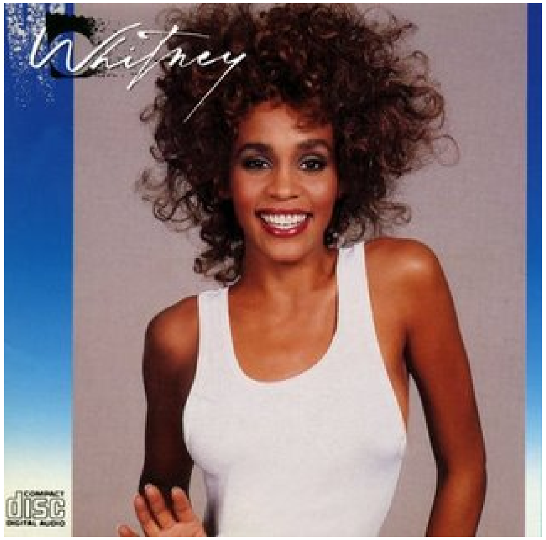
```
> mueller<-read.csv2("mueller2001.csv")
> tassentyp<-mueller$tassentyp
> hoehe<-mueller$hoehe
> ks.test(hoehe[tassentyp=="eingliedrig"],hoehe[tassentyp=="zweigliedrig"])

        Two-sample Kolmogorov-Smirnov test

data:  hoehe[tassentyp == "eingliedrig"] and hoehe[tassentyp == "zweigliedrig"]
D = 0.2519, p-value = 0.1020
alternative hypothesis: two-sided

Warning message:
In ks.test(hoehe[tassentyp == "eingliedrig"], hoehe[tassentyp ==  :
  kann bei Bindungen nicht die korrekten p-Werte berechnen
```



Christian-Albrechts-Universität zu Kiel



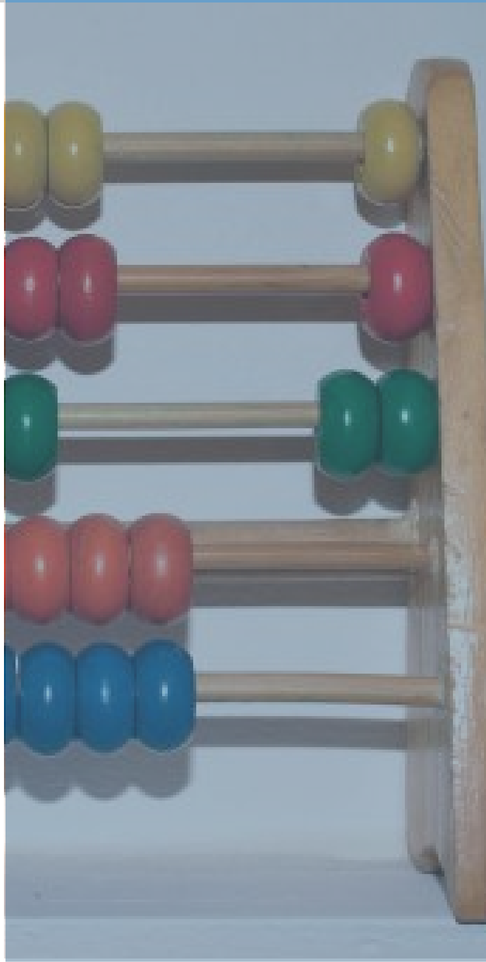
+



+

Test

Mann-Whitney-U-Test [1]
(=Wilcoxon rank-sum test)

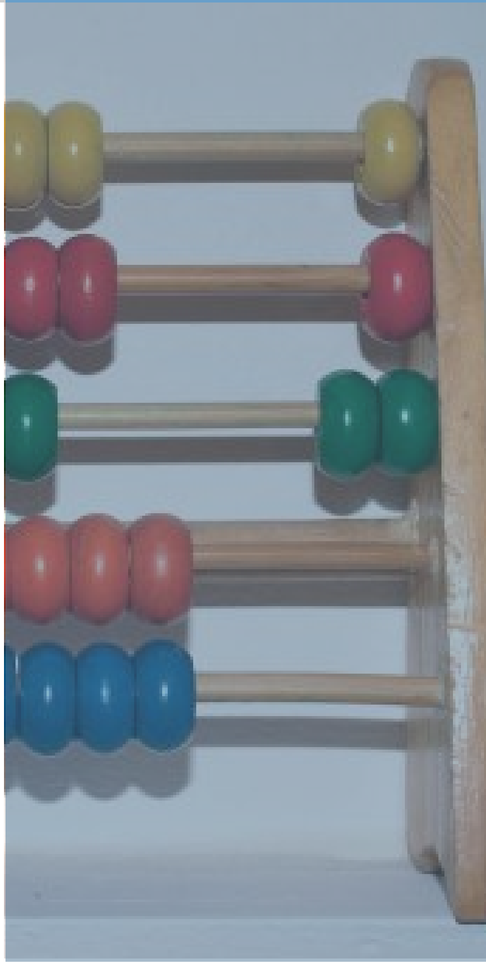


Test for differences of two distributions

Requirements: at least 1 interval- or ordinale scaled variable and 1 nominale scaled grouping variable

Procedure: The values were sorted and for every group the ranks were compared

Mann-Whitney-U-Test [2]



Example (after Müller-Scheeßel)

Chamber sizes of iron age chamber burials by sex

Chamber size	sex
11,7	m
4,4	f
35,9	m
8,0	f
23,0	m
5,1	f
9,2	m
15,8	f
26,1	m
7,3	f

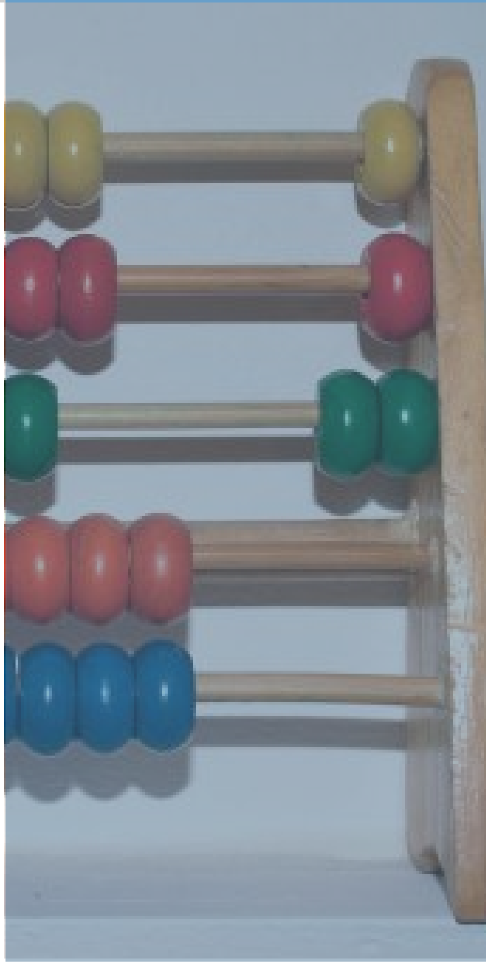
Question: Do the sizes differ in relation to the sex of the buried?

Mann-Whitney-U-Test [3]

Procedure

Determination of the rank of the graves according to size

Chamber size	sex	rank
11,7	m	5
4,4	f	10
35,9	m	1
8,0	f	7
23,0	m	3
5,1	f	9
9,2	m	6
15,8	f	4
26,1	m	2
7,3	f	8

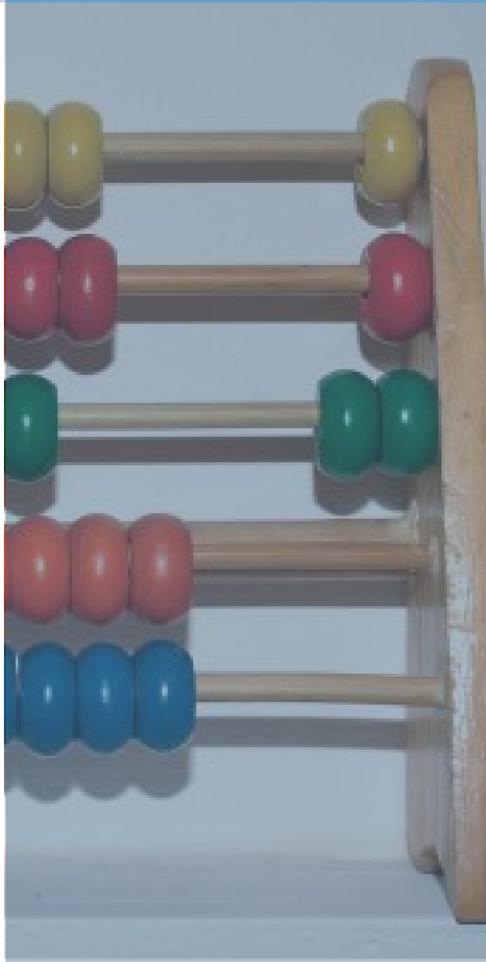


Mann-Whitney-U-Test [4]

Procedure

Sort according to rank

Chamber size	sex	rank
35,9	m	1
26,1	m	2
23,0	m	3
15,8	f	4
11,7	m	5
9,2	m	6
8,0	f	7
7,3	f	8
5,1	f	9
4,4	f	10

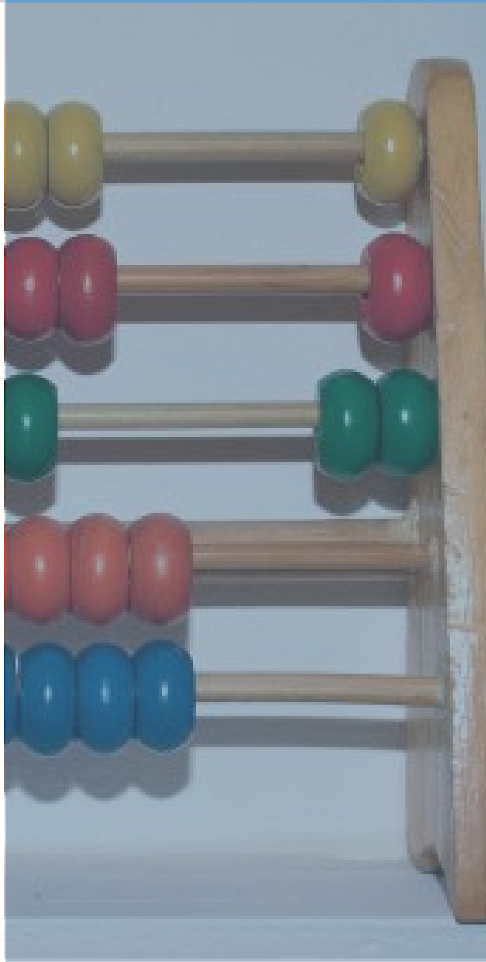


Mann-Whitney-U-Test [5]

Procedure

Count how many values of the opposite category are below the actual value

Chamber size	sex	rank	M below	F below
35,9	m	1		5
26,1	m	2		5
23,0	m	3		5
15,8	f	4	2	
11,7	m	5		4
9,2	m	6		4
8,0	f	7		
7,3	f	8		
5,1	f	9		
4,4	f	10		
Summe			2	23



Mann-Whitney-U-Test [6]

Procedure

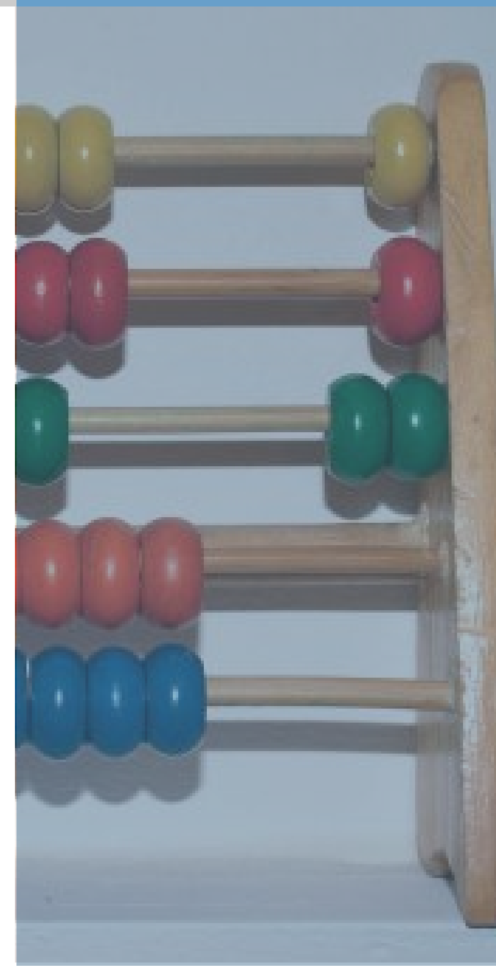
Number of male burials: 5
Number of female burials: 5
Rank sum of male burials: 23
Rank sum of female burial: 2

5*5=25=23+2

The smaller value will be evaluated: 2

Look up in a table (e.g. Shennan 1997, Table B):
Boundary value for significance 0.05 when n1=5 and n2=4: 2

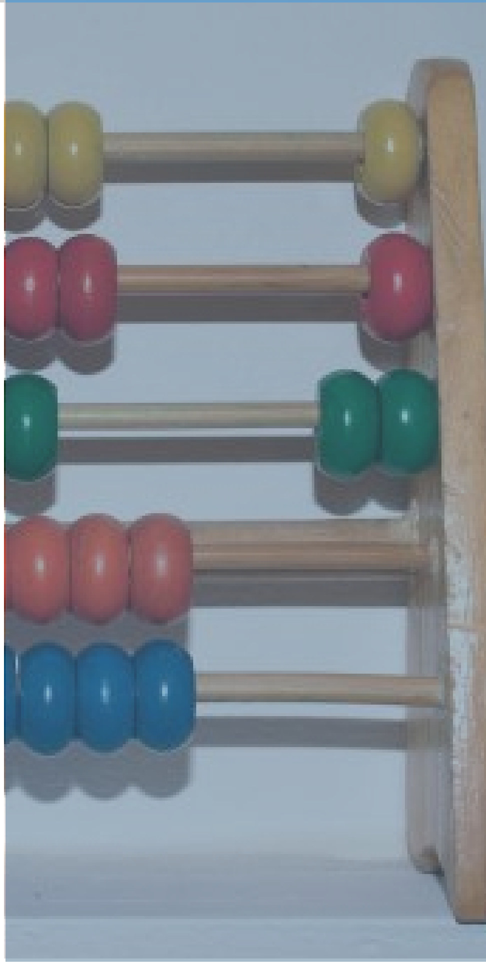
The chamber sizes do differ from each other significant.



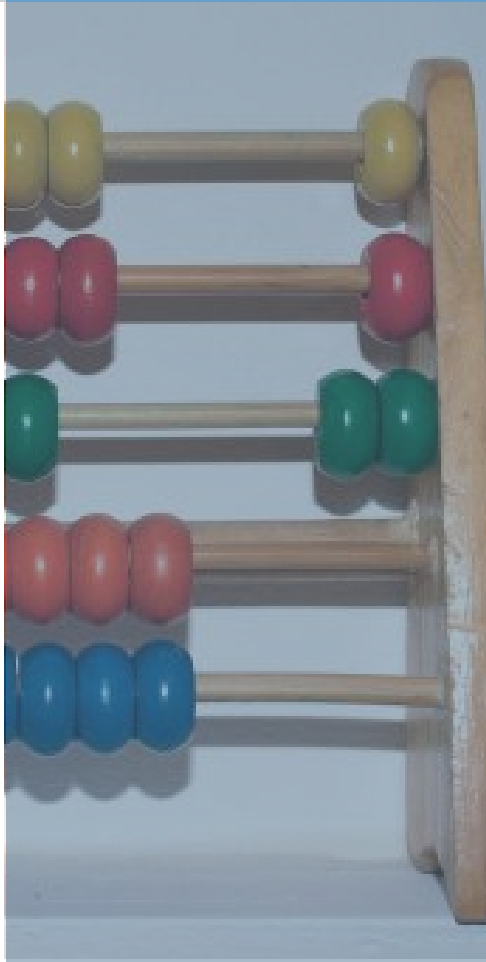
Mann-Whitney-U-Test [7]

Mann-Whitney-U-Test in R

```
> kammergroesse<-read.csv2("kammergroesse_mueller-scheessel.csv")
> kammergroesse
   kammergroesse geschlecht
1           35.9          m
2           26.1          m
3           23.0          m
4           15.8          w
5           11.7          m
6            9.2          m
7            8.0          w
8            7.3          w
9            5.1          w
10           4.4          w
> wilcox.test(kammergroesse$kammergroesse ~
kammergroesse$geschlecht)

     Wilcoxon rank sum test

data:  kammergroesse$kammergroesse by kammergroesse$geschlecht
W = 23, p-value = 0.03175
alternative hypothesis: true location shift is not equal to 0
```



Mann-Whitney-U-Test Aufgabe

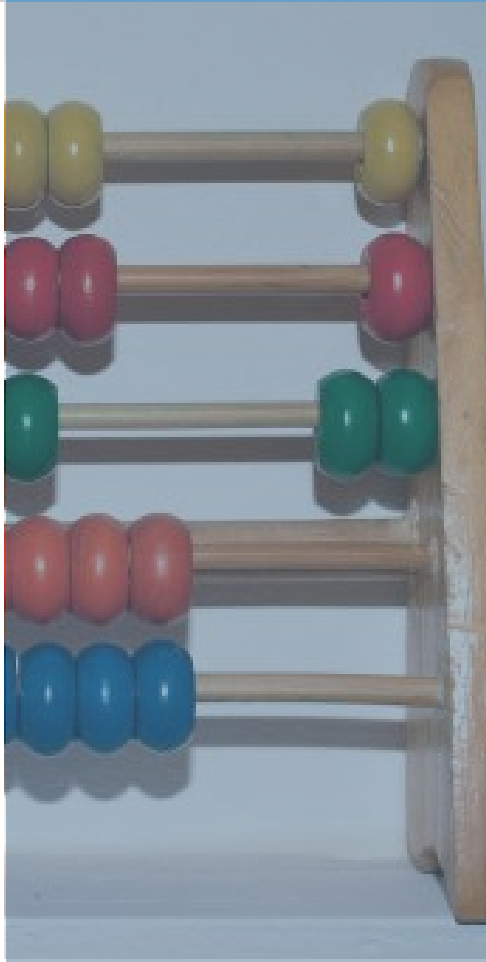


Length of flanged axes of types Bikun and Cegun (Cullberg 1968)

Analyse with the Mann-Whitney-U-Test if the length of flanged axes of the types Bikun and Cegun differ significant on a 0.05-level.

file: cullberg1968.csv

Mann-Whitney-U-Test Lösung



Length of flanged axes of types Bikun and Cegun (Cullberg 1968)

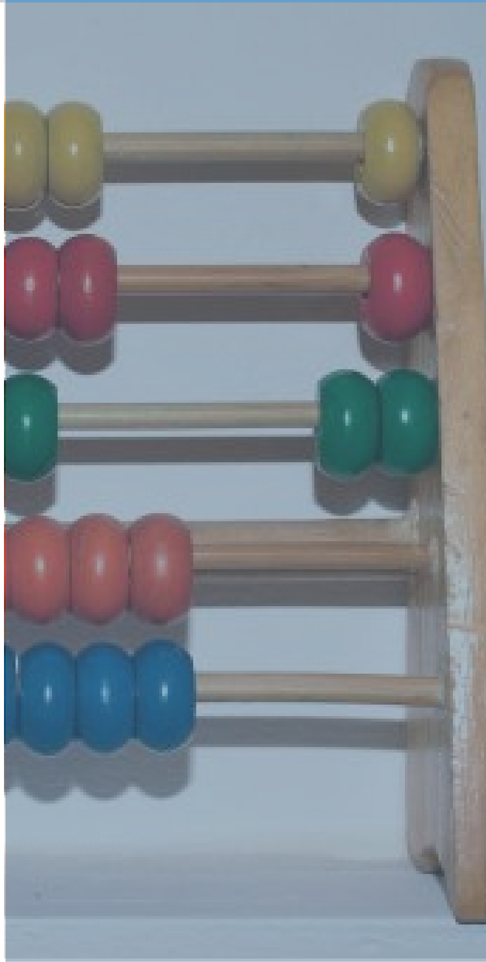
Analyse with the Mann-Whitney-U-Test if the length of flanged axes of the types Bikun and Cegun differ significant on a 0.05-level.

file: cullberg1968.csv

```
> cullberg<-read.csv2("cullberg1968.csv")
> laenge<-cullberg$laenge
> typ<-cullberg$typ
> wilcox.test(laenge[typ=="Bikun"],laenge[typ=="Cegun"])

     Wilcoxon rank sum test with continuity correction

data:  laenge[typ == "Bikun"] and laenge[typ == "Cegun"]
W = 17.5, p-value = 0.02673
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(laenge[typ == "Bikun"], laenge[typ ==
"Cegun"]) :
  kann bei Bindungen keinen exakten p-Wert Berechnen
```

Christian-Albrechts-Universität zu Kiel

Interpretation of significance tests



Pay attention also when the statistic seem to be clear

After the test as well as before the test: The interpretation determines the result!

Statistically significant ≠ archaeologically significant!

Statistical results stay statistical: significance is always probability that the choice of a hypothesis is correct, but there is also a probability that it is by chance...