Experimental Design and Data Analysis, Lecture 4

Eduard Belitser

VU Amsterdam

Lecture overview

1. Analysis of Variance (one-way ANOVA)
2. Kruskal-Wallis test
3. permutation tests in the setting of one-way ANOVA

one way ANOVA (analysis of variance)
completely randomized design

Setting

An experiment with:

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

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

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

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).

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.

One-way ANOVA

Data

sample 1: $Y_{11}, Y_{12}, \ldots, Y_{1N}$

sample 2: $Y_{21}, Y_{22}, \ldots, Y_{2N}$

$\vdots$

sample I: $Y_{I1}, Y_{I2}, \ldots, Y_{IN}$.

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

We want to test the null hypothesis $H_A : \mu_1 = \mu_2 = \cdots = \mu_I$ 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_A, implemented in R.

One-way ANOVA model

A categorical explanatory variable (also called factor) with I different categories/levels corresponds to I groups/populations/levels.

The one-way ANOVA model is: with $\mu_i = \mu + \alpha_i$,

$$Y_{ik} = \mu_i + e_{ik} = \mu + \alpha_i + e_{ik}, \quad i = 1, \ldots, I, \quad k = 1, \ldots, n_i,$$

- Y_{ik} is the k-th response measured in group i,
- μ is the common mean, α_i is the contribution of level i, $i = 1, \ldots, I$,

Assumption: the indep. errors $e_{ik} \sim N(0, \sigma^2)$, with unknown variance σ^2.

Balanced design: the same number of observations per group $n_i = N$, $i = 1, \ldots, I$, so that the total number of observations is $n = \sum_{i=1}^{I} n_i = NI$.

Note: if $I = 2$, this is the setting for the two sample t-test with equal variances.

Parameters $\mu, \alpha_1, \ldots, \alpha_I$ are not uniquely defined, one needs to specify one linear restriction on the parameters. Default parametrization in R is $\alpha_1 = 0$ (group 1 is the reference class). Other common parametrizations are $\mu = 0$ (then $\mu_i = \alpha_i$) or $\sum_{i=1}^{I} \alpha_i = 0$. The parametrization in R can be set by the command `contrasts`.

One-way ANOVA test

Setting: a one-way ANOVA model: $Y_{ij} = \mu + \alpha_i + e_{ij}$.

Hypotheses: $H_A : \alpha_1 = \ldots = \alpha_k = 0$ (no factor effect) versus H_1 : at least one $\alpha_i \neq 0$ (factor effect is present).

Test statistic: with $\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ik}$ and $\bar{Y}_{\cdot\cdot} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ik}$, under H_0,

$$F = \frac{\text{between-groups SS}}{\text{within-groups SS}} = \frac{\sum_{i=1}^{I} n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2/(n-I)} \sim F_{I-1,n-I},$$

the F-distribution with $I-1$ and $n-I$ degrees of freedom.

Larger values of $F = f$ give more evidence against H_0 in favor of H_1, hence we only reject H_A if F is large. The test is therefore always right-sided: compare the p-value $p_{right} = P(F > f)$ with a significance level α.

In R: the p-value is in `anova(lm(y~f),data=...)`, `f` is the factor.

In R: `summary(lm(y~f,data=...))` shows the coefficient estimates $\hat{\alpha}_i$'s in the `treatment` parameterization, to get these in the `sum` parametrization use (before `lm` command) `contrasts(f)=contr.sum`.

One-way ANOVA table

One-way ANOVA results are usually presented in an one-way ANOVA table:

Source	Df	Sum Sq	Mean Sq	F value	p-value
Factor A	$I-1$	SS_A	$SS_A/(I-1)$	$f = \frac{SS_A/(I-1)}{RSS/(n-I)}$	$P(F > f)$
Residuals	$n-I$	RSS	$RSS/(n-I)$		
Total	$n-1$	SS_T			

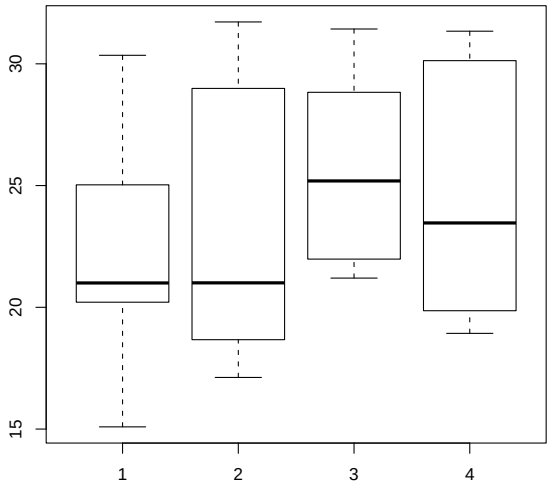
Here $RSS = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$, $SS_A = \sum_{i=1}^{I} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$, $SS_T = RSS + SS_A$.

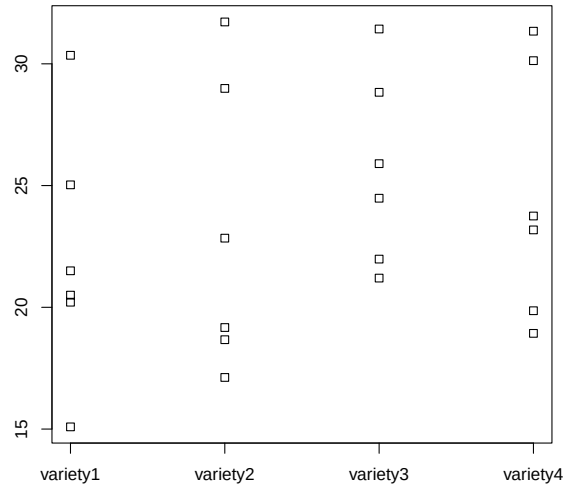
A one-way ANOVA table in R looks as follows:

```
            Df   Sum Sq   Mean Sq   F value   Pr(>F)
Factor      --   -------   -------   -------   ------
Residuals   --   -------   -------
```

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

If needed, create a data frame with a numeric column of responses $Y_{i,n}$ and a second factor column of the corresponding factor levels.

```
> 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)))  #create a data frame in the right format
> 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
```

The command `lm` creates an object of type `linear model` (many things can be extracted from it by using other functions), `yield~variety` is a model formula. Read it as: "explain yield using variety". The p-value for $H_A : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ (which is the same as $H_A : \mu_1 = \mu_2 = \mu_3 = \mu_4$) is 0.6512, hence H_A is not rejected, i.e., factor `variaty` is not significant.

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 parametrization, i.e., $\alpha_1 = 0$. In this case, R reports the estimates of $\mu_1 = \mu + \alpha_1 = \mu_1$, $\alpha_2 = \mu_2 - \mu_1$, $\ldots$, $\alpha_I = \mu_I - \mu_1$.

Thus, in the treatment contrasts, R takes the first level (here `variety1`, in alphabetical order) as a base level and compares the other levels to it. These estimates are $\hat{\mu}_1 = 22.1133$, $\hat{\mu}_2 - \hat{\mu}_1 = 0.9717$, $\hat{\mu}_3 - \hat{\mu}_1 = 3.5233$, $\hat{\mu}_4 - \hat{\mu}_1 = 2.4.183$. Then $\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$, $i = 1, \ldots, 4$, are just the group means and can also be obtained by command `fitted(melonaov)`. The column `Pr(>|t|)` gives the *p*-values for testing $\mu_1 = 0$ and $\alpha_i = \mu_i - \mu_1 = 0$, $i = 2, 3, 4$, respectively.

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
```

The 95% confidence intervals are 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 parametrization is sum parametrization. This gives a decomposition of the population means into the overall mean μ and factor effects $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ as

$$\mu_i = \mu + \alpha_i, \qquad i = 1, 2, \ldots, I, \quad \text{with the restiction } \sum_{i=1}^{I} \alpha_i = 0.$$

α_i's are expressing the deviations from the mean, and their average is zero.

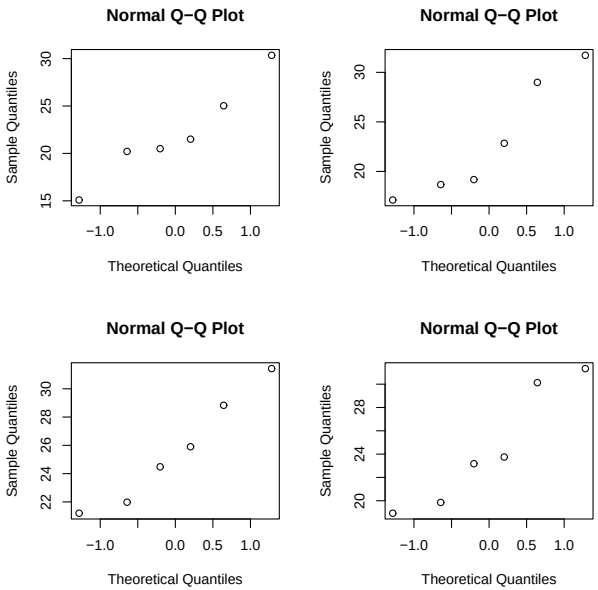
```
> contrasts(melonframe$variety)=contr.sum  #to specify sum-parametrization
> melonaov=lm(yield~variety,data=melonframe); summary(melonaov)
[ some output deleted ]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23. 8417     1.0446  22.825 8.55e-16 ***
variety1     -1.7283     1.8092  -0.955    0.351
variety2     -0.7567     1.8092  -0.418    0.680
variety3      1.7950     1.8092   0.992    0.333
```

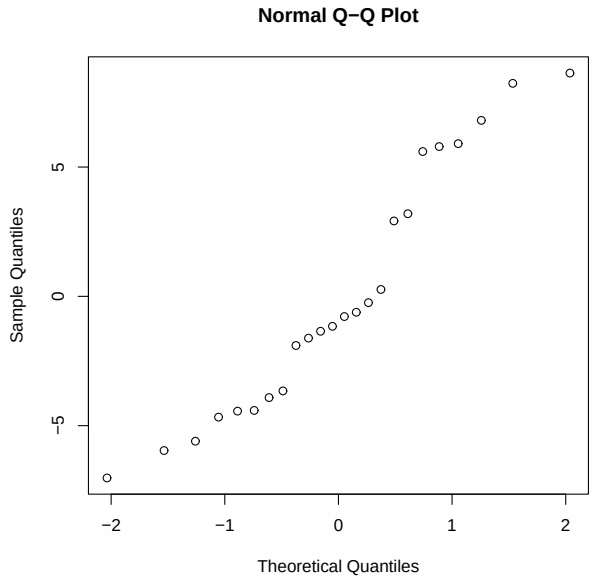
The 4 lines of the table give estimates of $\mu, \alpha_1, \alpha_2, \alpha_3$, now in sum-parametrization. The estimate for α_4 is omitted, but could be computed from $\sum_{i=1}^{4} \hat{\alpha}_i = 0$. We can compute the estimates for the μ_i's: $\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$, $i = 1, \ldots, 4$ (they must be the same as before).

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 $\log Y$);
  - using a different test;
  - omit some (outlying) data-points (careful!);
  - something else (there is no fix that always works).

Kruskal-Wallis test
(a nonparametric counterpart of ANOVA test)

Kruskal-Wallis test: design

The setting and design are the same as in the 1-way ANOVA (consider $n_i = N$, the balanced design). What if the normality assumption fails?

The Kruskal-Wallis test

- does not rely on the normality, it is based on ranks;
- is a nonparametric alternative to one-way ANOVA,
- is a generalization of the Mann-Whitney test for 2 samples;
- computes the sum of the ranks of $Y_{i,1}, \ldots, Y_{i,N}$ for each i within the total data. Under H_0 these N ranks should all lie randomly between 1 and NI.

Data

sample 1: $Y_{11}, Y_{12}, \ldots, Y_{1N}$
sample 2: $Y_{21}, Y_{22}, \ldots, Y_{2N}$
$\vdots$
sample I: $Y_{I1}, Y_{I2}, \ldots, Y_{IN}$

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

We test $H_0 : F_1 = \ldots = F_I$ versus H_1 : at least two distributions are different.

Kruskal-Wallis test: setting and analysis

Setting: measurements Y_{ik} for $i = 1, \ldots, I$ and $k = 1, \ldots, n_i$ from I different populations, Y_{ik} follows distribution F_i of population i.

Hypotheses: $H_0 : F_1 = \ldots = F_k$ versus $H_1 : F_i \neq F_j$ for some i, j.

Test statistic: $W = \frac{12}{n(n+1)} \sum_{i=1}^{I} n_i \bar{R}_i^2 - 3(n+1)$, where $N = n_1 + \ldots + n_I$ and $\bar{R}_i = \sum_{k=1}^{n_i} R_{ik}/n_i$ is the average pooled rank of the observations in sample i, R_{ik} are the pooled ranks.

Distribution of W under H_0: χ^2_{I-1} (approximately), the test is one sided.

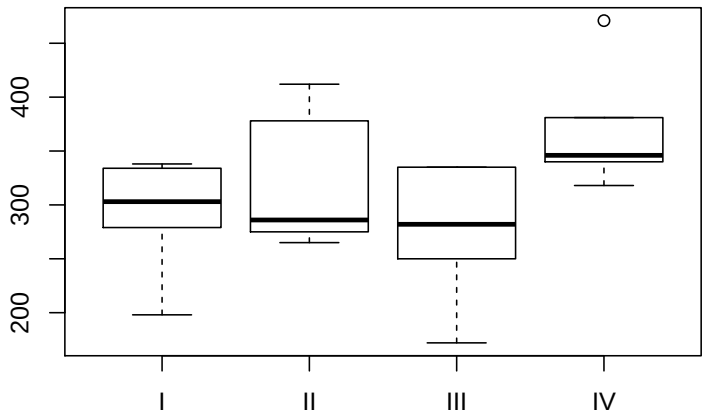
Assumption: all $n_i > 5$.

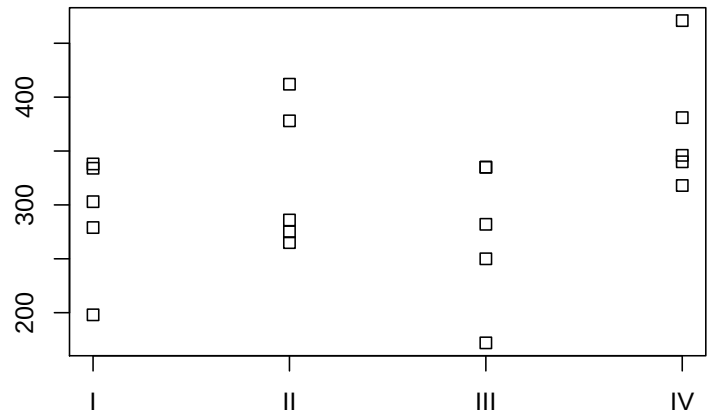
In R: `kruskal.test(y,f,data=...)`, where `y` is the outcome, `f` is the factor.

Analysis in R — data input and graphics

The dataset `ratdata.txt` contains the number of worms in rats in 4 different treatment groups.

```
> 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)
```





Analysis in R — data input

Create a data frame with the first columns containing all the outcomes $Y_{i,n}$ and the second column that indicates the levels of the factor factor.

```
> 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)

Now we perform the Kruskal-Wallis test.

```
> 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
```

The command `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, hence H_0 is not rejected.

Analysis in R — testing (2)

Compare the result of Kruskal-Wallis test with the ANOVA test results:

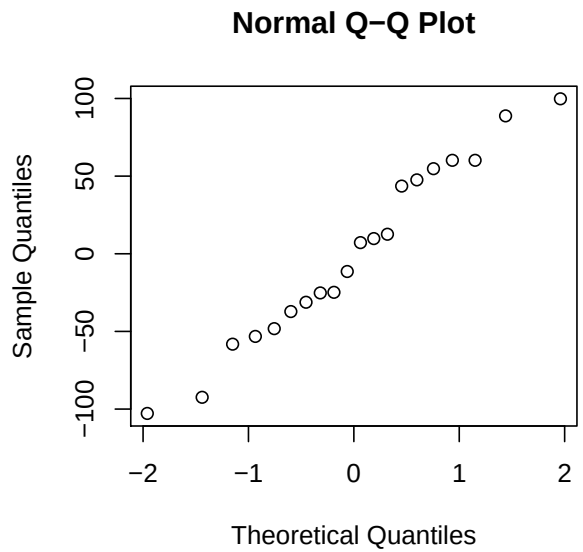
```
> rataov=lm(worms~group); anova(rataov)
Analysis of Variance Table

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
```

The one-way ANOVA also does not yield a significant difference.

```
> qqnorm(rataov$residuals)
```

The residuals do not seem to deviate significantly from normal, and both tests could be used here.



permutation tests for independent samples

Setting and design

Setting: an experiment with

- a numerical outcome Y,
- a factor that can be fixed at I levels ("label").

The same setting as 1-way ANOVA. The sample sizes for each label may differ.

EXAMPLE Medical experiment with outcome age at onset of a certain disease and label blood type.

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 label mutation probability set to 0.01, 0.02, 0.03, 0.04 or 0.05.

Design:

- Select I different labels
- Select N_i experimental units randomly from the population of label i.
- Perform the experiment $N_1 + N_2 + \ldots + N_I$ times, independently.

Analysis

Data

sample 1: $Y_{1,1}, Y_{1,2}, \ldots, Y_{1,N_1}$

sample 2: $Y_{2,1}, Y_{2,2}, \ldots, Y_{2,N_2}$

$\vdots$

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

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).

We choose a test statistic that expresses the conjectured differences between the I levels, and simulate the distribution of this statistic under H_0.

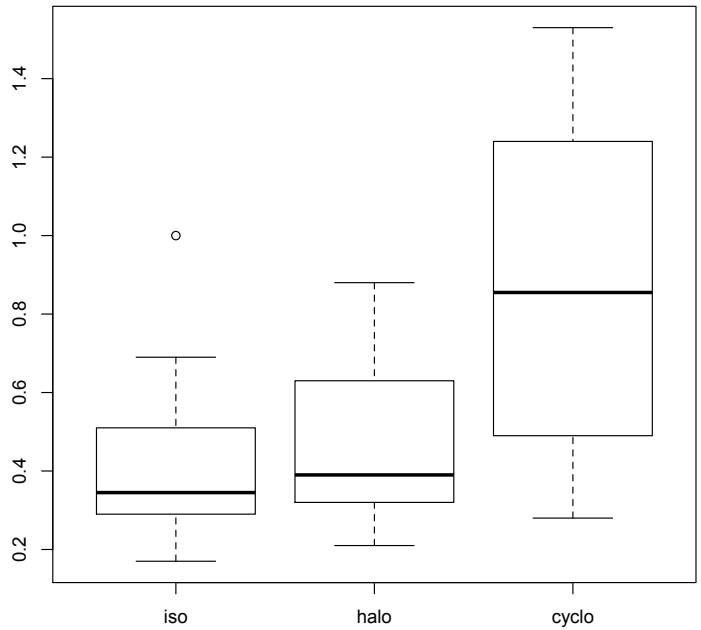
The same null hypothesis as in the Kruskal Wallis test, the difference between the Kruskal Wallis test and permutation tests is in the test statistic.

Analysis in R — data input and graphics

The dataset `dogs.txt` concerns measures of plasma epinephrine in dogs for three different anesthesia drugs ("iso","halo","cyclo").

```
> dogs=read.table("dogs.txt",header=TRUE)
> treat=factor(rep(1:3,c(10,10,10)),labels=c("iso","halo","cyclo"))
> dogsdata=data.frame(plasma=as.vector(as.matrix(dogs)),treat)
```

```
> head(dogsdata)
  plasma treat
1   0.28   iso
2   0.51   iso
3   1.00   iso
4   0.39   iso
5   0.29   iso
6   0.36   iso
> boxplot(plasma~treat,data=dogsdata)
```



Analysis in R — testing (1)

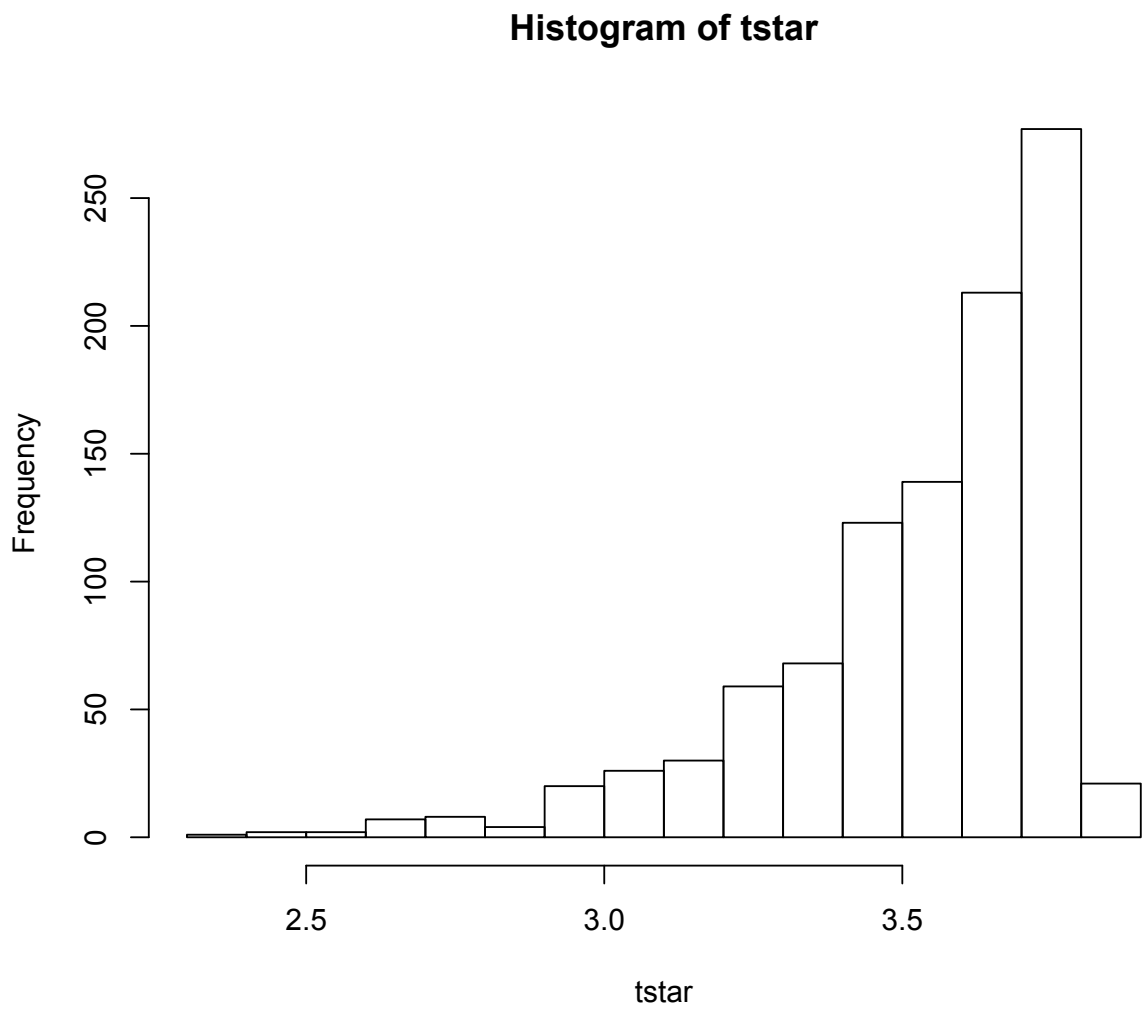
```
> attach(dogsdata)
> mystat=function(x) sum(residuals(x)^2)
> B=1000
> tstar=numeric(B)
> for (i in 1:B) {
+   treatstar=sample(treat)    ## permuting the labels
+   tstar[i]=mystat(lm(plasma~treatstar)) }
> myt=mystat(lm(plasma~treat))
```

The above test statistic is the sum the squared residuals. This can be programmed efficiently as `sum(residuals(lm(data~labels))^2)`. Note that we do not use the *p*-values of `lm`, we find p-values in a bootstrap fashion.

Analysis in R —testing (2)

```
> hist(tstar)
> myt
[1] 2.72474
> pl=sum(tstar<myt)/B
> pr=sum(tstar>myt)/B
> 2*min(pl,pr)
[1] 0.03
```

The treatment is clearly significant. This is (hopefully) in line with your results using 1-way ANOVA and Kruskal-Wallis test in the corresponding assignment.



Discussion

- A permutation test for independent samples can be performed with any test statistic that expresses difference between the samples. As an alternative to the summed squared deviations from the average per label one can look at differences in mean per label, differences in scale, etc.
- An alternative to the permutation test for independent samples is the Kruskal-Wallis test.
- Nearly all hypotheses concerning the dependence of some quantity on different levels of a "treatment" can be investigated using some sort of permutation.
- By permuting the categories of either the row or column factor in a contingency table, one can test the null hypothesis of no dependence between these two factors.
- In fact a permutation test is a bootstrap test, because the distribution of the test statistic is approximated by simulation.

To wrap up

Today we learned:

- One-way ANOVA
- Kruskal-Wallis test
- permutation tests in the setting of one-way ANOVA

Next time: 2-way ANOVA, factorial design, multiple comparisons.