Experimental Design and Data Analysis, Lecture 3

Eduard Belitser

VU Amsterdam

Lecture overview

1. two paired samples (normal and not normal)
   - paired t-test
   - Pearson's correlation test
   - Spearman's rank correlation test
   - permutation test
2. two independent samples (normal and not normal)
   - two samples t-test
   - Mann-Whitney test
   - Kolmogorov-Smirnov test

two paired samples

Setting A

Setting:
An experiment with two numerical outcomes per experimental unit. Interest is in a possible difference between the two outcomes.

EXAMPLE Measurement of blood pressure of a person before and after a drug treatment.

EXAMPLE Comparing pain relief by a dedicated drug or by a placebo. Both treatments are applied to every individual (with recovery time in between, order assumed to have no effect).

EXAMPLE Comparing two car tire brands by putting both brands of tire on the same car and measuring the tires' wear.

Design A

- Take a random sample of experimental units from the relevant population.
- Measure the two outcomes on each unit.
- The two outcomes are related, because measured on the same experimental unit.
- The experiment should be set up so that any other type of "dependence" is eliminated and a difference in outcomes is due to the "treatment" only.

EXAMPLE If subjects must perform two tasks, then they should be allowed sufficient time between the tasks to recover and forget.

Remark. If a learning effect (the first measurement influences the second) is suspected, then, if possible, randomize the order of the two treatments within the units. The analysis must then follow the cross over design (studied later), not the paired samples design as discussed here.

Analysis A — paired t-test

- Data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$.
- The paired t-test: the differences $Z_1 = X_1 - Y_1, \ldots, Z_N = X_N - Y_N$ are assumed to be a random sample from a normal population $N(\mu, \sigma^2)$.
- We test the null hypothesis $H_0 : \mu = 0$.
- The test statistic is

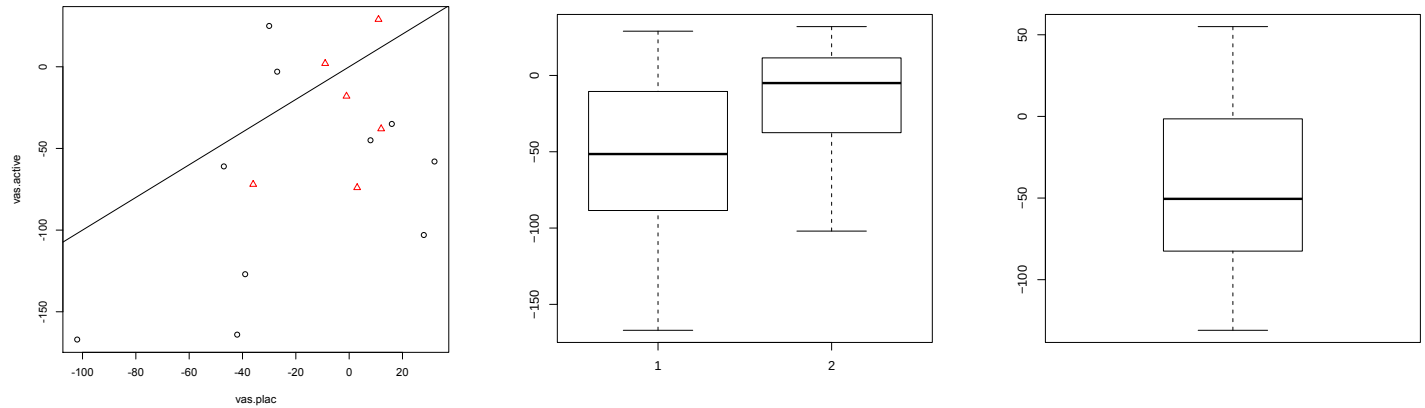
$$T = \frac{\bar{Z}_N}{S_N},$$

  where $\bar{Z}_N$ is the average of the differences $Z_i = X_i - Y_i$ and S_N is the sample standard deviation. Under H_0, T has the t_{N-1}-distribution.
- The analysis is simply a one sample analysis on the differences, and μ is the difference of the means of the X-population and the Y-population.

Analysis A in R — graphics

The rows of the data set `ashina.txt` correspond to 16 subjects and give measures of pain (for chronic headache) when treated with an active drug or a placebo.

```
> ashina=read.table("ashina.txt",header=TRUE); ashina
   vas.active vas.plac grp
1        -167     -102   1
2        -127      -39   1
[ some output deleted ]
16        -72      -36   2
> plot(vas.active~vas.plac,pch=grp,col=grp,data=ashina); abline(0,1)
> boxplot(ashina[,1],ashina[,2]); boxplot(ashina[,1]-ashina[,2])
```



(The third column of the data.frame `ashina` indicates the order of measurement (1=placebo first, 2=active first). This is used in the first plot (only) to determine the plotting character; this plot does not suggest that the order is important.)

Analysis A in R — estimation and testing (1)

The paired t-test:

```
> t.test(ashina[,1],ashina[,2],paired=TRUE)
        Paired t-test
data:  ashina[, 1] and ashina[, 2]
t = -3.2269, df = 15, p-value = 0.005644
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -71.1946 -14.5554
sample estimates:
mean of the differences
                -42.875
```

Conclusion: H_0 is rejected, mean of the differences is different from 0.

(A possible effect of the ordering of the measurements is ignored. Without the option `paired=TRUE` the function `t.test` with 2 arguments assumes that the 2 samples are independent.)

Analysis A in R — estimation and testing (2)

The one sample t-test applied to the differences:

```
> t.test(ashina[,1]-ashina[,2])
        One Sample t-test
data:  ashina[, 1] - ashina[, 2]
t = -3.2269, df = 15, p-value = 0.005644
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -71.1946 -14.5554
sample estimates:
mean of x
  -42.875
```

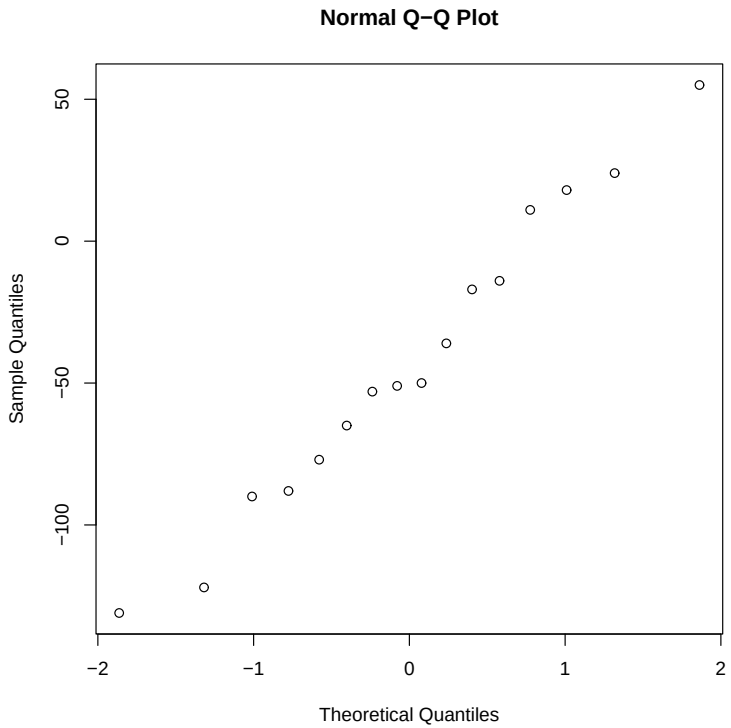
Conclusion: H_0 is rejected, mean of the differences is different from 0.

(With 1 argument the function `t.test` performs a one sample t-test. Applied to the differences this is equivalent to a paired two sample t-test – the shown p-values are identical.)

Analysis A in R — diagnostics

Check the normality assumption on the differences:

```
> qqnorm(ashina[,1]-ashina[,2])
> shapiro.test(ashina[,1]-ashina[,2])  ## gives $p$-value 0.9377
```



(No reason to suspect that the differences are not taken from a normal population.)

Setting and design B

Setting:
An experiment with two numerical outcomes (say X and Y) per experimental unit. Interest is in a possible dependence between the two outcomes per unit.

EXAMPLE Relation between shoe size and body mass index of a person.

EXAMPLE Relation between average course grade and number of students taking the course for courses at the VU.

EXAMPLE Relation between amount of precipitation and sun hours for different cities in Europe.

Design:

- Take a random sample of experimental units from the relevant population.
- Measure the two quantities on each unit. (The two outcomes are in principal related, because measured on the same experimental unit.)
- However, we possibly have measured unrelated quantities of the units and we want to test whether these quantities are correlated.

Analysis B1 — Pearson's correlation test

- Data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$.
- The Pearson correlation test assumes normality of the both X_i's and Y_i's. (Rather, the asympt. normality of the sample correlation $\hat{\rho}$.)
- The test is based on the sample correlation coefficient (which estimates of the "true" correlation $\rho = \mathrm{cor}(X, Y)$):

$$\hat{\rho} = \hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X}_N)(Y_i - \bar{Y}_N)}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X}_N)^2)\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2}}.$$

- We test the null hypothesis $H_0 : \rho = \rho_0 = 0$ that the correlation between the two populations is $\rho_0 = 0$. The test statistic is given by

$$T_\rho = \frac{\hat{\rho} - \rho_0}{\left(\frac{1-\hat{\rho}^2}{n-2}\right)^{1/2}} = \frac{\hat{\rho}}{\left(\frac{1-\hat{\rho}^2}{n-2}\right)^{1/2}},$$

which has under $H_0 : \rho = 0$ a t-distribution with $n-2$ degrees of freedom.

Analysis B2 — Spearman's rank correlation test

- Data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$.
- Spearman's rank correlation test does not assume normality. The test considers the ranks in the two samples, and compares the ordering of the ranks in the X_i and the Y_i. If the data are rank correlated, these sequences of ranks will run (approximately) in parallel or in opposite order.
- The test statistic is based on the correlation coefficient $\tilde{\rho}$ of the rank vectors.
- We test the null hypothesis $H_0 : \tilde{\rho} = 0$ that the rank correlation between the two populations is 0.

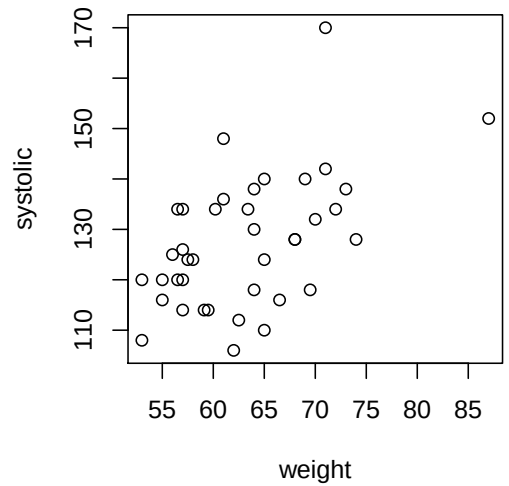
Analysis B in R — data input and graphics

Consider the data frame `peruvians.txt`, where the rows correspond to 39 men that moved from a native culture to a modern society. Amongst others, weight and systolic blood pressure were measured.

```
> peruvians=read.table("peruvians.txt",header=TRUE); peruvians
   age migration weight length chin  arm calf wrist systolic diastolic
1   21         1   71.0   1629  8.0  7.0 12.7    88      170        76
2   22         6   56.5   1569  3.3  5.0  8.0    64      120        60
      [ some output deleted ]
39  54        40   87.0   1542 11.3 11.7 11.3    92      152        88
```

```
> attach(peruvians)
> plot(systolic~weight)
```

Based on this picture, we expect dependence between `systolic` and `weight`.



Analysis B1 in R — estimation and testing

```
> cor.test(systolic,weight)

Pearson's product-moment correlation

data:  systolic and weight
t = 3.7164, df = 37, p-value = 0.0006654
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2463759 0.7186619
sample estimates:
      cor
0.5213643
```

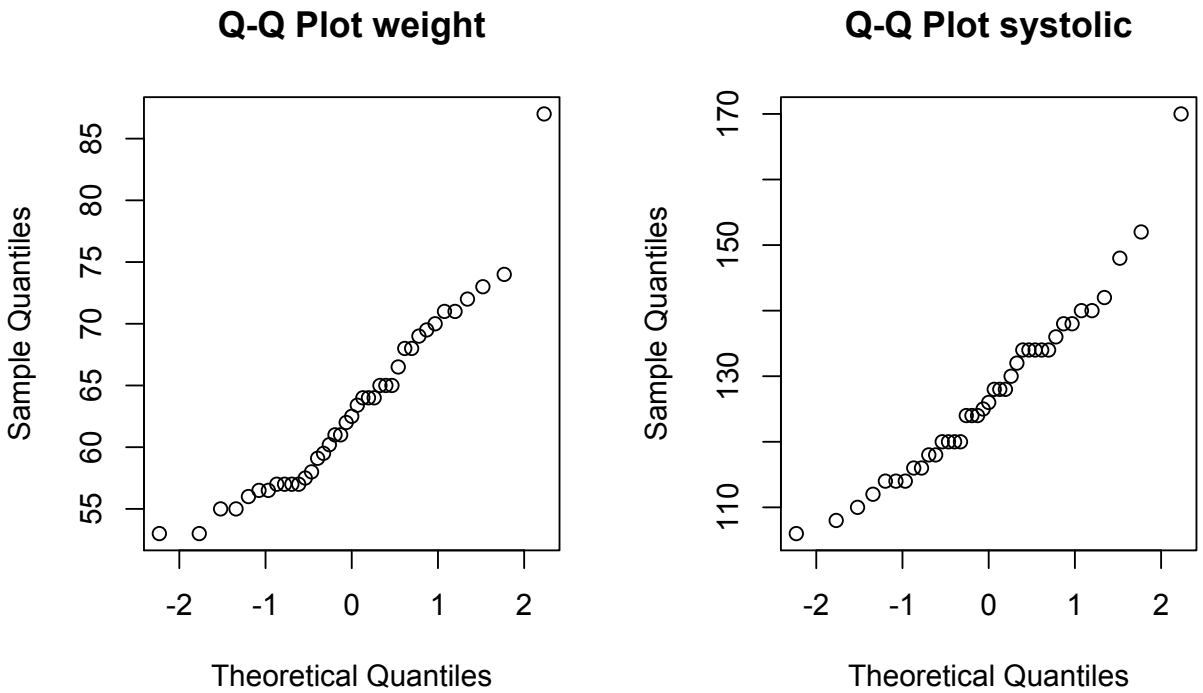
Conclusion: there is significant correlation, if normality is assumed.

The default for `cor.test` is Pearson's correlation test, based on normality.

Analysis B1 in R — diagnostics

Check the normality assumption on the two samples:

```
> qqnorm(weight,main="Q-Q Plot weight")
> qqnorm(systolic,main="Q-Q Plot systolic")
```



QQ-plots show that normality is not plausible for the weight sample. Hence, use the rank correlation test of Spearman.

Analysis B2 in R — estimation and testing

```
> cor.test(systolic,weight,method="spearman")

        Spearman's rank correlation rho

data:  systolic and weight
S = 5322.352, p-value = 0.003119
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4613004

Warning message:
In cor.test.default(systolic, weight, method = "spearman") :
  Cannot compute exact p-values with ties
```

Conclusion: there is indeed significant rank correlation.

(There is a warning about ties, which means that some values occur multiple times in `weight` and/or `systolic`. Therefore *R* uses an approximation for the *p*-value.)

permutation tests for two paired samples

Setting and design

Setting:

- An experiment with a numerical outcome measured according to two conditions per experimental unit;
- Interest is in a possible difference between the two outcomes per unit.

EXAMPLE Difference in average course grade for mathematical courses and informatics courses for BA-students at the VU.

EXAMPLE Difference in pain relief by an active drug and a placebo for patients.

Design (the standard paired samples design):

- Take a random sample of experimental units from the relevant population.
- Measure the two outcomes on each unit.

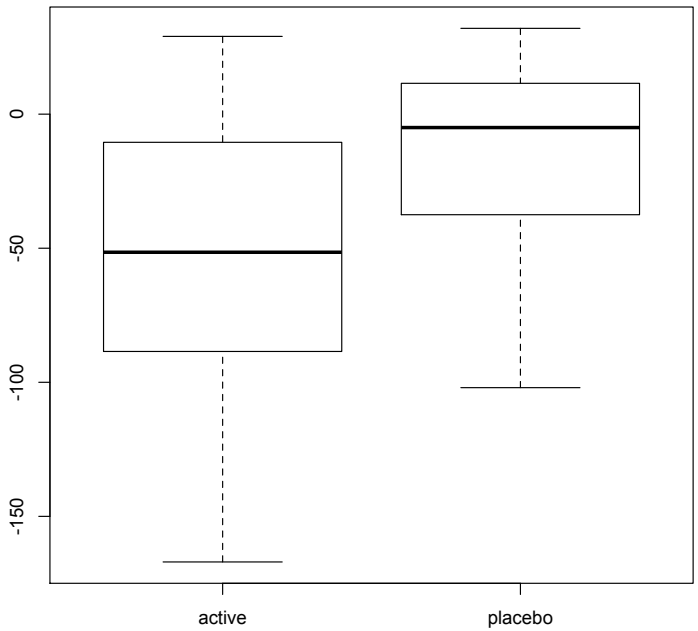
Analysis

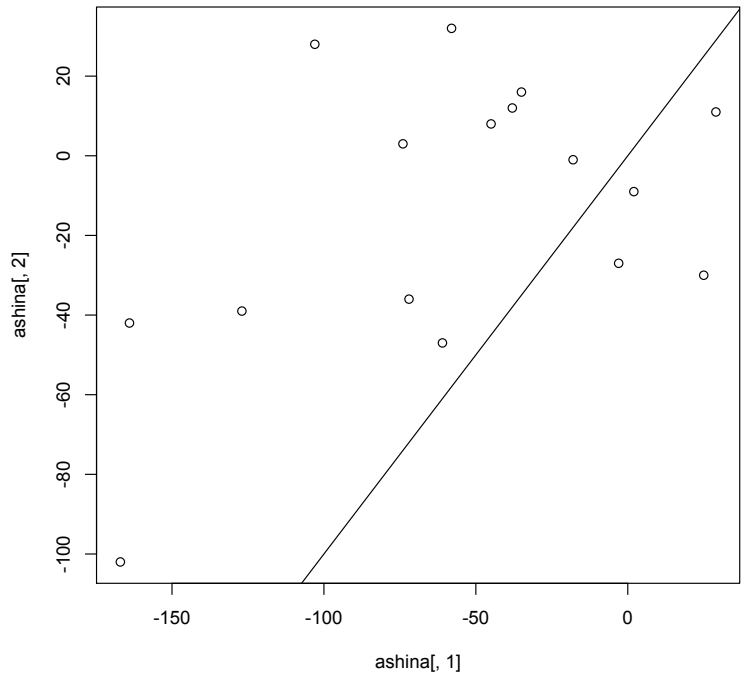
- Data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$.
- In a permutation test we do not assume normality.
- We can use any test statistic $T = T(X_1, Y_1, \ldots, X_N, Y_N)$ to test the null hypothesis of no difference between the distribution of X_i and that of Y_i within samples. The choice depends on the difference conjectured.
- Like in a bootstrap test, we simulate the distribution of T under H_0, using B surrogate T^*-values. Repeat B times (for $i = 1, \ldots, B$):
  - generate (X_j^*, Y_j^*) by generating a permutation of the original (X_j, Y_j) (relabeling) for $j = 1, \ldots, N$, i.e., choose between (X_j, Y_j) and (Y_j, X_j) with equal probability.
  - compute $T_i^* = T(X_1^*, Y_1^*, \ldots, X_N^*, Y_N^*)$
- Under H_0 of no difference between the distributions of X and Y within pairs permuting the labels does not change the distribution of T.

Analysis in R — data input and graphics

Recall dataset `ashina.txt` (drug or placebo against headache for 16 subjects).

```
> ashina=read.table("ashina.txt",header=TRUE)
> boxplot(ashina[,1],ashina[,2],names=c("active","placebo"))
> plot(ashina[,1],ashina[,2]); abline(0,1)
```





(Based on this picture we expect the active medicine to yield better pain relief.)

Analysis in R — testing (1)

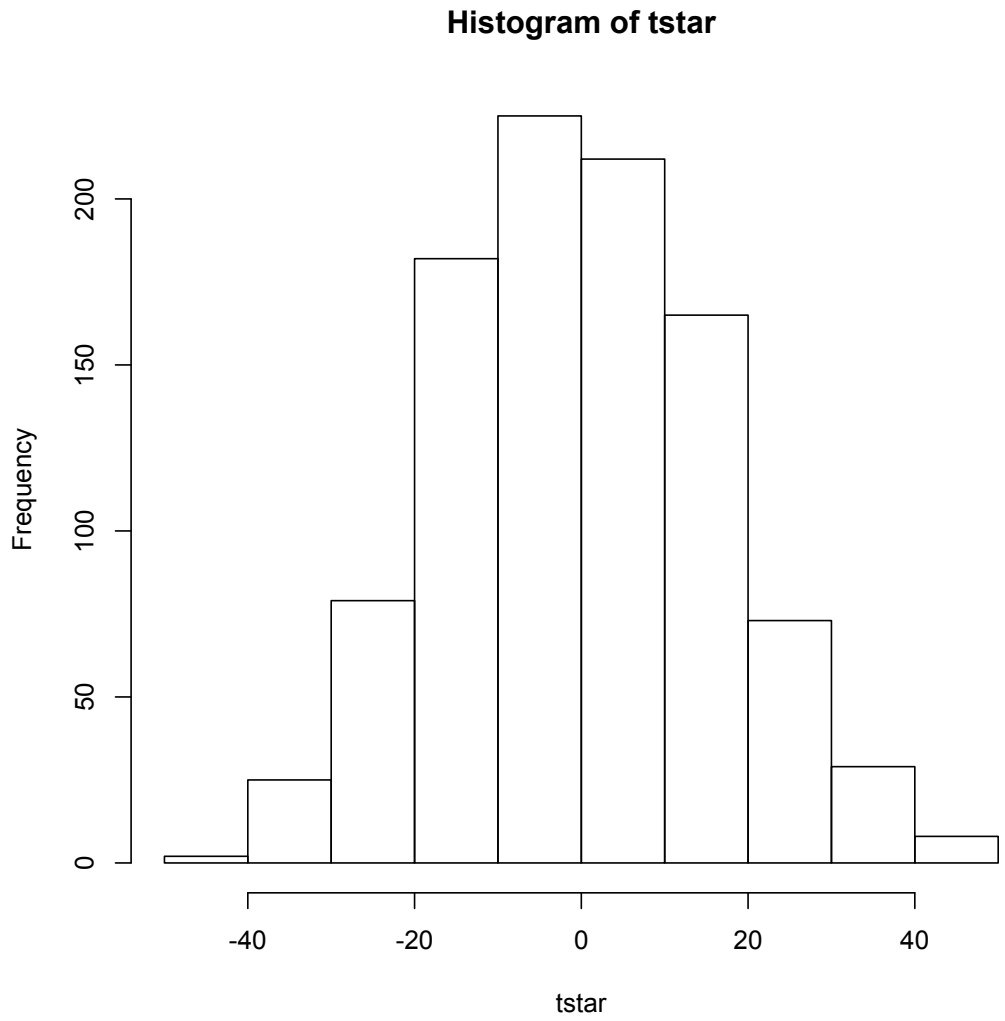
```
> mystat=function(x,y) {mean(x-y)}
> B=1000
> tstar=numeric(B)
> for (i in 1:B)
+ {
+   ashinastar=t(apply(cbind(ashina[,1],ashina[,2]),1,sample))
+   tstar[i]=mystat(ashinastar[,1],ashinastar[,2])
+ }
> myt=mystat(ashina[,1],ashina[,2])
```

(Instead of computing all $2^{16} = 65536$ possible permutations, we generate 1000 randomly chosen permutations to estimate the distribution of our test statistic under H_0. The function `apply` applies a function to either all rows or all columns in a `matrix`.)

Analysis in R — testing (2)

```
> myt
[1] -42.875
> hist(tstar)
> pl=sum(tstar<myt)/B
> pr=sum(tstar>myt)/B
> p=2*min(pl,pr)
> p
[1] 0.008
```

Conclusion: there is indeed a significant difference between the active drug and the placebo.



Discussion

- A permutation test for two paired samples can be performed with any test statistic that expresses difference between the X and Y within pairs. (The mean of differences $Z_i = X_i - Y_i$ is most common to consider, but one may as well consider the median of the Z_i's. Then the test is a bootstrap version of the sign test on the median of Z_i equal to 0.)
- Nonparametric alternatives to the permutation test for two paired samples are the sign test and the Wilcoxon signed rank test applied to the differences (cf. the previous lecture).

two independent samples

Setting and design

Setting: an experiment with

- one numerical outcome per experimental unit,
- two groups of experimental units.

Interest is in a possible difference between the two populations. medskip

EXAMPLE Comparing the weight of newborn children in two countries, The Netherlands and Chile.

EXAMPLE Measurement of total yield from an agricultural plot for two different fertilizers.

Design:

- Take a random sample of experimental units of size M from the first population and a random sample of size N from the second population;
- Measure the outcome on each unit.

The numbers M and N need not be the same. (Taking the number M and N equal is preferable since it maximizes the power of two sample tests.)

Analysis A: two samples t-test

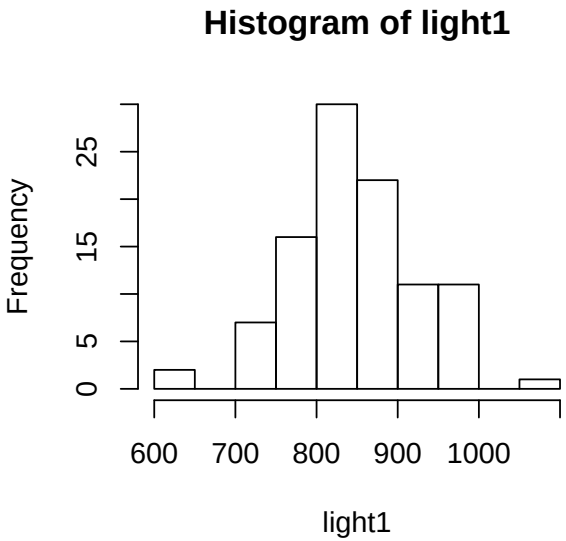
- Data $(X_1, \ldots, X_M)$ and $(Y_1, \ldots, Y_N)$.
- The two samples t-test assumes that both samples $X_1, \ldots, X_M$ and $Y_1, \ldots, Y_N$ come from a normal population. Denote the mean of the first population by μ and the mean of the second by ν.
- We test the null hypothesis $H_0 : \mu = \nu$ that the means of the populations are the same.
- The test statistic is

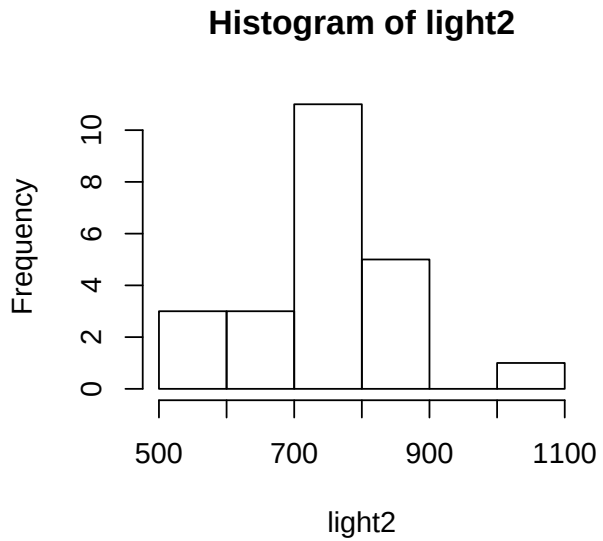
$$T = \frac{\overline{X}_M - \overline{Y}_N}{S_{N,M}}, \quad \text{which has the } t_{N+M-2}\text{-distribution under } H_0.$$

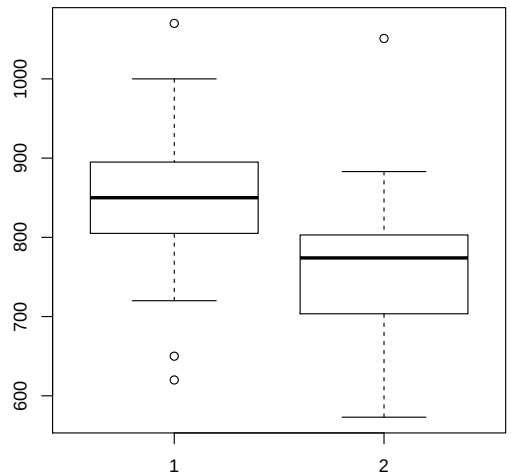
Analysis A in R — data input and graphics

Consider two data sets of measurements of the speed of light (minus 299000) by Michelson in 1879 and in 1882.

```
>light1=scan("light1.txt"); light2=scan("light2.txt")
> hist(light1); hist(light2); boxplot(light1,light2)
```







Analysis A in R — estimation and testing

The two samples t-test:

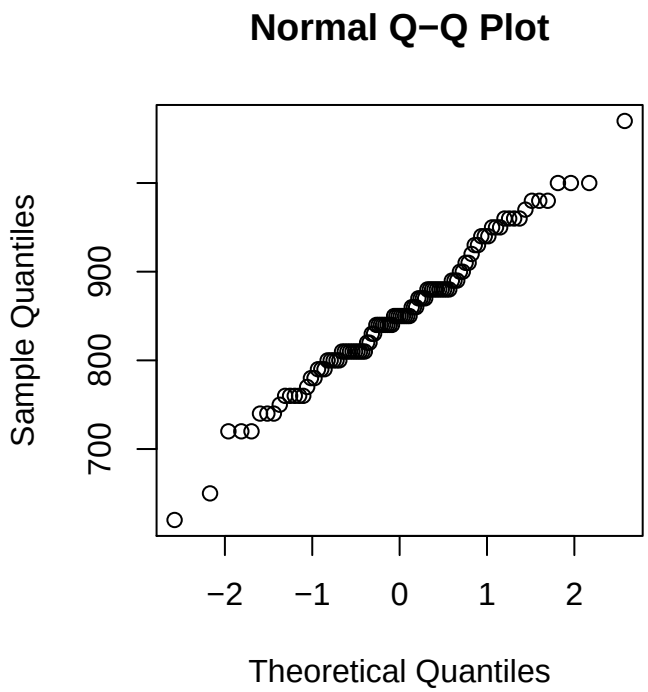
```
> t.test(light1,light2)

        Welch Two Sample t-test

data:  light1 and light2
t = 4.0598, df = 27.754, p-value = 0.0003625
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  47.63387 144.73135
sample estimates:
mean of x mean of y
 852.4000  756.2174
```

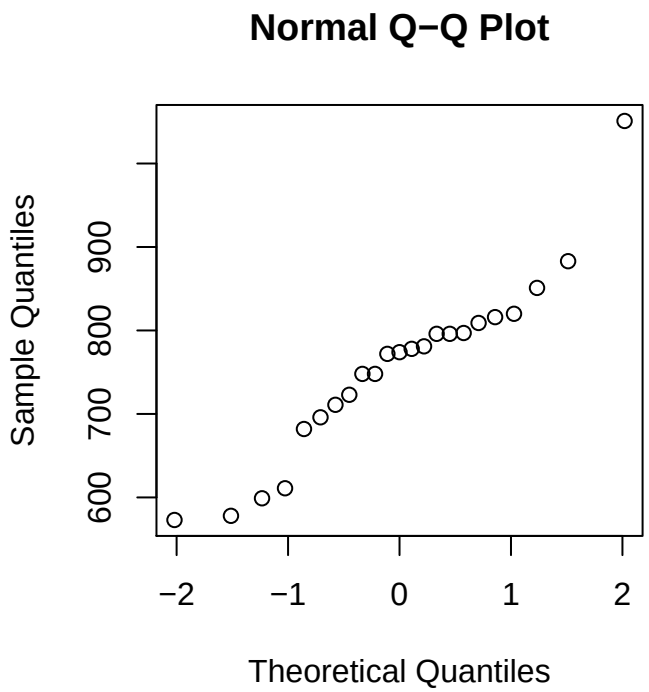
Conclusion: H_0 of equal means is rejected.

By default `t.test` with two arguments performs the two samples t-test for independent samples.

Analysis A in R — diagnostics

```
> qqnorm(light1)
> qqnorm(light2)
```





Normality of the second sample is actually doubtful.

Analysis B: the Mann-Whitney test

- Data $(X_1, \ldots, X_M)$ and $(Y_1, \ldots, Y_N)$.
- The Mann-Whitney test assumes that the sample $X_1, \ldots, X_M$ stems from population F and sample $Y_1, \ldots, Y_N$ stems from population G.
- We test the null hypothesis $H_0 : F = G$ that the populations are the same.
- The Mann-Whitney test is again based on ranks. It considers the M ranks $R_1, \ldots, R_M$ of $X_1, \ldots, X_M$ in the combined sample $(X_1, \ldots, X_M, Y_1, \ldots, Y_N)$ of length $M + N$. If $F = G$ these M rank numbers should lie randomly between 1 and $M + N$. The test statistic is

$$T = \sum_{i=1}^{M} R_i, \quad \text{the distribution of } T \text{ under } H_0 \text{ is (approximately) known.}$$

- Large values of T indicate that F is shifted towards the right from G, i.e. that X-values are bigger than Y-values.

If responses are continuous, a significant result of Mann-Whitney test shows a difference in medians, actually this test is only consistent against the alternative $H_1 : P(X > Y) \neq P(Y > X)$ (or $P(X > Y) + 0.5P(X = Y) \neq 0.5$).

Analysis B in R — testing

```
> wilcox.test(light1,light2)

        Wilcoxon rank sum test with continuity correction

data:  light1 and light2
W = 1829, p-value = 1.056e-05
alternative hypothesis: true location shift is not equal to 0
```

Conclusion: H_0 of equal medians is rejected. The underlying distribution of `light1` is shifted to the right from that of `light2`.

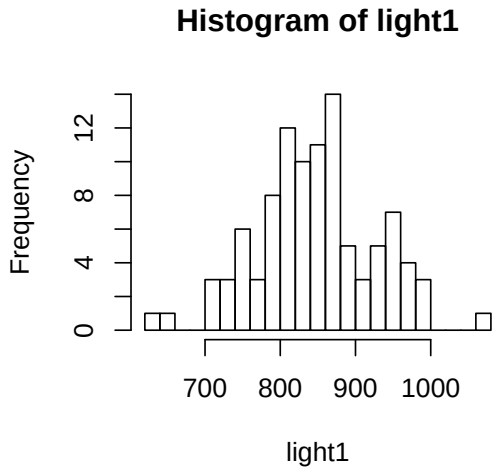
When given two arguments `wilcox.test` will perform the Mann-Whitney test for two samples. The Mann-Whitney test is especially suited for detecting shift differences — differences in location — between two populations.

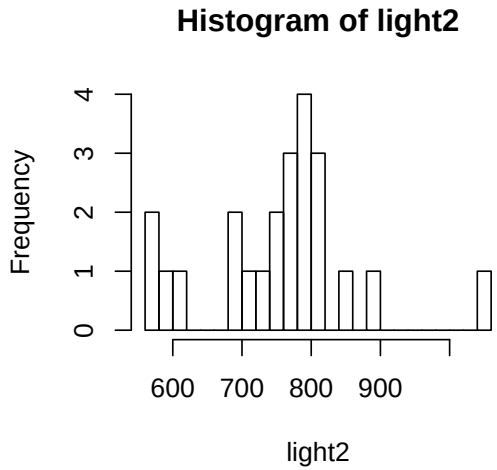
Analysis C: Kolmogorov-Smirnov test

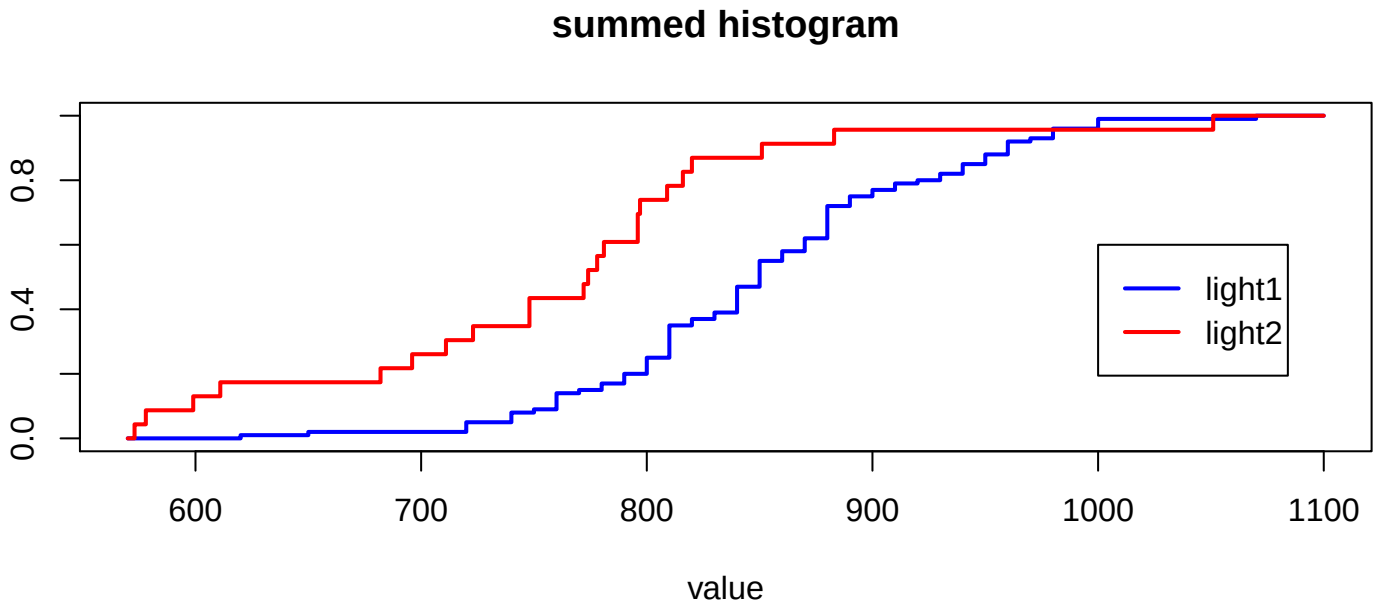
- Data $(X_1, \ldots, X_M)$ and $(Y_1, \ldots, Y_N)$.
- The Kolmogorov-Smirnov test assumes that the sample $X_1, \ldots, X_M$ stems from population F and sample $Y_1, \ldots, Y_N$ stems from population G.
- We test the null hypothesis $H_0 : F = G$ that the populations are the same.
- The Kolmogorov-Smirnov test is based on the differences in the histograms of the two samples.
- The test statistic computes the maximal vertical difference in summed histograms (empirical distribution functions). Its distribution under H_0 is known (e.g., in R).

Analysis C in R — graphics

```
> hist(light1)
> hist(light2)
```







Analysis C in R — testing

```
> ks.test(light1,light2)

        Two-sample Kolmogorov-Smirnov test

data:  light1 and light2
D = 0.5391, p-value = 3.803e-05
alternative hypothesis: two-sided

Warning message:
In ks.test(light1, light2) : cannot compute exact p-values with ties
```

Conclusion: H_0 of equal means is rejected. The mean of `light1` is larger.

(There is a warning about ties again. *R* uses an approximation for computing the *p*-value.)

To finish

Today we discussed: two samples tests; for paired and independent samples, and for normal and not normal cases.

Next time: k samples, one way ANOVA.