UNIVERSITÄT BERN

Statistical methods for archaeological data analysis I: Basic methods

07 - Parametric tests

Martin Hinz

Institut für Archäologische Wissenschaften, Universität Bern

21.04.2021

Repetition: Population and sample [1]

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!

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

Parametric tests, advantages and disadvantages:

Advantage: Tests have general a higher power.

Disadvantages: Are generally not appropriate if no statements about the distribution are possible or the distribution fits no for parametric tests.

Also require usually a bigger samples size.

Possible requirements for parametric tests

Certain distribution

The data must follow certain distributions, i.e. originate from phenomenon of specific kind.

Example: t-Test - Normal distribution

Certain similarities

The data to compare must be equivalent in respect to certain parameters.

Example: ANOVA - variance homogenity

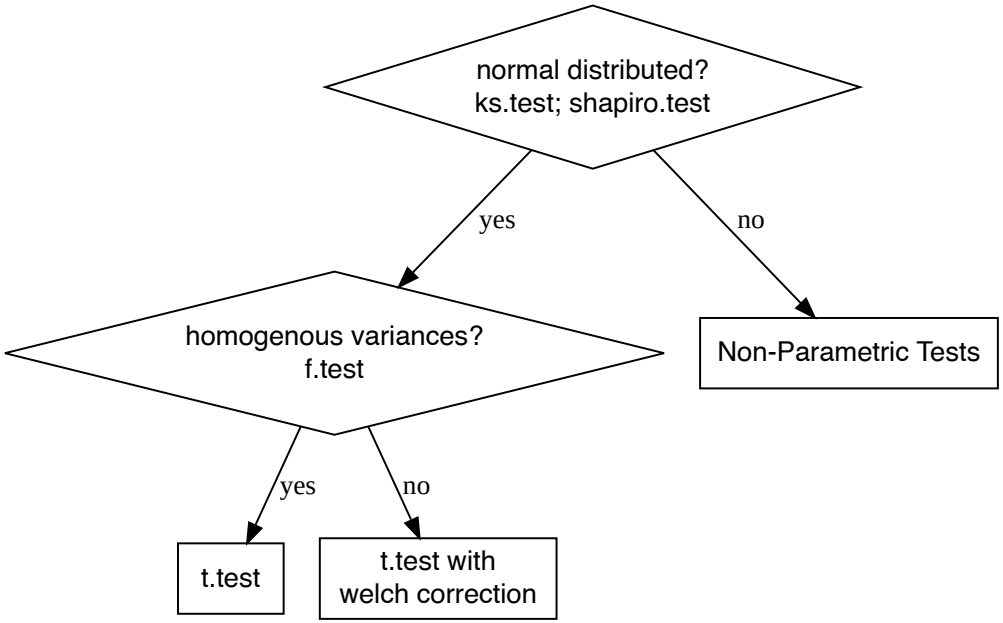
Certain scale level

Nearly in all cases at least one variable has to be interval scaled or higher

Example: F-Test - Test for variances, therefore at least interval scaled

(Deviation from the mean, which can only be calculated in case of interval or higher scaled data)

Possible test tree



Test for normal distribution

The good old KS-Test

eg. length of silex blades (simulated) - normal distributed?

```
blade_length<-c(14.9, 24.0, 8.7, 29.3, 25.5, 23.9, 22.4,
                12.7, 8.7, 25.1, 25.6, 14.7, 23.0, 23.2,
                26.5, 11.1, 15.2, 20.6, 20.1, 25.1)

ks.test(blade_length,"pnorm",mean(blade_length),sd(blade_length))
```

```
## Warning in ks.test(blade_length, "pnorm", mean(blade_length), sd(blade_length)):
## ties should not be present for the Kolmogorov-Smirnov test

##
##     One-sample Kolmogorov-Smirnov test
##
## data:  blade_length
## D = 0.19659, p-value = 0.4221
## alternative hypothesis: two-sided
```

Result is not significant, distribution does not deviate significant from normal distribution.

But: KS-test is rather conservative, Null hypothesis (normal distributed) will only be rejected if very strong deviation exists.

Test for normal distribution

The shapiro.test

The better test for normal distribution

requirements: $x_1 \dots x_n$ is a independend sample of a metric scaled variable

H_0 The population is normal distributed

H_1 the population is not normal distributed

```
shapiro.test(blade_length)
```

```
##
##     Shapiro-Wilk normality test
##
## data:  blade_length
## W = 0.90199, p-value = 0.04494
```

Test is just significant, so the distribution differs from normal distribution.

But: t-Test is rather robust (at least with sufficient sample sice), so we might proceed (but document the result!)

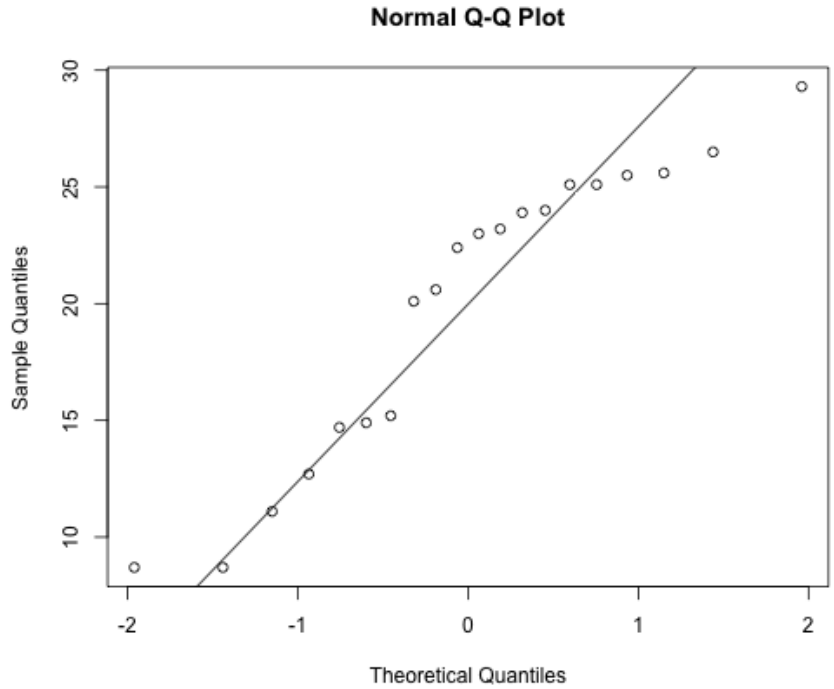
Test for normal distribution

visual: QQ-plot

Visual control of normal distribution

The quantile of the actual distribution is plotted against the quantile of a normal distribution.

```
qqnorm(blade_length)
qqline(blade_length)
```



Excercise test for normal distribution

Length of the handles of amphora of type Dressel 10 (Ihm 1978).

The length of the handles of different amphora are given. Test with the appropriate methods if the variable is normal distributed

- **file**: henkel_amphoren.csv

F-Test

F-Test [1]

test for variance homogenity of two samples

requirements: two independend normal distributed sample of a metric scaled variable

H_0 Both samples have the same variance (dispersion)

H_1 Both samples have a different variance (dispersion)

Basic idea: if both variances are equal, their quotient should be 1

$$s_1^2 = s_2^2;\ then\ \frac{s_1^2}{s_2^2} = 1$$

The quotient will be compared to a tabled threshold (eg. Shennan) according to degree of freedom ($df_1 = n_1 - 1; df_2 = n_2 - 1$) and the desired significance level.

If the calculated quotient > treshhold, than H_0 will be rejected, otherwise not.

Significant: unequal variances

not significant: we can assume homogenuous variances (for further tests)

F-Test [2]

example blade length

site 1; site 2

$n_1 = 20; \bar{x}_1 = 20.015$

$variance\ s_1^2 = \frac{\sum_{i=1}^{n} x_i - \bar{x}^2}{n-1} = \frac{\sum_{i=1}^{20} x_i - 20.015}{20-1} = 40.20871$

$n_2 = 25; \bar{x}_2 = 20.492$

$variance\ s_2^2 = \frac{\sum_{i=1}^{25} x_i - 20.492}{25-1} = 33.0641$

$F = \frac{s_1^2}{s_2^2} = \frac{40.20871}{33.0641} = 1.216$

df_1 = 20−1 = 19, df_2 = 25 − 1 = 24; Sign.level=0.05

threshold at df_1 = 19, df_2 = 24, α=0.05: 2.114

1.216 < 2.114; not significant, the variances do not differ significantly from each other

F-Test [3]

F-Test in R

blade_length.csv

```
blade_length <- read.csv("blade_length.csv")
var.test(length~site,data=blade_length)
```

```
##
##     F test to compare two variances
##
## data:  length by site
## F = 1.2161, num df = 19, denom df = 24, p-value = 0.643
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5185518 2.9822271
## sample estimates:
## ratio of variances
##           1.216084
```

Result is **not** significant, the variances are not significantly different

Excercise F-test

(logarithmic) sizes of ritual enclosures at the Society Islands (Example by Shennan)

Given are the (logarithmic) sizes of ritual enclosures in two valleys at the Society Islands.

Please check whether the variances in both valleys are different!

file: marae.csv

t-Test

t-Test, homogenuous variances [1]

Test for the comparison of the means of two samples.

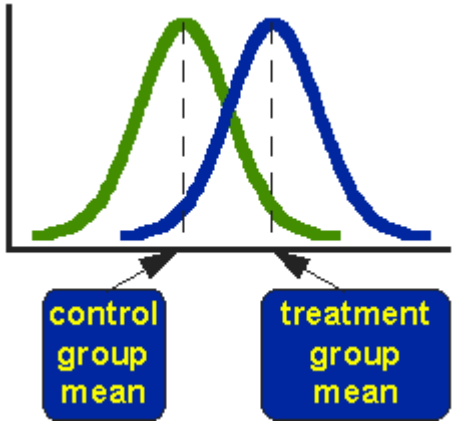
If the means do differ significantly, it is assumed that both samples come from different population (in the statistical sense).

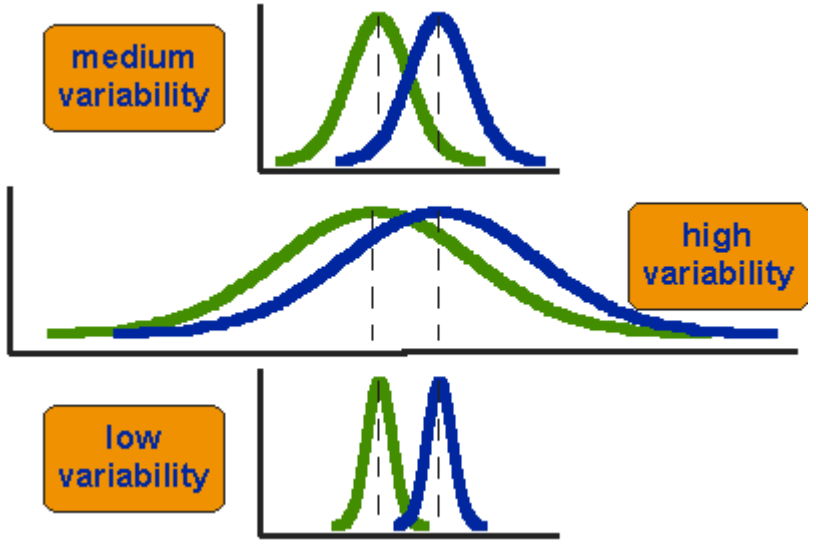
requirements: two independend normal distributed sample of a metric scaled variable with homogenous variances

H_0 The populations of both samples have the same mean

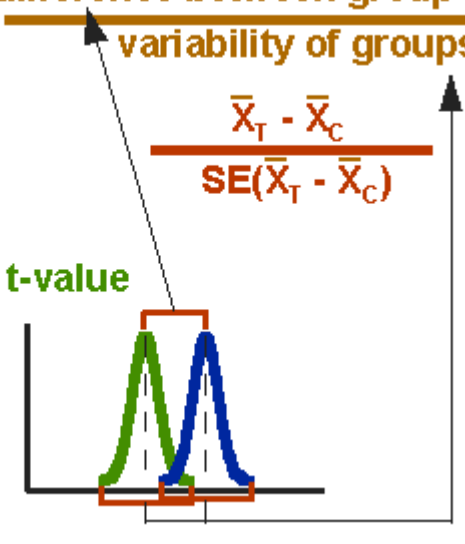
H_1 The populations of both samples have a different mean

Basic idea: If the means of both samples are within the standard error of the estimations of the mean of the according populations, than both populations could be potentially the same. Else not.





$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)} = \text{t-value}$$



t-Test, homogenuous variances [2]

Calculation 'by hand'

example blade length

site 1, site 2

$n_1 = 20; \bar{x_1} = 20.015; s_1^2 = 40.20871$
$n_2 = 25; \bar{x_2} = 20.492; s_2^2 = 33.0641$

$$t = \frac{difference\ between\ group\ means}{variability\ of\ groups}$$

$$t = \frac{\bar{x_1} - \bar{x_2}}{SE_{(\bar{x_1} - \bar{x_2})}}$$

$$SE = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$

$$SE_{(\bar{x_1} - \bar{x_2})} = \sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}$$

$$S^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}$$

$$S^2 = \frac{(20 - 1) * 40.20871 + (25 - 1) * 33.0641}{20 + 25 - 2} = 36.22102$$

$$SE_{(\bar{x_1} - \bar{x_2})} = \sqrt{\frac{36.22102}{20} + \frac{36.22102}{25}} = 1.805517$$

t-Test, homogenuous variances [3]

Calculation 'by hand'

$$t = \frac{20.015 - 20.492}{1.805517} = -0,26419$$

$$df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2 = 20 + 25 - 2 = 43$$

Looking up in table (eg. Shennan): df = 43; sig. level = 0.05

possible differences bigger - smaller, therefore two tailed question

threshold: 2.021

not significant, we can not reject the null hypothesis

There is no significant difference in the means of both samples, the could originate from the same population (statistically speaking).

t-Test, homogenuous variances [4]

Test for the comparison of the means of two samples.

If the means do differ significantly, it is assumed that both samples come from different population (in the statistical sense).

requirements: two independend normal distributed sample of a metric scaled variable with homogenous variances

H_0 The populations of both samples have the same mean

H_1 The populations of both samples have a different mean

```
t.test(length ~ site, data=blade_length, var.equal=T)
```

```
##
##     Two Sample t-test
##
## data:  length by site
## t = -0.26419, df = 43, p-value = 0.7929
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.118172  3.164172
## sample estimates:
## mean in group site1 mean in group site2
##              20.015              20.492
```

not significant, we can not reject the null hypothesis

Excercise t-test

Again the (logarithmic) sizes of ritual enclosures at the Society Islands (Example by Shennan)

Given are the (logarithmic) sizes of ritual enclosures in two valleys at the Society Islands.

Please check whether both valleys are different in respect to the mean sizes of the enclosures!

file: marae.csv

Welch-Test (t-Test for inhomogenous variances with welch correction)

If homogenity precondition is not fulfilled

approximation of the results by correction of degrees of freedom

blade_length3.csv

```
blade_length3 <- read.csv("blade_length3.csv")

var.test(length ~ site, data = blade_length3)
```

```
##
##     F test to compare two variances
##
## data:  length by site
## F = 2.8288, num df = 19, denom df = 19, p-value = 0.02854
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.119663 7.146758
## sample estimates:
## ratio of variances
##           2.828774
```

```
t.test(length ~ site, data = blade_length3)
```

```
##
##     Welch Two Sample t-test
##
## data:  length by site
## t = -1.8368, df = 30.941, p-value = 0.07586
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.394619  0.334619
## sample estimates:
## mean in group site1 mean in group site3
##              20.015              23.045
```

Welch-Test (t-Test for inhomogenous variances with welch correction)

Comparison of test power

Tests with lesser preconditions to the data often have lesser power

```
A <- rnorm(20)
B <- rnorm(20) + 0.5
```

```
t.test(A,B, var.equal=T)
```

```
##
##     Two Sample t-test
##
## data:  A and B
## t = -1.7675, df = 38, p-value = 0.08517
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2281221  0.0831989
## sample estimates:
##   mean of x   mean of y
## -0.08246632  0.48999529
```

```
t.test(A,B, data=sim_data)
```

```
##
##     Welch Two Sample t-test
##
## data:  A and B
## t = -1.7675, df = 35.995, p-value = 0.08562
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.22932376  0.08440054
## sample estimates:
##   mean of x   mean of y
## -0.08246632  0.48999529
```

In most of the cases (with small sample size), the Welch test will be slightly more unsure to reject the null hypothesis.

Multiple Tests [1]

The following is important for all statistical tests!

What, if one compares more than two groups?

Given are the hypothetical sizes of TRB sites in Schleswig-Holstein by region and wetness.

```
settlements <- read.csv("settlements.csv")
settlements$wetness <- factor(settlements$wetness)
head(settlements)
```

```
##     size                regions wetness
## 1 41.161              Steinburg   humid
## 2 26.497                   Kiel   humid
## 3 39.894           Dithmarschen   humid
## 4 17.783             Neumünster   humid
## 5 36.158 Rendsburg-Eckernförde   humid
## 6 27.181                Stormarn   humid
```

Question: Do the sizes of the settlements differ significantly in relation to the wetness?

How to proceed?

Multiple Tests [2]

Intuitive, but problematic answer: We test all groups against each other if there is a significant difference.

Problem: The more often we test, the more likely is a significant result 'by chance'.

Example: We test 3 groups, therefore we need 3 tests:

medium ↔ humid, humid ↔ arid, arid ↔ medium

The probability, that the alternative hypothesis is wrong even with significant result, is with each test 0.05. With three tests, it becomes 0.15!

With 100 tests, the expectation value becomes 5.0 (meaning, we expect 5 tests to show a wrong positive significance)!

Multiple Tests [3]

Solution 1, valid for all tests: we correct the p-values, eg. with the Bonferroni correction

The whole sequence of tests is regarded as one test. Therefore, the total p-value should be 0.05. Therefore, we devide for the individual tests the 0.05 by number of tests, to get the p-value for the individual tests.

Example: We test 3 groups, therefore we need 3 tests:

medium $\leftrightarrow$ humid: p-value = 0.2869584 humid $\leftrightarrow$ arid: p-value = 1.9394078×10^{-7} arid $\leftrightarrow$ medium: p-value = 3.4368554×10^{-5}

```
p.adjust(c(0.2869584, 1.939408e-07, 3.436855e-05), method = "bonferroni")
```

```
## [1] 8.608752e-01 5.818224e-07 1.031056e-04
```

The first result is not significant (admittedly, it was not before either).

ANOVA

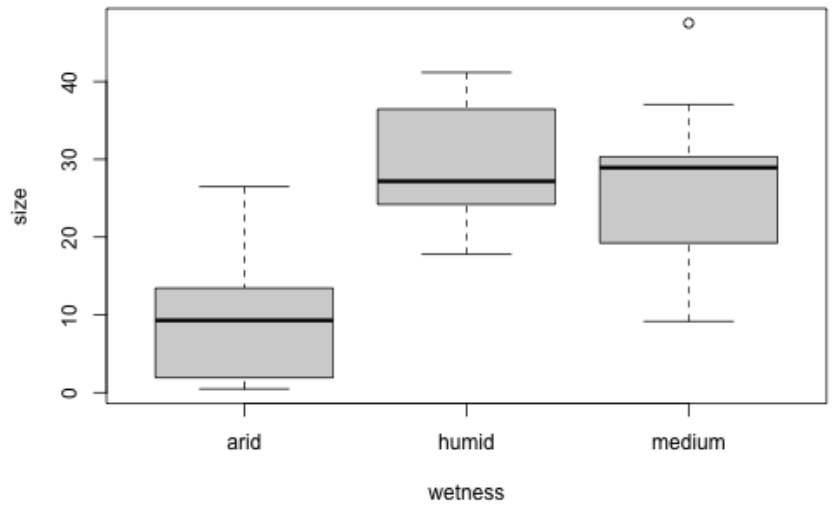
Multiple Tests [4]

ANOVA: Comparison of multiple groups

dependent variable: the variable to test (here size)

independent variable: the grouping variable (here wetness)

```
boxplot(size~wetness, data=settlements)
```




```
summary(aov(size~wetness, data=settlements))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## wetness      2   3564  1782.1   23.12 1.7e-07 ***
## Residuals   42   3238    77.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There are significant differences, but where?

ANOVA [2]

More detailled result:

We analyse the result as linear model

```
summary.lm(aov(size~wetness, data=settlements))
```

```
##
## Call:
## aov(formula = size ~ wetness, data = settlements)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.097  -7.252   1.236   4.627  21.244
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.388      2.267   4.141 0.000163 ***
## wetnesshumid    20.386      3.206   6.359 1.21e-07 ***
## wetnessmedium   16.880      3.206   5.265 4.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.78 on 42 degrees of freedom
## Multiple R-squared:  0.524,    Adjusted R-squared:  0.5013
## F-statistic: 23.12 on 2 and 42 DF,  p-value: 1.698e-07
```

The first group (arid) is the control group. The others are compared to this. If we are interested in the differences between medium and other soils, we have to change the controll group to 'medium'.

ANOVA [3]

changing the control group:

```
wetness.new<-relevel(settlements$wetness, ref ="medium")
summary.lm(aov(settlements$size~wetness.new))
```

```
##
## Call:
## aov(formula = settlements$size ~ wetness.new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.097  -7.252   1.236   4.627  21.244
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         26.268      2.267  11.587 1.16e-14 ***
## wetness.newarid    -16.880      3.206  -5.265 4.48e-06 ***
## wetness.newhumid     3.506      3.206   1.094     0.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.78 on 42 degrees of freedom
## Multiple R-squared:  0.524,    Adjusted R-squared:  0.5013
## F-statistic: 23.12 on 2 and 42 DF,  p-value: 1.698e-07
```

There is a significant difference between the control (medium) and arid soils.

Site note: Overview with multiple t-tests (shortcut!)

```
pairwise.t.test(settlements$size,settlements$wetness,
p.adjust="bonferroni")
```

```
##
##     Pairwise comparisons using t tests with pooled SD
##
## data:  settlements$size and settlements$wetness
##
##        arid    humid
## humid  3.6e-07 -
## medium 1.3e-05 0.84
##
## P value adjustment method: bonferroni
```

ANOVA [4]

two-factorial ANOVA

If we are interested in the influence of two grouping variables

```
boxplot(size~regions, data=settlements)
```

```
interaction.plot(settlements$wetness, settlements$regions,
settlements$size)
```

ANOVA [5]

```
summary(aov(size~wetness*regions, data=settlements))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## wetness          2   3564  1782.1  29.331 3.08e-06 ***
## regions         13   1372   105.6   1.738    0.142
## wetness:regions 12    832    69.4   1.142    0.391
## Residuals       17   1033    60.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notation in the formula

*: Interaction of the grouping variables are considered

+: grouping variables are considered as indipendent

ANOVA [5]

All the flavours of ANOVA

uni-factorial ANOVA

one independent factor (grouping variable), one dependent variable (measurement)

two-/multi-factorial ANOVA

two/multiple independent factor (grouping variable), one dependent variable (measurement)

multivariate ANOVA (MANOVA)

two/multiple independent factor (grouping variable), multiple dependent variable (measurements)