Experimental Design and Data Analysis, Lecture 4

Eduard Belitser

VU Amsterdam

Lecture overview

1. two paired samples
   - permutation test
2. two independent samples
   - two samples t-test
   - Mann-Whitney test
   - Kolmogorov-Smirnov test
3. k independent samples
   - Analysis of Variance (1-way ANOVA)
   - Kruskal-Wallis test

permutation tests for two paired samples

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

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

(This is the standard paired samples design.)

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.

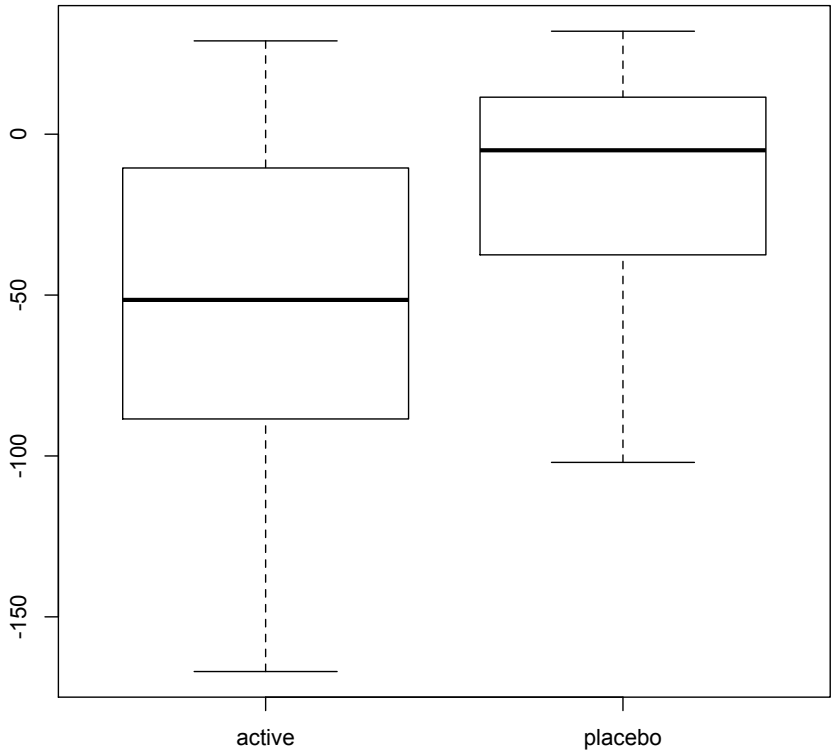
Analysis in R: data input

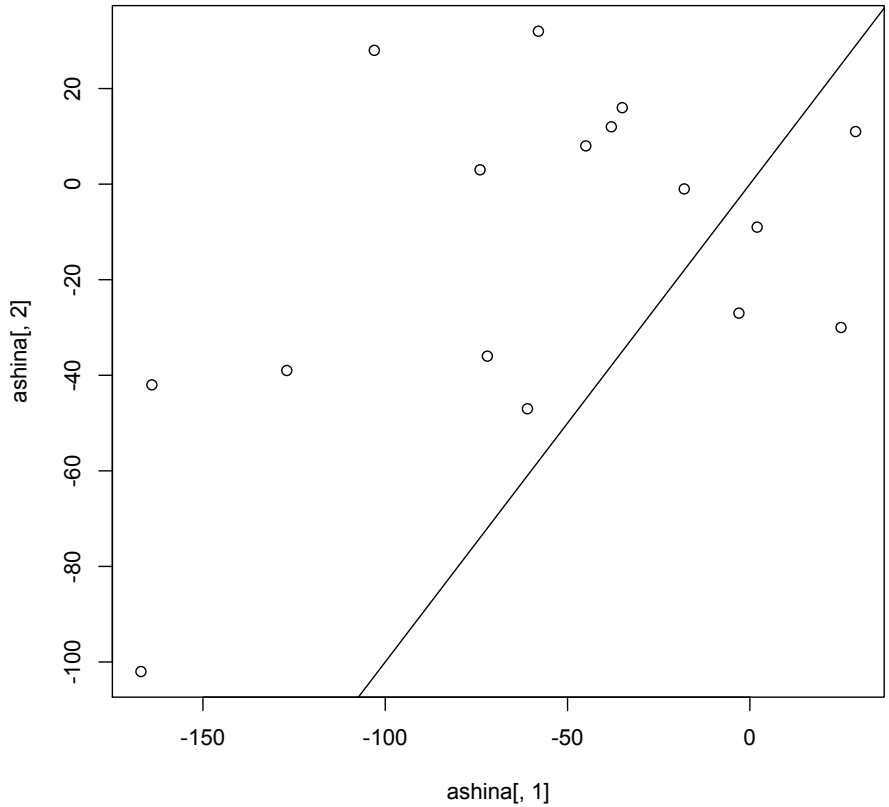
Create the two samples as parallel vectors, e.g. as two columns of a data frame.

```
> ashina=read.table("ashina.txt",header=TRUE)
> ashina
   vas.active vas.plac grp
1        -167     -102   1
2        -127      -39   1
3         -58       32   1
4        -103       28   1
5         -35       16   1
6        -164      -42   1
7          -3      -27   1
8          25      -30   1
9         -61      -47   1
10        -45        8   1
11        -38       12   2
12         29       11   2
13          2       -9   2
14        -18       -1   2
15        -74        3   2
16        -72      -36   2
```

Analysis in R: graphics

```
> 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])
```

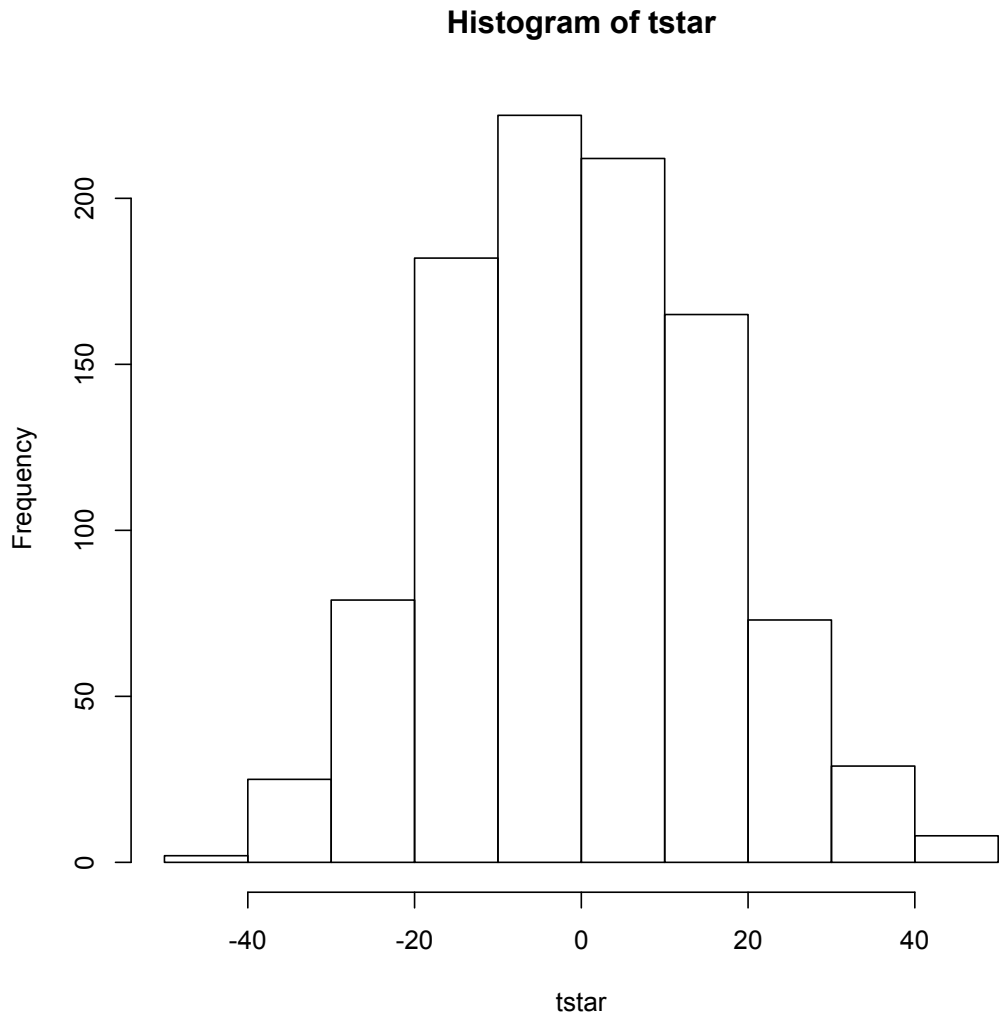
(Instead of computing all $2^{16} = 65536$ possible permutations, we generate 1000 randomly chosen permutations to estimate the distribution of our test statistic under H_0.

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 $Z_i = X_i - X_i$ 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. lecture 3).

two independent samples

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.

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

EXAMPLE Measurement of the time it takes to find a certain document in a web design for male and female users.

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

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

$$T = \frac{\overline{X}_M - \overline{Y}_N}{S_{N,M}}$$

which has the t_{N+M-2}-distribution under H_0.

We estimate the population means μ and ν.

Analysis A in R — data input

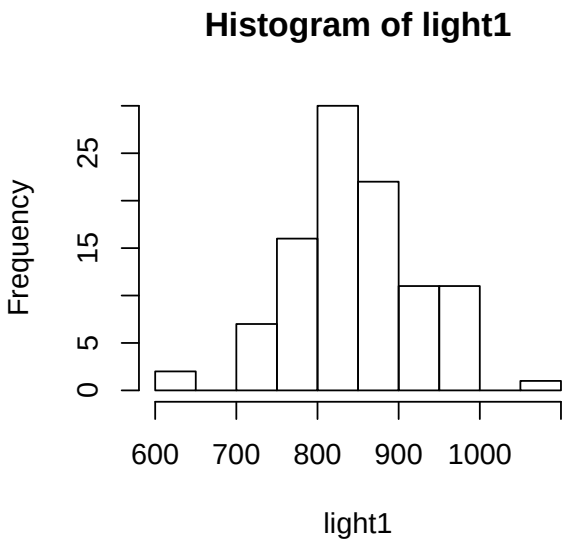
Create the two samples as two different vectors.

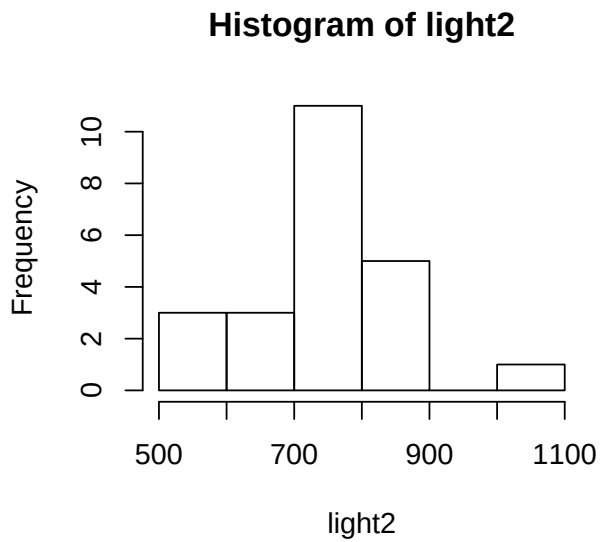
```
> light1=scan("light1.txt")
Read 100 items
> light2=scan("light2.txt")
Read 23 items
> light1
  [1]  850  740  900 1070  930  850  950  980  980  880
 [11] 1000  980  930  650  760  810 1000 1000  960  960
 [21]  960  940  960  940  880  800  850  880  900  840
 [31]  830  790  810  880  880  830  800  790  760  800
 [41]  880  880  880  860  720  720  620  860  970  950
 [51]  880  910  850  870  840  840  850  840  840  840
 [61]  890  810  810  820  800  770  760  740  750  760
 [71]  910  920  890  860  880  720  840  850  850  780
 [81]  890  840  780  810  760  810  790  810  820  850
 [91]  870  870  810  740  810  940  950  800  810  870
> light2
 [1]  883  816  778  796  682  711  611  599 1051  781
[11]  578  796  774  820  772  696  573  748  748  797
[21]  851  809  723
```

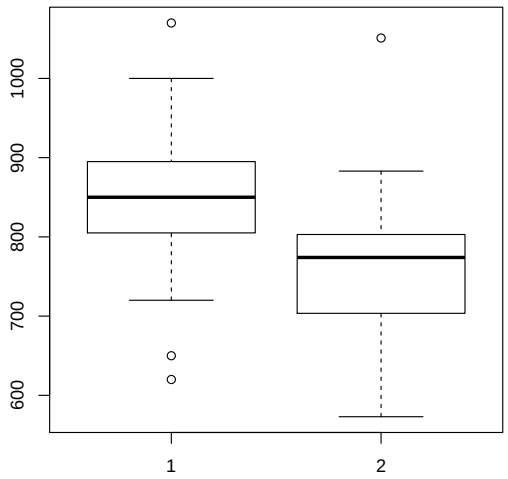
(The data are measurements of the speed of light (minus 299000) by Michelson in 1879 and in 1882.)

Analysis A in R — graphics

```
> hist(light1)
> hist(light2)
> boxplot(light1,light2)
```







Analysis A in R — estimation and testing

The two samples t-test:

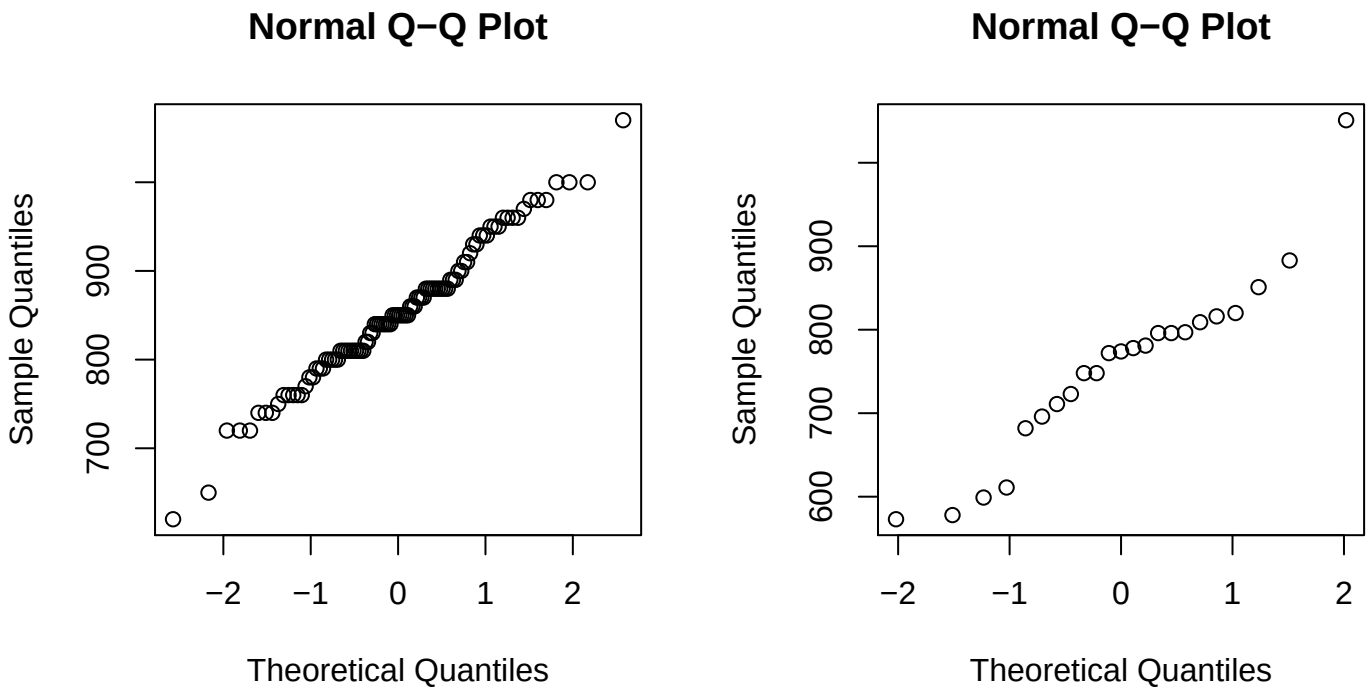
```
> t.test(light1,light2)

        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. The mean of `light1` is larger.

(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)
```



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

Analysis B

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

$$T = \sum_{i=1}^{M} R_i.$$

The distribution of T under H_0 is known (e.g. in R).

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

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 means 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

Data $(X_1, \ldots, X_M)$ and $(Y_1, \ldots, Y_N)$.

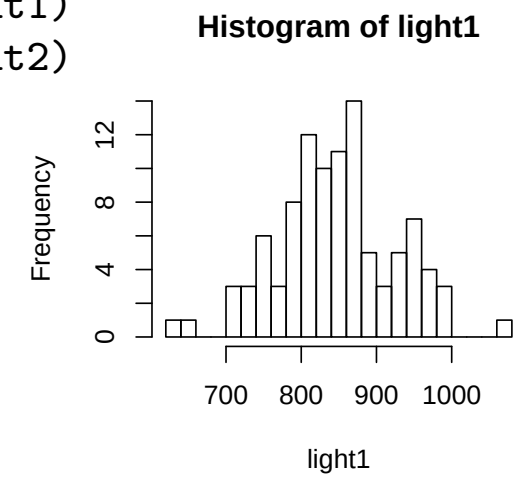
The Kolmogorov-Smirnov test assumes that the sample $X_1, \ldots, X_M$ stems from population F and sample $Y_1, \ldots, Y_N$ stems from population G.

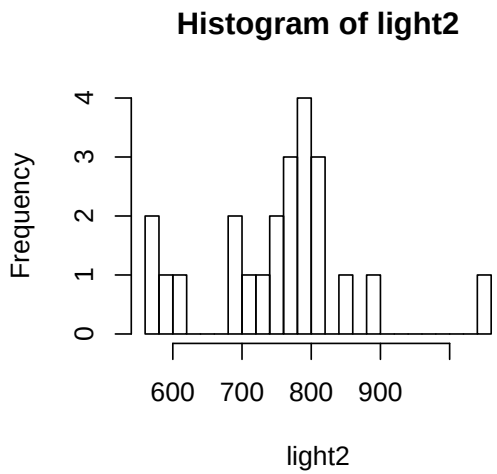
We test the null hypothesis $H_0 : F = G$ that the populations are the same.

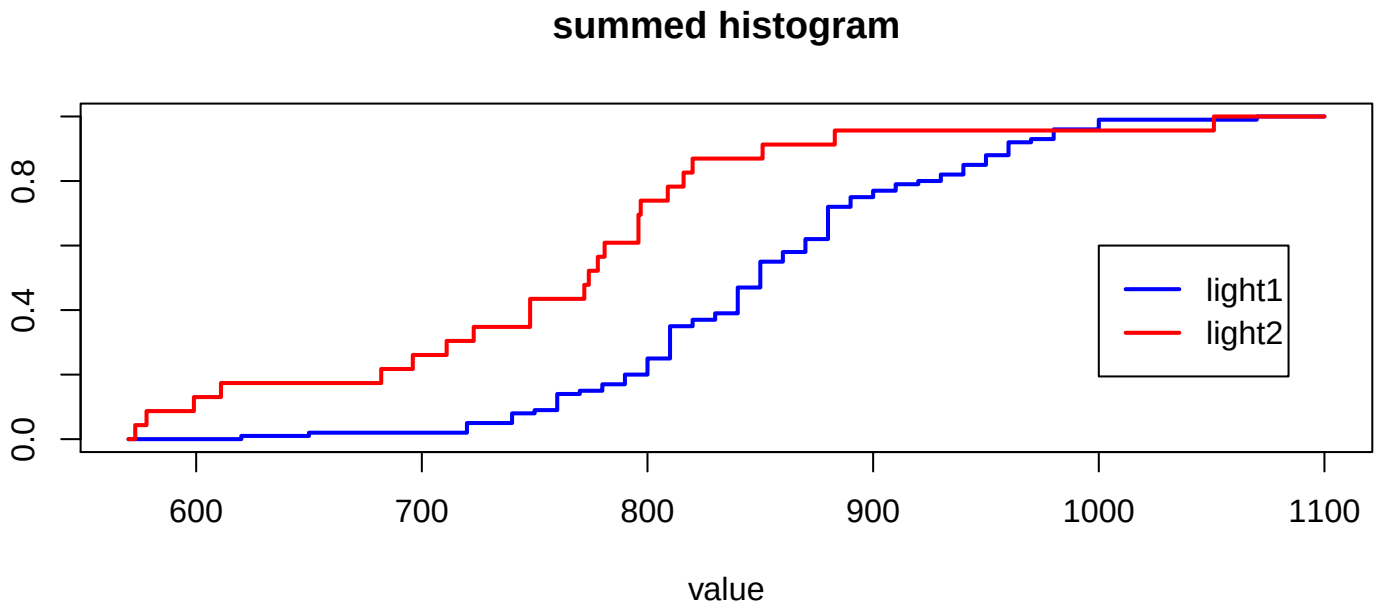
The Kolmogorov-Smirnov test is based on the differences in the histograms of the two samples. The test statistic computes the maximal vertical difference in *summed histograms* (empirical distribution functions). Its distribution under H_0 is known (e.g. in R).

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.)

one way analysis of variance
(completely randomized design)

Setting

An experiment with:

- a numerical outcome Y;
- a factor that can be fixed at I *levels* ("treatment").

EXAMPLE Agricultural experiment with outcome total yield from a plot and treatment type of fertilizer.

EXAMPLE Experiment where a subject must press a green or red button if there is a car in a picture shown on a screen, with outcome reaction time and treatment presence or not of an auditory stimulus.

EXAMPLE Quality of a genetic algorithm to determine the minimal value of a criterion function with outcome CPU time needed to find true minimum and treatment mutation probability set to 0.01, 0.02, 0.03, 0.04 or 0.05.

EXAMPLE Outcome time to develop mold on bread and treatment temperature of the environment fixed to 15, 19 or 22 degrees (garage, bedroom, living room).

If $I = 2$, this is just the two-sample problem, and we could perform a t-test.

Design

- Select NI experimental units randomly from the population of interest.
- Assign level i of the factor to a random set of N units ($i = 1, 2, \ldots, I$).
- Perform the experiment NI times, independently.

Randomization in R.

```
> I=4; N=5
> rep(1:I,N)
 [1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4
> sample(rep(1:I,N))
 [1] 3 4 2 1 1 4 3 4 3 1 3 2 3 2 1 4 2 4 2 1
```

Use level 3 for unit 1, level 4 for unit 2, etc.

(Using an equal number of units N for each level (called balanced design) is preferable, but not necessary.)

Analysis

Data

sample 1: $Y_{1,1}, Y_{1,2}, \ldots, Y_{1,N}$.
sample 2: $Y_{2,1}, Y_{2,2}, \ldots, Y_{2,N}$
$\vdots$
sample I: $Y_{I,1}, Y_{I,2}, \ldots, Y_{I,N}$

Assume that these are sampled independently from I normal populations with (possibly different) population means $\mu_1, \mu_2, \ldots, \mu_I$, and with equal population variances.

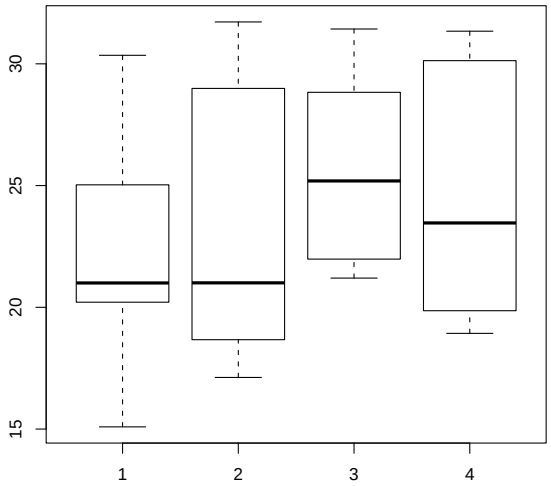
We test the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ versus the alternative $H_1 : \mu_i \neq \mu_j$ for some (i, j).

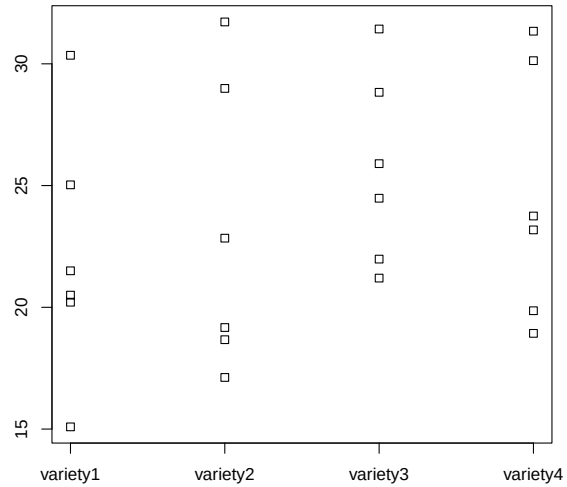
The test statistic is a bit complicated. It is, together with its distribution under H_0, implemented in R.

We estimate the means μ_i.

Analysis in R — graphics

```
> melon=read.table("melon.txt",header=TRUE)
> melon
  variety1 variety2 variety3 variety4
1    15.09    17.12    21.20    18.93
2    20.21    19.17    28.83    31.34
3    30.35    28.99    31.43    30.13
4    25.03    22.84    25.90    23.18
5    20.50    31.72    21.98    19.86
6    21.50    18.67    24.48    23.75
> boxplot(melon); stripchart(melon,vertical=TRUE)
```





Analysis in R — data input

Create a data-frame with each outcome $Y_{i,n}$ on a separate line and a second column that indicates the level of its factor.

```
> melon
  variety1 variety2 variety3 variety4
1    15.09    17.12    21.20    18.93
2    20.21    19.17    28.83    31.34
3    30.35    28.99    31.43    30.13
4    25.03    22.84    25.90    23.18
5    20.50    31.72    21.98    19.86
6    21.50    18.67    24.48    23.75
> melonframe=data.frame(yield=as.vector(as.matrix(melon)),
+           variety=factor(rep(1:4,each=6)))
> melonframe[1:5,]
  yield variety
1 15.09       1
2 20.21       1
3 30.35       1
4 25.03       1
5 20.50       1
> is.factor(melonframe$variety); is.numeric(melonframe$variety)
[1] TRUE
[1] FALSE
```

Analysis in R — testing

```
> melonaov=lm(yield~variety,data=melonframe)
> anova(melonaov)
Analysis of Variance Table
Response: yield
          Df Sum Sq Mean Sq F value Pr(>F)
variety    3  43.55  14.516  0.5543 0.6512
Residuals 20 523.73  26.186
```

(`lm` creates an object of type `linear model`. Its properties can be extracted with other functions.
`yield~variety` is a *model formula*. Read it as: "explain yield using variety".
The p-value for testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ is 0.6512: H_0 is not rejected.)

Analysis in R — estimation (1)

```
> summary(melonaov)
[ some output deleted ]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.1133     2.0891  10.585 1.21e-09 ***
variety2      0.9717     2.9545   0.329    0.746
variety3      3.5233     2.9545   1.193    0.247
variety4      2.4183     2.9545   0.819    0.423
```

By default R uses treatment contrasts: it takes the first level (here `variety1`) as a base level and compares the other levels to it.

(estimates: $\hat{\mu}_1 = 22.1133$; $\hat{\mu}_2 - \hat{\mu}_1 = 0.9717$; $\hat{\mu}_3 - \hat{\mu}_1 = 3.5233$; $\hat{\mu}_4 - \hat{\mu}_1 = 2.4.183$.
p-values: $(H_0 : \mu_1 = 0)$: 1.21-09; $(H_0 : \mu_2 = \mu_1)$: 0.746; $(H_0 : \mu_3 = \mu_1)$: 0.247; $(H_0 : \mu_4 = \mu_1)$: 0.423.)

Analysis in R — estimation (2)

```
> confint(melonaov)
                2.5 %    97.5 %
(Intercept) 17.755509 26.471158
variety2    -5.191228  7.134561
variety3    -2.639561  9.686228
variety4    -3.744561  8.581228
```

(95% confidence intervals: for μ_1: [17.755509, 26.471158]; for $\mu_2 - \mu_1$: [-5.191228, 7.134561], for $\mu_3 - \mu_1$: [-2.639561, 9.686228], for $\mu_4 - \mu_1$: [-3.744561, 8.581228].)

Analysis in R — estimation (3)

An alternative to the (default) treatment contrasts are sum contrasts. These give a decomposition of the population means into the overall mean μ and factor effects $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ as

$$\mu_i = \mu + \alpha_i, \qquad i = 1, 2, \ldots, I.$$

The effects are deviations from the mean; their average is zero: $\sum_i \alpha_i = 0$.

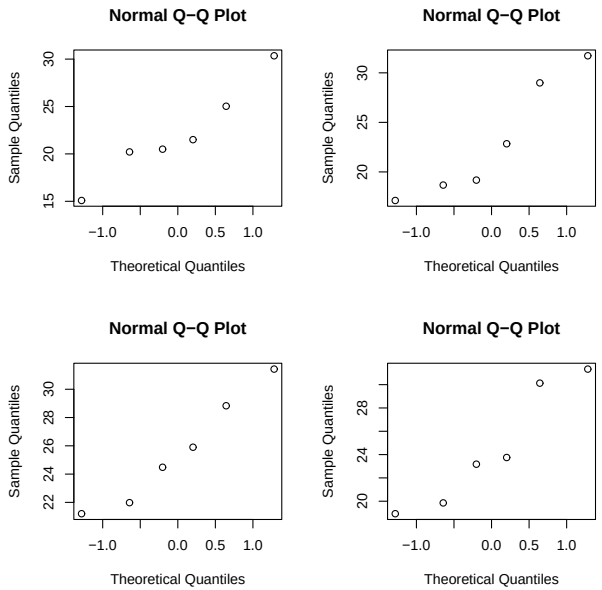
```
> contrasts(melonframe$variety)=contr.sum
> melonaov=lm(yield~variety,data=melonframe)
> summary(melonaov)
[ some output deleted ]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.8417     1.0446  22.825 8.55e-16 ***
variety1     -1.7283     1.8092  -0.955    0.351
variety2     -0.7567     1.8092  -0.418    0.680
variety3      1.7950     1.8092   0.992    0.333
```

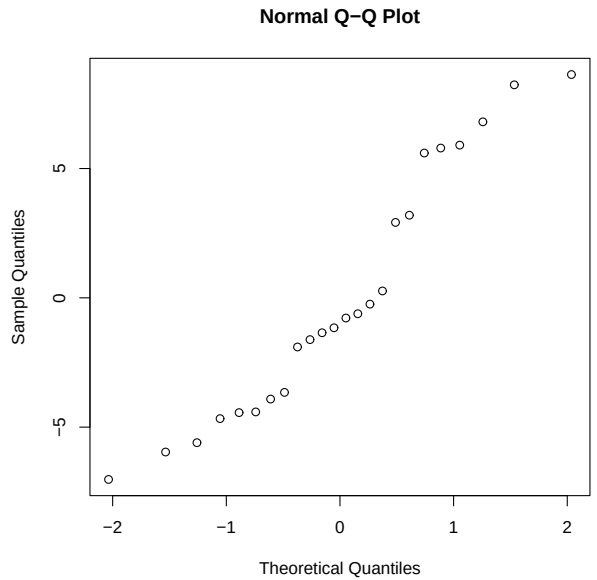
(The 4 lines of the table refer to $\mu, \alpha_1, \alpha_2, \alpha_3$. The 4th factor effect α_4 is omitted, but could be computed from $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$.)

Analysis in R — diagnostics

We can use the data to check whether the assumption of normality of the populations is not totally untrue.
The residuals $\hat{e}_{i,n} = Y_{i,n} - \hat{\mu}_i$ are the data corrected for the different population means and ought to look normal.

```
> par(mfrow=c(2,2)); for (i in 1:4) qqnorm(melon[,i])
> par(mfrow=c(1,1)); qqnorm(residuals(melonaov))
```





(Because the 4 samples are small, separate QQ-plots are not so useful. The second plot, using `residuals`, uses all 24 points, but corrected for being sampled from different populations.)

If the assumptions fail?

The design of the experiment ensures that the data are independent random samples from the populations.

However, the populations might be nonnormal or have different variances.

If the number of data points is large, then the p-value should still be accurate.

In the other case, consider:

- transforming the data (e.g. use $\log Y$) — see Assignment 3.
- using a different test.
- omit some (outlying) data-points (careful!).
- something else (there is no fix that always works).

Kruskal-Wallis test
a nonparametric test

Analysis

The Setting and Design are equal to the 1-way ANOVA case.

Data

sample 1: $Y_{1,1}, Y_{1,2}, \ldots, Y_{1,N}$.
sample 2: $Y_{2,1}, Y_{2,2}, \ldots, Y_{2,N}$
$\vdots$
sample I: $Y_{I,1}, Y_{I,2}, \ldots, Y_{I,N}$

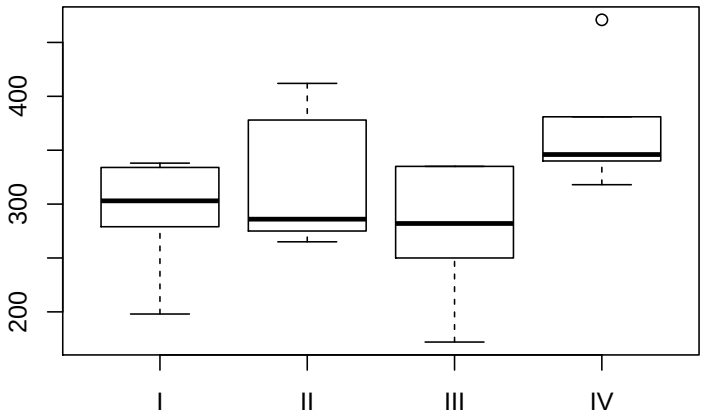
Assume that these are sampled independently from I populations $F_1, \ldots, F_I$ which are possibly different.

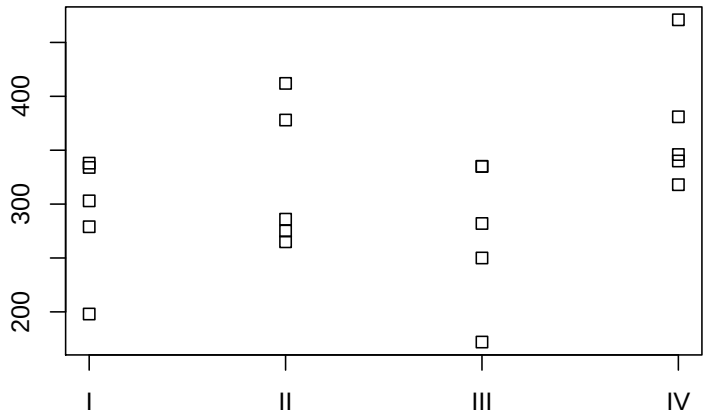
We test the null hypothesis $H_0 : F_1 = F_2 = \cdots = F_I$ versus the alternative $H_1 : F_i \neq F_j$ for some (i, j).

The Kruskal-Wallis test is a generalization of the Mann-Whitney test for 2 samples. It computes the sum of the ranks of $Y_{i,1}, \ldots, Y_{i,N}$ within the total data for each i. Under H_0 these N ranks should all lie randomly between 1 and NI.

Analysis in R — graphics

```
> ratdata=read.table("ratdata.txt",header=TRUE)
> ratdata
    I  II III  IV
1 279 378 172 381
2 338 275 335 346
3 334 412 335 340
4 198 265 282 471
5 303 286 250 318
> boxplot(ratdata); stripchart(ratdata,vertical=TRUE)
```





(Data are number of worms in rats in 4 different treatment groups.)

Analysis in R — data input

Create a data-frame with each outcome $Y_{i,n}$ on a separate line and a second column that indicates the level of its factor.

```
> ratdata
    I  II III  IV
1 279 378 172 381
2 338 275 335 346
3 334 412 335 340
4 198 265 282 471
5 303 286 250 318
> ratframe=data.frame(worms=as.vector(as.matrix(ratdata)),
+                     group=as.factor(rep(1:4,each=5)))
> ratframe[1:6,]
  worms group
1   279     1
2   338     1
3   334     1
4   198     1
5   303     1
6   378     2
> is.factor(ratframe$group); is.numeric(ratframe$group)
[1] TRUE
[1] FALSE
```

Analysis in R — testing (1)

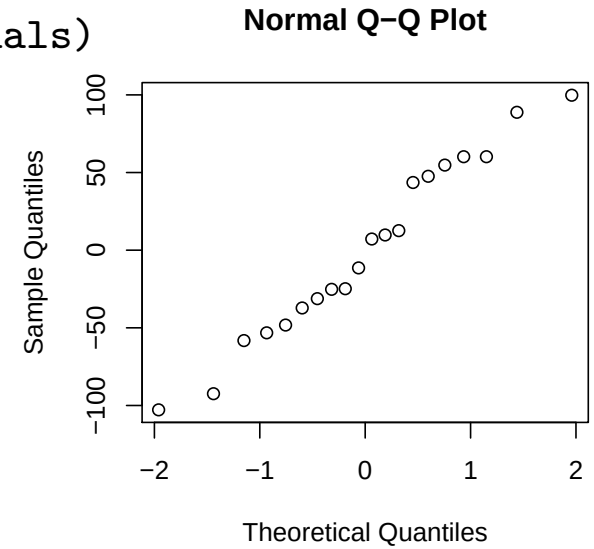
```
> attach(ratframe)
> kruskal.test(worms,group)

        Kruskal-Wallis rank sum test

data:  worms and group
Kruskal-Wallis chi-squared = 6.2047, df = 3, p-value = 0.1021
```

(`kruskal.test` performs the Kruskal-Wallis test and yields a p-value.
The p-value for testing $H_0 : F_1 = F_2 = F_3 = F_4$ is 0.1021: H_0 is not rejected.)

Analysis in R — testing (2)

Compare the ANOVA results:

```
> rataov=lm(worms~group)
> anova(rataov)
Analysis of Variance Table

Response: worms
          Df Sum Sq Mean Sq F value Pr(>F)
group      3  27234  9078.1  2.2712 0.1195
Residuals 16  63954  3997.1
> qqnorm(rataov$residuals)
```



(1-way ANOVA also doesn't yield a significant difference. The residuals don't seem to deviate significantly from normal, and both tests could be used here.)

to finish

To wrap up

Today we saw:

1. two paired samples
   - permutation test
2. two independent samples
   - two samples t-test
   - Mann-Whitney test
   - Kolmogorov-Smirnov test
3. k independent samples
   - Analysis of Variance (1-way ANOVA)
   - Kruskal-Wallis test

Next time: permutation test k samples, 2-way ANOVA, factorial design, multiple comparisons