Statistical methods for archaeological data analysis I: Basic methods

05 - Nonparametric Tests

Martin Hinz

Institut für Archäologische Wissenschaften, Universität Bern

01.04.2025

You can download a pdf of this presentation.

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)

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 [2]

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:

\(\mu:\) mean of the population

\(\sigma\): standard deviation of the population

\(\bar{x}\): mean of the sample

\(s\): 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?)

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 compositionisthe 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:

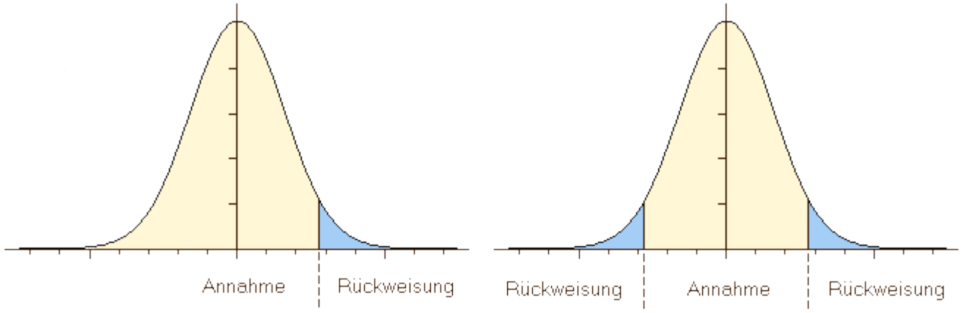
Is the number of grave goods in female burials higher than in male?

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 (onetailed, two-tailed).



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 \(\alpha\).

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, \(\alpha\)-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, \(\beta\)-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

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.

\(\chi^2\) test

(- tai)^2

\(\chi^2\) test [1]

Possible Questions

Do settlements tend to be situated on rather good soil or is the distribution random?

Conclusions about settlement behaviour and economy would be possible

Do older individuals have more shoe-last celt as grave goods than younger?

If shoe-last celt would be signs of social rank than this situation would make conclusions possible about heredity or acquisition of social rank during life time.

Tests for nominal scaled variables are possible!

Therefore of particular value for archaeology because we have often to deal with such data.

\(\chi^2\) test [2]

Test for independence of two distributions

Requirements: at least 1 nominal scaled variable (one sample case) and 1 nominal scaled grouping variable (two sample case)

Procedure with one sample: observed values are compared with expected values given a certain distribution, no expected value should be < 5; n should be > 50

Procedure with two samples: observed values of both distributions are compared with expected values if the samples would be even distributed, no expected value should be < 5; n should be > 50

If sample size is too small: Fishers Exact Test

Test statistics: \(\chi^2\)

Significance depend on degree of freedom (df)

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	female	total
cremation			201
inhumation			197
total	216	182	398

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	female	total
cremation	123		201
inhumation			197
total	216	182	398

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	female	total
cremation	123	78	201
inhumation	93	104	197
total	216	182	398

df=1: if one value is chosen all other can be calculated with the help of the margins

(number of columns – 1)*(number of rows – 1)

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	female	uncertain	total
cremation				201
inhumation				197
total	196	179	23	398

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	female	uncertain	total
cremation		78		201
inhumation				197
total	196	179	23	398

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	female	uncertain	total
cremation	113	78		201
inhumation				197
total	196	179	23	398

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	**female**	**uncertain**	**total**
cremation	113	78	10	201
inhumation	83	101	13	197
total	196	179	23	398

df=2: if two values are chosen all other can be calculated with the help of the margins

(number of columns – 1)*(number of rows – 1)

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	**female**	**uncertain**	**total**
cremation				201
inhumation				197
other				30
total	201	187	40	428

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	female	uncertain	total
cremation		78		201
inhumation	83		13	197
other		8		30
total	201	187	40	428

Excursus degree of freedom

Number of slots free to vary given the margin sums

	male	female	uncertain	total
cremation	113	78	10	201
inhumation	83	101	13	197
other	5	8	17	30
total	201	187	40	428

\(\chi^2\) test [3]

Test for one sample (example after Shennan)

Numbers of neolithic settlements by soil type in eastern france

Soil type	Number of settlements
Rendzina	26
Alluvial	9
Brown earth	18
total	53

Question: Is there a significant preference for a soil type?

We calculate two versions:

1. even distributed

2. even distributed with consideration of the proportion of the soil types on the total area

\(\chi^2\) test [4]

Version 1: even distributed

Soil type	Number of settlements	Proportion of soil type	expected number of settlements
Rendzina	26	1/3	17.6667
Alluvial	9	1/3	17.6667
Brown earth	18	1/3	17.6667
total	53	1	53

\(H_0\): The settlements are evenly distributed on all soil types.

\(H_1\): The settlements are **not** evenly distributed on all soil types.

χ^2 test [5]

Version 1: even distributed

Soil type	Number of settlements	Proportion of soil type	expected number of settlements
Rendzina	26	1/3	17.6667
Alluvial	9	1/3	17.6667
Brown earth	18	1/3	17.6667
total	53	1	53

Formula for χ^2:

$\chi^2=\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$

O_i: number of **observed** cases

E_i: number of **expected** cases

χ^2: symbol for the test statistic chi-squared

χ^2 test [6]

Procedure: Calculation of the X^2-value

$\chi^2=\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$

Soil type	Number of observed cases	Number of expected cases	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Rendzina	26	17.6667	8.3333	69,4444	3.9308
Alluvial	9	17.6667	-8.6667	75,1117	4.2516
Brown earth	18	17.6667	0.3333	0.1111	0.0063
total	53	53			**8.18868**

Look up in a table (e.g. Shennan): Df=2 (2 colums (expected, observed), 3 categories)

Level of significance: 0.05

Boundary value: 5.99145

Significant result: The distribution is uneven!

\(\chi^2\) test [7]

Version 2: even distributed with consideration of the proportion of the soil types on the total area

Soil type	Number of settlements	Proportion of soil type	expected number of settlements
Rendzina	26	32%	16.69
Alluvial	9	25%	13.25
Brown earth	18	34%	22.79
total	53	1	53

\(\chi^2=\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}\)

χ^2 test [8]

Procedure: Calculation of the X^2-value

$\chi^2=\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$

Soil type	Number of observed cases	Number of expected cases	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Rendzina	26	16.69	9.04	81.7216	4.8185
Alluvial	9	13.25	-4.25	18.0625	1.1363
Brown earth	18	22.79	-4.79	22.9441	1.007
total	53	53			**7.1885**

Look up in a table (e.g. Shennan): Df=2 (2 colums (expected, observed), 3 categories)

Level of significance: 0.05

Boundary value: 5.99145

Significant result: The distribution is uneven also if we consider the proportions of the soil types!

\(\chi^2\) test [9]

```
siedlungen<-c(26,9,18)
names(siedlungen)<-c("Rendzina","Alluvial","Braunerde")
siedlungen
```

```
##  Rendzina  Alluvial Braunerde
##        26         9        18
```

Version 1: even distributed

```
chisq.test(siedlungen)
```

```
##
##      Chi-squared test for given probabilities
##
## data:  siedlungen
## X-squared = 8, df = 2, p-value = 0.02
```

Version 2: even distributed with consideration of the proportion of the soil types on the total area

```
chisq.test(siedlungen,p=c(0.32,0.25,0.43))
```

```
##
##      Chi-squared test for given probabilities
##
## data:  siedlungen
## X-squared = 7, df = 2, p-value = 0.03
```

\(\chi^2\) test [10]

Two sample case (Test for independence)

(example after Hinz, beautified)

Comparison of amber in graves and settlements

Classic 2x2 situation

Type of site	amber		total
	+	-	
settlement	6	18	24
grave	132	44	176
total	138	62	200

Is amber primary a grave good?

df=1

Level of significance = 0.05

\(\chi^2\) test [11]

Procedure: Calculation of the expected values

Multiply the margins and divide the result by the total number

Type of site	amber		total
	+	-	
settlement	24*138/200 = 16.56	24*62/200=7.44	24
grave	138*176/200=121.44	62*176/200=54,56	176
total	138	62	200

\(\chi^2\) test [12]

Procedure: Calculation of the expected values

Multiply the margins and divide the result by the total number

Type of site	amber		total
	+	-	
settlement	O=6 vs. E=16.56	O=18 vs. E=7.44	24
grave	O=132 vs. E=121.44	O=44 vs. E=54.56	176
total	138	62	200

\(\chi^2=\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}\)

χ^2 test [13]

Procedure: Calculation of the expected values

$\chi^2=\sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$

Type of site	amber		total
	+	-	
settlement	(6-16.56)^2/16.56=6.73	(18-7.44)^2/7.44=14.99	24
grave	(132-121.44)^2/121.44=0.92	(44-54.56)^2/54.56=2.04	176
total	138	62	200

Is amber primary a grave good?

Df=1, Level of significance = 0.05;

χ^2=24,68; boundary value (df=1 and p=0.05): 3.84146

The difference in the distribution is significantly not by chance. Both variables are associated!

\(\chi^2\) test [14]

```
amber<-matrix(c(6,132,18,44),ncol=2)
colnames(amber)<-c("with amber","without amber")
rownames(amber)<-c("settlement","grave")
amber
```

```
##            with amber without amber
## settlement          6            18
## grave             132            44
```

```
chisq.test(amber)
```

```
##
##      Pearson's Chi-squared test with Yates' continuity correction
##
## data:  amber
## X-squared = 22, df = 1, p-value = 2e-06
```

\(\chi^2\) test [15]

```
chisq.test(amber)
```

```
##
##      Pearson's Chi-squared test with Yates' continuity correction
##
## data:  amber
## X-squared = 22, df = 1, p-value = 2e-06
```

```
chisq.test(amber,correct=F)
```

```
##
##      Pearson's Chi-squared test
##
## data:  amber
## X-squared = 25, df = 1, p-value = 7e-07
```

Correct: Yates correction for small samples → $(|O-E|-0,5)^2/E$

\(\chi^2\) excercise

Animal bones from middle and late neolithic strata in Wolkenwehe (Mischka et al. 2005)

The following counts are given

layer	Domestic animal	Wild animal
202 (late neolithic)	159	32
203 (middle neolithic)	84	54

Analyse if the observed differences are statistically significant!

Kolmogorov–Smirnov test

https://commons.wikimedia.org/wiki/File:Andrej_Nikolajewitsch_Kolmogorov.jpg

https://de.wikipedia.org/wiki/Datei:Smirnoff_vodka.jpg

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

- graeberbrz.csv

	rich	poor
infans I	6	23
infans II	8	21
juvenilus	11	25
adultus	29	36
maturus	19	27
senilis	3	4
Sum	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

	rich	rich_ratio	poor	poor_ratio
infans I	6	0.079	23	0.169
infans II	8	0.105	21	0.154
juvenilus	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
Sum	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

	rich	rich_ratio	rich_cumsum	poor	poor_ratio	poor_cumsum
infans I	6	0.079	0.079	23	0.169	0.169
infans II	8	0.105	0.184	21	0.154	0.324
juvenilus	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
Sum	76	1.000	3.263	136	1.000	3.743

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)

	rich_cumsum	poor_cumsum	difference
infans I	0.079	0.169	0.090
infans II	0.184	0.324	0.139
juvenilus	0.329	0.507	0.178
adultus	0.711	0.772	0.062
maturus	0.961	0.971	0.010
senilis	1.000	1.000	0.000

Find the 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

Formula:

\(boundary-value = f * \sqrt{\frac{n_1 + n_2}{n_1 * n_2} }\)

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: \(boundary-value = 1.36 * \sqrt{\frac{76 + 136}{76 * 136} } = 0.195\)

0.195 > 0.178, difference is not significant

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

Kolmogorov-Smirnov-Test [8]

KS-Test in R, prepare the dataset

```
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
head(alter)
```

```
## [1] 1 1 1 1 1 1
```

```
reichtum<-graeberbrz$reichtum
head(reichtum)
```

```
## [1] "reich" "reich" "reich" "reich" "reich" "reich"
```

Kolmogorov-Smirnov-Test [9]

KS-Test in R, the test itself

```
ks.test(alter[reichtum=="arm"],
        alter[reichtum=="reich"]
        )
```

```
## Warning in ks.test.default(alter[reichtum == "arm"], alter[reichtum ==
## "reich"]): p-value will be approximate in the presence of ties

##
##      Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  alter[reichtum == "arm"] and alter[reichtum == "reich"]
## D = 0.178, p-value = 0.09
## alternative hypothesis: two-sided
```

Excercise

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

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!

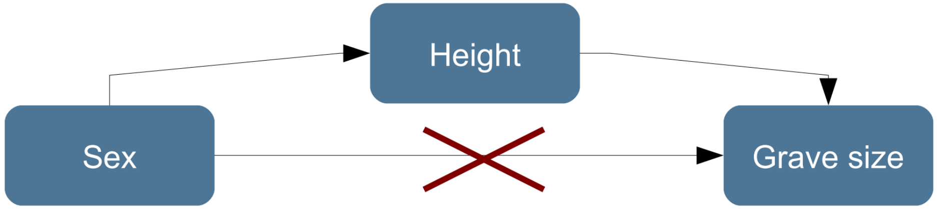
Statistical results stay statistical: significance is always probability that the choice of a hypothesis is correct, but there is also a probability that it is by chance...

Statistical association not mean causal association!

Example after Shennan: Grave size and sex

Although there is a statistically significant association between grave size and sex this could be caused by a third factor (here height)

A conclusion which says that grave size are determined by sex would be wrong!


Height
Sex
Grave size