Statistical inference (MVE155/MSG200)

Analysis of variance (ANOVA)

One-way ANOVA: Example

Question: Does diet effect coagulation time of blood?

Set up: 4 different diets, A, B, C and D, compared. 24 animals allocated randomly to the 4 diets - 6 animals per diet.
Is there some evidence that there is difference between the treatments?
$\rightarrow$ Analysis of variance (ANOVA)

Hypotheses:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

against

$$H_1 : \text{the means are not all equal}$$

The idea of ANOVA is to partition the overall variability to two (or more) components. How?

One-way ANOVA

We have I independent populations with means $\mu_1, ..., \mu_I$ and a sample of size n from each population. We want to compare the means of the populations, i.e. we want to test

$$H_0 : \mu_1 = ... = \mu_I \quad \text{against} \quad H_1 : \text{not all the } \mu_i'\text{s are equal}$$

The samples are $(y_{11}, ..., y_{1n})$,..., $(y_{I1}, ..., y_{In})$.

The response variable Y has I different levels, and we assume that

$$Y_{ik} \sim N(\mu_i, \sigma), i = 1, ..., I, k = 1, .., n.$$

Note that the variances are assumed to be equal,
$\sigma_1^2 = ... = \sigma_I^2 = \sigma^2$.

Each observation can be written as

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik} = \mu + \alpha_i + \delta Z_{ik},$$

where $i = 1, ..., I, k = 1, ..., n$, and

- μ is the overall population mean.
- $\alpha_i = \mu_i - \mu$ is the effect of the factor/treatment and μ_i is the mean in group i.
- Note that $\alpha_1 + ... + \alpha_I = 0$.
- $\epsilon_{ik} = \delta Z_{ik}$ is a noise term, where Z_{ik}'s are iid $N(0, 1)$-variables.

μ, μ_i, and α_i, $i = 1, ..., I$ can be estimated by

- $\hat{\mu} = \frac{1}{nI} \sum_i \sum_k y_{ik} = \bar{y}_{..}$
- $\hat{\mu}_i = \frac{1}{n} \sum_k y_{ik} = \bar{y}_{i.}$
- $\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..}$

Let Y_{ik} be the kth ($k = 1, ..., n$) (random) observation in the group i ($i = 1, ..., I$). It can also be written as

$$Y_{ik} = \bar{Y}_{..} + (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ik} - \bar{Y}_{i.}),$$

Equivalently,

$$(Y_{ik} - \bar{Y}_{..}) = (\bar{Y}_{i.} - \bar{Y}_{..}) + (Y_{ik} - \bar{Y}_{i.}),$$

The first term on the right describes the difference between the groups and the second term the difference within a group.

By taking a square in each side and summing over all the observations and groups, we obtain

$$\sum_{i=1}^{I}\sum_{k=1}^{n}(Y_{ik}-\bar{Y}_{..})^2=\sum_{i=1}^{I}\sum_{k=1}^{n}(\bar{Y}_{i.}-\bar{Y}_{..})^2+\sum_{i=1}^{I}\sum_{k=1}^{n}(Y_{ik}-\bar{Y}_{i.})^2$$

or

$$SS_T = SS_A + SS_E,$$

where

- SS_T is the total sum of squares
- SS_A is the factor (treatment) sum of squares
- SS_E is the error (residual) sum of squares.

The numbers of degrees of freedom for SS_T, SS_A, and SS_E are $df_T = nI - 1 = N - 1$, $df_A = I - 1$, and $df_E = nI - I = I(n-1)$, respectively.

As the test statistic we use

$$F = \frac{SS_A/df_A}{SS_E/df_E} = \frac{MS_A}{MS_E},$$

which, under H_0, is F-distributed with degrees of freedom $df_A = I - 1$ and $df_E = I(n-1)$.

$MS_E = SS_E/df_E = S_p^2$ gives an unbiased estimator for the variance σ^2 and can be viewed as the pooled sample variance.

When H_0 is true,

$$\mathbb{E}(MS_A) = \mathbb{E}(SS_A/df_A) = \sigma^2,$$

and even MS_A is an unbiased estimator for σ^2. However, when H_0 is not true,

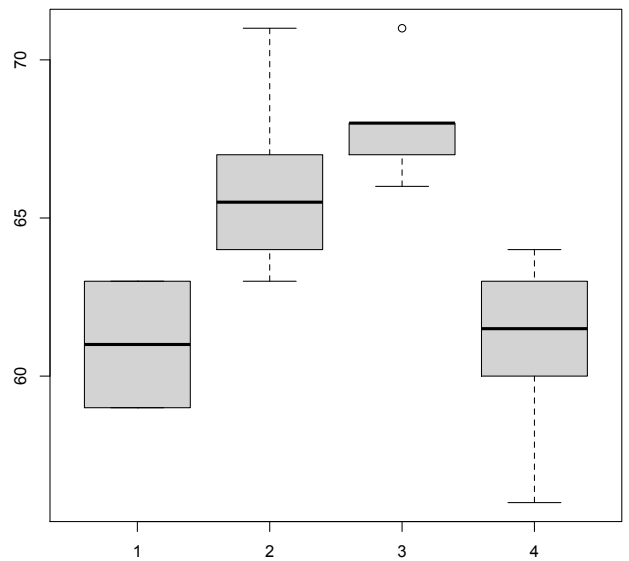
$$\mathbb{E}(MS_A) = \sigma^2 + \frac{n}{I-1}\sum_i \alpha_i^2 > \sigma^2 = \mathbb{E}(MS_E)$$

and we reject the null hypothesis if $\frac{MS_A}{MS_E}$ is large.

ANOVA: example

In our earlier example (does the diet affect the coagulation time of blood?), we have the following data:

Diet 1	Diet 2	Diet 3	Diet 4
$62^{(20)}$	$63^{(12)}$	$68^{(16)}$	$56^{(23)}$
$60^{(2)}$	$67^{(9)}$	$66^{(7)}$	$62^{(3)}$
$63^{(11)}$	$71^{(15)}$	$71^{(1)}$	$60^{(6)}$
$59^{(10)}$	$64^{(14)}$	$67^{(17)}$	$61^{(18)}$
$63^{(5)}$	$65^{(4)}$	$68^{(13)}$	$63^{(22)}$
$59^{(24)}$	$66^{(8)}$	$68^{(21)}$	$64^{(19)}$



ANOVA: example

In R, we can perform ANOVA by the following function:

$$\text{aov(coagulation} \sim \text{diet)}$$

resulting in the ANOVA tabel

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	3	228	76.0	13.57	4.66e-05
Residuals	20	112	5.6		

The p-value (in R, write 1-pf(13.57,3,20)) is very small, smaller than 0.05, and the null hypothesis can be rejected. It seems that the diet has effect on the coagulation time of blood.

Pairwise comparisons

The p-value of the F-test tells us whether the population means differ.

Why cannot we just perform $I(I-1)/2$ pairwise t-tests on the significance level α?
$\rightarrow$ There is a risk that at least one of the comparisons becomes significant by chance.

Pairwise comparisons

We perform c independent tests, where in each case, the null hypothesis is true. Let X_c be the number of tests where the null hypothesis is rejected on significance level α_c.

The overall significance level is defined as

$$\alpha = \mathrm{P}(X_c \geq 1 | H_0).$$

Since the tests are independent, $X_c \sim Bin(c, \alpha_c)$, which gives

$$\mathrm{P}(X_c \geq 1 | H_0) = 1 - \mathrm{P}(X_c = 0) = 1 - (1 - \alpha_c)^c.$$

If α_c is small, $\alpha = 1 - (1 - \alpha_c)^c \approx c\alpha_c$ (by using two first terms of Taylor series).

Bonferroni test

To obtain the overall significance level α when performing c independent tests, the individual tests should have the significance level α/c. In our ANOVA case the number of tests is $\binom{I}{2} = \frac{1}{2}I(I-1)$.

Each pairwise comparison $(i \neq j, i, j = 1, ..., I)$

$$H_0 : \mu_i = \mu_j \quad \text{against} \quad H_1 : \mu_i \neq \mu_j$$

can be performed by

- two sample t-tests at significance level $\frac{2\alpha}{I(I-1)}$ and $I(n-1)$ degrees of freedom, or
- using the corresponding $100(1-\alpha)\%$ simultaneous confidence intervals.

In both cases, the pooled sample variance $s_p^2 = ms_E$ is used.

Bonferroni test: example

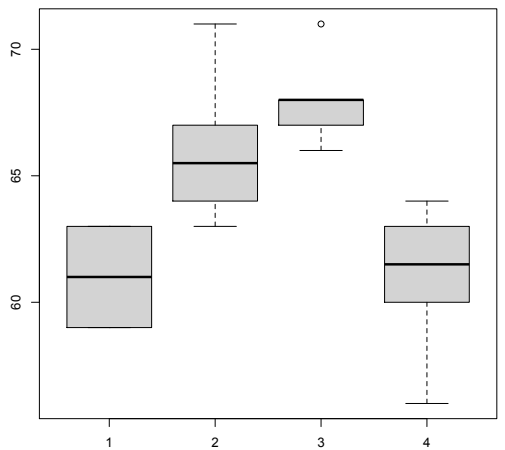
We can perform the Bonferroni-test in R by

```
pairwise.t.test(coagulation, diet, p.adjust.method="bonferroni")
```

and obtain

Pairwise comparisons using t tests with pooled SD

	1	2	3
2	0.00934	-	-
3	0.00031	0.95266	-
4	1.00000	0.00934	0.00031



There seems to be significant difference (p-value<0.05) between diets mellan 1 and 2, 1 and 3, 2 and 4, and 3 and 4.

Remarks

To be able to rely on the results based on the F-test, we need

1) Additive model

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik},$$

where μ is the overall mean, α_i is the deviation produced by the "treatment" i, and ϵ_{ik} error.

2) ϵ_{ik}'s are iid (same variance σ^2)
3) Normal assumption for the errors: $\epsilon_{ti} \sim N(0, \sigma)$

ANOVA is quite robust to

- moderate non-normality
- moderate inequality of group variances

but NOT to

- dependence of errors for an unrandomized design
- outliers

Kruskal-Wallis test

Kruskal-Wallis test is a non-parametric alternative for ANOVA when we cannot assume normality or equal variances. It tests the null hypothesis

H_0 : the underlying I independent population distributions are equal

against that they are not all equal.

Note that if the null hypothesis is rejected, it can be due to the distributions having different locations (mean, median m) or different variances.

If one can assume an identically shaped and scaled distribution for all groups, except for any difference in medians, then

$$H_0 : m_i = m_j \text{ for all i,j}$$

against that at least one of the medians differs from the others.

Kruskal-Wallis test

As in the rank sum test, we pool the samples to obtain one big sample of size $N = In$, and let r_{ik} be the pooled ranks of the observations y_{ik}. Then,

$$\sum_i \sum_k r_{ik} = 1 + \ldots + N = \frac{N(N+1)}{2}$$

and the overall mean rank

$$\bar{r}_{\cdot\cdot} = \frac{N(N+1)}{2N} = \frac{N+1}{2}.$$

Kruskal-Wallis test statistic

The test statistic is

$$w = \frac{(N-1)\sum_i n(\bar{r}_{i.} - \bar{r}_{..})^2}{\sum_i \sum_k (r_{ik} - \bar{r}_{..})^2} = \frac{12n}{N(N+1)} \sum_i \left(\bar{r}_{i.} - \frac{N+1}{2} \right)^2$$

measures the difference between the sample means

$$\bar{r}_{i.} = \frac{r_{i1} + \ldots + r_{in}}{n}$$

of the ranks. The corresponding random variable

$$W \approx \chi^2_{I-1}$$

when H_0 is true and $I = 3$ and $n \geq 5$ or $I > 3$ and $n \geq 4$.

Kruskal-Wallis test: example

In R, we can perform the Kruskal-Wallis test for the coagulation data by writing kruskal.test(coagulation $\sim$ diet) and obtain the following result

Kruskal-Wallis rank sum test

data: coagulation by diet
Kruskal-Wallis chi-squared = 17.027, df = 3, p-value = 0.0006977

Two-way ANOVA

We have two factors affecting the response variable Y, factor A having I levels and factor B having J levels. The observations can be written as

$$\{y_{ijk}, i = 1, ..., I, j = 1, ..., J, k = 1, ...n\}$$

and we assume that

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \sigma Z_{ijk},$$

where

- μ is the overall mean
- α_i is the effect by factor A, $i = 1, ..., I$
- β_j is the effect by factor B, $j = 1, ..., J$
- δ_{ij} is the interaction effect
- iid $Z_{ijk} \sim N(0, 1)$
- Note that all the variances are assumed to be equal to σ^2.

μ, α_i, β_j, and δ_{ij}, $i = 1, ..., I, j = 1, ..., J$, can be estimated by

- $\hat{\mu} = \bar{y}_{...} = \frac{1}{nIJ} \sum_i \sum_j \sum_k y_{ijk}$
- $\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}$, where $\bar{y}_{i..} = \frac{1}{nJ} \sum_j \sum_k y_{ijk}$
- $\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}$, where $\bar{y}_{.j.} = \frac{1}{nI} \sum_i \sum_k y_{ijk}$
- $\hat{\delta}_{ij} = \bar{y}_{ij.} - \bar{y}_{...} - \hat{\alpha}_i - \hat{\beta}_j = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$, where $\bar{y}_{ij.} = \frac{1}{n} \sum_k y_{ijk}$.

F-tests for two-way ANOVA

Similar to one-way ANOVA, we can write

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}(Y_{ijk}-\bar{Y}_{...})^2 = nJ\sum_{i=1}^{I}\hat{\alpha}^2 + nI\sum_{j=1}^{J}\hat{\beta}_j^2 + n\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{\delta}_{ij}^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n}\hat{\epsilon}_{ijk}^2$$

or

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E.$$

The number of degrees of freedom of these sums of squares (from left to right) are

- $df_T = nIJ - 1 = N - 1$
- $df_A = I - 1$
- $df_B = J - 1$
- $df_{AB} = (I-1)(J-1)$
- $df_E = nIJ - IJ = IJ(n-1)$.

F-tests for two-way ANOVA

In this case, we have three test statistics which each have the given F-distribution under H_0:

- $\frac{SS_A/df_A}{SS_E/df_E} = \frac{MS_A}{MS_E} \sim F_{df_A, df_E}$ (tests $H_0 : \alpha_1 = ... = \alpha_I = 0$)
- $\frac{SS_B/df_A}{SS_E/df_E} = \frac{MS_B}{MS_E} \sim F_{df_B, df_E}$ (tests $H_0 : \beta_1 = ... = \beta_J = 0$)
- $\frac{SS_{AB}/df_{AB}}{SS_E/df_E} = \frac{MS_{AB}}{MS_E} \sim F_{df_{AB}, df_E}$ (tests H_0 : no interaction between A and B)

The common variance σ^2 can be estimated by the pooled sample variance

$$s_p^2 = ms_E = \frac{\sum_i \sum_j \sum_k \hat{e}_{ijk}^2}{IJ(n-1)}$$

which gives an unbiased estimate.

Two-way ANOVA without replicates

When we have only one observation per each combination of the factor A and factor B, we are not able to estimate the interaction effect. Then, we use the additive model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \sigma Z_{ij},$$

where Z_{ij}'s are iid $N(0,1)$.

The analysis is done as in two-way ANOVA but without the interaction effect resulting in two test statistics

- $\frac{SS_A/df_A}{SS_E/df_E} = \frac{MS_A}{df_E} \sim F_{df_A, df_E}$ (tests $H_0 : \alpha_1 = ... = \alpha_I = 0$)
- $\frac{SS_B/df_A}{SS_E/df_E} = \frac{MS_B}{df_E} \sim F_{df_B, df_E}$ (tests $H_0 : \beta_1 = ... = \beta_J = 0$)

Randomized block design

A process of the manufacture of penicillin investigated

Response variable: yield

Four variants of the process: A, B, C and D

As raw material they used corn liquor and the properties of it vary quite a lot. This was expected to cause differences in the yield.

We had five blends of the material and each blend was enough for 4 experiments (one for each process)
$\rightarrow$ run $I = 4$ processes (treatments) within $J = 5$ blends (blocks) of the raw material and randomize the order of the experiments within each blend (block).

Randomiced block design

- We test the null hypothesis

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

against

$$H_1 : \text{all the means are not the same.}$$

- Note that the block effect is not included in the hypothesis, the main interest is on the process type.
- Can be performed as two-way ANOVA without the interaction term.
- Block what you can, randomize what you cannot.

Randomiserad blockdesign: example

ANOVA-table (aov(yield $\sim$ process + block) in R):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
process	3	70	23.3	1.24	0.338
block	4	264	66.0	3.51	0.041
Residuals	12	226	18.8		

Result:

- There do not seem to be any differences between the processes A, B, C and D (p-value 0.338)
- Which blend of corn liquor (block) is used affects the result significantly (p-value 0.041).

Friedman test

Friedman test is a non-parametric alternative for randomized block design (two-way ANOVA) with I groups (treatments), J blocks, and $n = 1$, which does not require normally distributed errors (and data). The null hypothesis is that there is no difference between the group means.

We rank the observations within each block j so that the ranks of $y_{1j}, ..., y_{Ij}$ are $(r_{1j}, ..., r_{Ij})$ and $r_{1j} + ... + r_{Ij} = I(I+1)/2$. Then, the average ranks within treatments (levels of the factor) are

$$\bar{r}_{i.} = \frac{r_{i1} + ... + r_{iJ}}{J}, \quad \bar{r}_{..} = \frac{\bar{r}_{1.} + ... + \bar{r}_{I.}}{I} = \frac{I(I+1)}{2I} = \frac{I+1}{2}.$$

The Friedman test statistic is

$$q = \frac{12J}{I(I+1)} \sum_{i=1}^{I} \left(\bar{r}_{i.} - \frac{I+1}{2} \right)^2,$$

which (the random version of it) is approximately χ^2_{I-1}-distributed when H_0 is true.

Friedman test

Seven ($I = 7$) treatments (including "no treatment" and placebo) against itching were compared. Each treatment was applied to $J = 10$ male volunteers after an itching condition had been initiated by a certain injection. In the table below, the duration of the itching is given in seconds (and ranks in parantheses).

BG	174	263	105	199	141	108	141
rank	5	7	1	6	3.5	2	3.5

Subject	Tr 1	Tr 2	Tr 3	Tr 4	Tr 5	Tr 5	Tr 7
BG	174 (5)	263 (7)	105 (1)	199 (6)	141 (3.5)	108 (2)	141 (3.5)
JF	224 (6)	213 5)	103 (1)	143 (2)	168 (3)	341 (7)	184 (4)
BS	260 (7)	231 (6)	145 (4)	113 (2)	78 (1)	159 (5)	125 (3)
SI	225 (6)	291 (7)	103 (1)	225 (4)	164 (3)	135 (2)	227 (5)
BW	165 (3)	168 (4)	144 (2)	176 (5)	127 (1)	239 (7)	194 (6)
TS	237 (7)	121 (3)	94 (1)	144 (5)	114 (2)	136 (4)	155 (6)
GM	191 (7)	137 (5)	35 (1)	87 (2)	96 (3)	140 (6)	121 (4)
SS	100 (1)	102 (2)	133 (5)	120 (3)	222 (7)	134 (6)	129 (4)
MU	115 (5)	89 (3)	83 (2)	100 (4)	165 (6)	185 (7)	79 (1)
OS	189 (4)	433 (7)	237 (5)	173 (2)	168 (1)	188 (3)	317 (6)
$\sum r_{i.}$	51	49	23	35	30.5	49	42.5
$\bar{r}_{i.}$	5.10	4.90	2.30	3.50	3.05	4.90	4.25

The test statistics gets the value 14.86 which is significant at 5% level ($\chi^2_6(0.05) = 12.59$). The effects of the treatments seem to differ.