Lecture 6 – Program

- Analysis of variance
- Experimental design

ANOVA = ANalysis Of VAriance

1. Comparison of several groups
2. Variation within and between groups
3. One-way layout and t-test
4. Connection to regression
5. Parameterization
6. Two-way layout
7. Interaction
8. Higher-way layouts

Two-sample t-tests: Comparison of two groups

Example:
Two treatments: placebo and medication

Group 1: placebo

Group 2: new medication

Is blood pressure lower with medication?

One-way ANOVA: Comparison of k groups

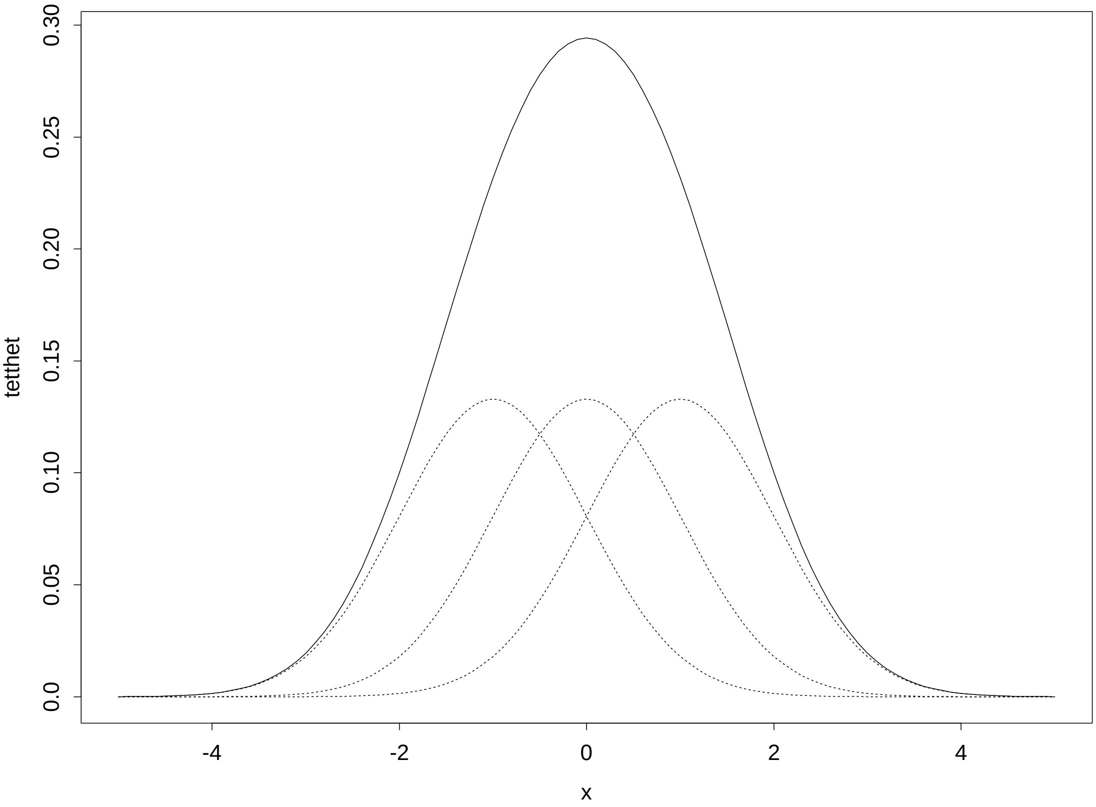
Example: Three different medications

Group j: medication no. j

Is there a difference between the medications?

If yes, which results in lowest blood pressure?

Decomposing the variation:



Total variance $=$

Variance *within* groups

$+$ Variance *between* groups

Important quantities and notation

y_{ij} = observation number i in group j

$(i = 1, ..., n_j \quad j = 1, ..., k)$

We assume that all observations are independent and that $y_{ij} \sim N(\mu_j, \sigma^2)$

$\bar{y}_{\cdot j}$ = mean in group j

$\bar{y}_{\cdot\cdot}$ = total mean.

Sum of squares:

Total:	SS_{tot}	$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot\cdot})^2$
Between :	SS_{tre}	$= \sum_{j=1}^{k} n_j (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$
Within :	SS_{res}	$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{\cdot j})^2$

Important decomposition:

$$SS_{tot} = SS_{tre} + SS_{res}$$

Test of $H_0 : \mu_1 = \cdots = \mu_k$

Unbiased estimator of σ^2:

$\widehat{\sigma}^2 = MS_{res} = SS_{res}/(n-k)$

($n =$ total number of observations)

Under the null hypothesis σ^2 can also be estimated by

$MS_{tre} = SS_{tre}/(k-1)$

If the statistic

$$F = \frac{MS_{tre}}{MS_{res}} = \frac{SS_{tre}/(k-1)}{SS_{res}/(n-k)}$$

is much larger than 1, H_0 is not reasonable.

F is F-distributed with $k-1$ and $n-k$ degrees of freedom under H_0. This result is used to compute the p-value of the test.

Relation to two sample t-test

Number of groups is $k = 2$

Will test $H_0 : \mu_1 = \mu_2$

The test statistic

$$t = \frac{\bar{y}_{\cdot 1} - \bar{y}_{\cdot 2}}{se(\bar{y}_{\cdot 1} - \bar{y}_{\cdot 2})}$$

is t-distributed with $n_1 + n_2 - 2$ degrees of freedom under H_0.

May show that $t^2 = F$

t^2 is F-distributed with 1 and $n_1+n_2-2 = n-2$ degrees of freedom under H_0.

The usual (two-sided) t-test for two samples is a special case of the F-test in a one-way ANOVA.

ANOVA-table for one-way layout:

Source	SS	df	MS	F	p-value.
Treatment	SS_{tre}	$k-1$	MS_{tre}	$F = \frac{MS_{tre}}{MS_{res}}$	p
Residual	SS_{res}	$n-k$	MS_{res}		
Total	SS_{tot}	$n-1$			

p-value is obtained from:

$$p = P(F_{k-1,n-k} > \text{observed value of } F)$$

Example (B&S, page 26):

Comparing blood coagulation times for rats given four diets

Diet	No. obs.	Mean	Sd
A	4	61	1.8
B	6	66	2.8
C	6	68	1.7
D	8	61	2.6

Anova-table:

Source	SS	df	MS	F	p-value
Diet	228	3	76.0	13.57	<0.0001
Residual	112	20	5.6		
Total	340	23			

R commands (diet coded as factor):

```
fit<-lm(time~diet, data=rats)
anova(fit)
```

ANOVA as multiple regression

Reorder the observations (with covariates) in the form $y_1, y_2, ..., y_n$, where:

- the first n_1 belong to group 1,
- the next n_2 belong to group 2,
- etc.

Let x_{ij} be an indicator (dummy) equal to 1 if y_i is in group j and equal to 0 otherwise.

Then the model can be expressed as

$$y_i = \mu_1 x_{i1} + \mu_2 x_{i2} + ... + \mu_k x_{ik} + \varepsilon_i$$

Here the errors are independent and $\varepsilon_i \sim N(0, \sigma^2)$.

In other words: a linear multiple regression without intercept.

Various parameterizations

1. Without intercept:

$$y_i = \mu_1 x_{i1} + \mu_2 x_{i2} + ... + \mu_k x_{ik} + \varepsilon_i$$

2. With group 1 as reference:

$$y_i = \mu_1 + (\mu_2 - \mu_1) x_{i2} + ... + (\mu_k - \mu_1) x_{ik} + \varepsilon_i$$

3. As deviations from the grand mean $\mu = (\mu_1 + ... + \mu_k)/k$:

$$y_i = \mu + (\mu_1 - \mu) x_{i1} + ... + (\mu_k - \mu) x_{ik} + \varepsilon_i$$

Option 2, called **treatment-contrast**, is default in R.

Option 3, called **sum-contrast**, is commonly used for ANOVA, and may be specified in R by the command:

```
options(contrasts=c("contr.sum","contr.poly"))
```

Two-way ANOVA

Two categorical variables (or factors) A and B

Factor A has r levels, factor B has c levels

One observation for each combination of the levels of the factors

y_{ij} = observation with level i for A and j for B

Model (only main effects):

$$y_{ij} = \mu + a_i + b_j + \varepsilon_{ij}$$

Decomposition of sum of squares:

$SS_{tot} = SS_A + SS_B + SS_{res}$

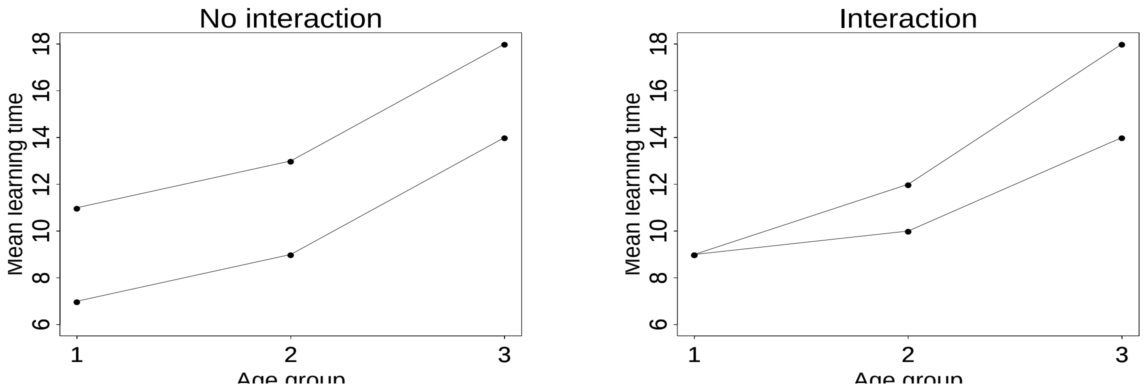
ANOVA-table:

Source	SS	df	MS	F	p-value
A	SS_A	$r-1$	MS_A	MS_A/MS_{res}	p_A
B	SS_B	$c-1$	MS_B	MS_B/MS_{res}	p_B
Res	SS_{res}	$n-c-r+1$	MS_{res}		
Tot	SS_{tot}	$n-1$			

Two-way layouts and interaction

The expected response at level i for factor A and level j for factor B may differ from the sum of the main effects $a_i + b_j$.

Graphically this shows up as non-parallel lines in a plot of the expected values (B&S p. 64):



Model for interaction:

$$y_{ij} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ij}$$

$(ab)_{ij}$ = interaction

Two-way layout, contd.

A model for a two-way layout including intercept, main effects and interactions can *not* be estimated if there is only one observation per combination of factor levels or cell, (i, j).

To estimate interaction we need replications:

$$y_{ijk} = k\text{th observation at levels } A = i \text{ and } B = j$$

Balanced design: Same number of replications m per combination of levels (i, j).

With a balanced design there is a unique decomposition of the sum of squares

$$SS_{tot} = SS_A + SS_B + SS_{AB} + SS_{res}$$

where SS_A og SS_B are defined earlier and SS_{AB} is the sum of squares for interaction.

ANOVA-table for balanced two-way layout with replications

Source	SS	df	MS	F	p-value
A	SS_A	$r-1$	MS_A	MS_A/MS_{res}	p_A
B	SS_B	$c-1$	MS_B	MS_B/MS_{res}	p_B
AB	SS_{AB}	$(r-1)(c-1)$	MS_{AB}	MS_{AB}/MS_{res}	p_{AB}
Residual	SS_{res}	$n-rc$	MS_{res}		
Total	SS_{tot}	$n-1$			

Relevant hypotheses:

$$\begin{array}{lll} \mathrm{H}_{AB}: & (ab)_{ij} = 0 & \text{No interaction} \\ \mathrm{H}_{A}: & a_i = 0 & \text{No main effect of A} \\ \mathrm{H}_{B}: & b_j = 0 & \text{No main effect of B} \end{array}$$

Non-balanced designs

ANOVA is used a lot for observational studies, and then it is usually difficult to obtain a balanced design.

In an non-balanced design the number of observations are not the same for all combinations (i, j).

The decomposition of sum of squares is *not* unique.

Usually one can estimate both main- and interaction effects, but the situation is not so neat as in a balanced design.

But one should try to adjust for confounding variables, even if they are correlated.

Higher-way layouts

E.g. three factors A, B og C.

Data:

$y_{ijkl} =$ replication l with levels $A = i, B = j$ og $C = k$

Model:

$$y_{ijkl} = \mu + a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} + \varepsilon_{ijkl}$$

ANOVA-table:

Source	SS	$df^\star$	MS	F	p
A	SS_A		MS_A	F_A	p_A
B	SS_B		MS_B	F_B	p_B
C	SS_C		MS_C	F_C	p_C
AB	SS_{AB}		MS_{AB}	F_{AB}	p_{AB}
AC	SS_{AC}		MS_{AC}	F_{AC}	p_{AC}
BC	SS_{BC}		MS_{BC}	F_{BC}	p_{BC}
ABC	SS_{ABC}		MS_{ABC}	F_{ABC}	p_{ABC}
Residual	SS_{res}		MS_{res}		
Total	SS_{tot}	$n-1$			

*) can be found in computer print-outs

The decomposition is unique when the design is balanced, but main- and interaction effects can be estimated and tested in more general situations.

Experimental design

1. Sample size and power calculations

2. Randomization

3. Blocking

4. Simultaneous variation of factors versus one at a time

Sample size and power calculations

Example: Two normal samples, σ known

n_j observations in group $j = 1, 2$.

Question :

How large must n_1 and n_2 be in order that the probability is "large" for rejecting
$H_0 : \mu_1 = \mu_2$ when $\mu_2 - \mu_1 = \Delta$?

Here Δ is a user-specified difference of "substantial importance".

It is "optimal" to choose the same size for both samples, i.e. $n_1 = n_2 = n/2$ where n is the total number of observations.

Test statistic:

$$Z = \frac{\bar{y}_2 - \bar{y}_1}{se(\bar{y}_1 - \bar{y}_2)} \sim \mathsf{N}(\sqrt{n}\Delta/(2\sigma), 1).$$

Reject two-sided hypothesis
at 5% level if $|Z| > 1.96$

Reject one-sided hypothesis
at 2.5% level if $Z > 1.96$

Consider one-sided test
(for pedagogical reasons).

Can express Z as

$$Z = Z_0 + \sqrt{n}\Delta/(2\sigma)$$

where $Z_0 \sim \mathsf{N}(0, 1)$.

If we want probability of rejection to exceed 80%, we should have:

$$0.80 \leq P(Z > 1.96) = P\left(Z_0 > 1.96 - \sqrt{n}\frac{\Delta}{2\sigma}\right).$$

For $Z_0 \sim \mathrm{N}(0,1)$ we have

$$P(Z_0 > -0.84) = 0.80$$

Therefore we should have

$$-0.84 > 1.96 - \sqrt{n}\frac{\Delta}{2\sigma}$$

which gives

$$n \geq \frac{4(1.96 + 0.84)^2\sigma^2}{\Delta^2},$$

E.g if $\mu_2 - \mu_1 = \Delta = \sigma$

$$n \geq 4 * 2.8^2 = 31.36, \text{ dvs. } n \geq 32.$$

Usually σ is unknown, and we will have to use a t-test. This means that n must be slightly larger.

Sample size and power calculations in R

Example: Two sample t-test, σ unknown

Want power 80% for $\mu_2 - \mu_1 = \Delta = \sigma$

R command:

```
power.t.test(n = NULL, delta = 1, sd=1, power=0.80)

     Two-sample t test power calculation

              n = 16.71477
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

 NOTE: n is number in *each* group
```

R can do power and sample size calculations for a number of tests. Give the command `help.search("power")` to get information on these

Summary of power calculations

- Parameter of interest, θ

- Nullhypothesis $H_0 : \theta = \theta_0$

- Test statistics V

- Reject with level α if $V > v_0 =$ critical value

- Power function $\gamma_n(\theta) = P(V > v_0 \,|\, \theta, n)$

- Alternative of interest to θ_0 is θ_1.

- Wants power $1 - \beta$ for rejecting H_0 under the alternative of interest, i.e.
  n so large that $\gamma_n(\theta_1) > 1 - \beta$

Randomization

Want to compare effects of several treatments

Randomization means that we randomly assign the units to the treatments

Why randomize?

- To avoid systematic assignment to treatments, which can entail biased estimates of treatment effects

- In addition: The errors will be symmetrically distributed, so that the approximation to the normal distribution is good.

Randomization remove bias

Example:
Comparison of placebo and treatment

x_{i1} dummy variable indicating whether unit i receives treatment

x_{i2} confounding covariate (often not observed)

Assume "true" model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

A two-sample t-test is the same as running the simple linear regression

$$y_i = a + b x_{1i} + \epsilon_i$$

We know that we then estimate

$$b = \beta_1 + \beta_2 \tau \frac{v_1}{v_2}$$

where $\tau = \text{corr}(x_{i1}, x_{i2})$ and v_j is the standard deviation for x_{ij} (cf. R-exercise 3)

The estimate of the treatment effect is **biased** if $\tau \neq 0$ and $\beta_2 \neq 0$.

By **randomizing**, the treatment x_{i1} and the confounding covariate x_{i2} are independent

Then $\tau = 0$ and the estimate of the treatment effect is **unbiased** even if $\beta_2 \neq 0$.

Randomization and symmetric distribution of errors

Continue the example with placebo and treatment

Numerator of t-statistics is

$$\bar{y}_2 - \bar{y}_1 = \frac{2}{n}\sum_{i=1}^{n/2}(y_i - y_{i+n/2})$$

(assuming $x_{i1} = 1$ for the first $n/2$ units)

We may write:

$$y_i - y_{i+n/2} = \beta_1 + \beta_2(x_{i2} - x_{i+n/2,2}) + (\varepsilon_i - \varepsilon_{i+n/2})$$

Even if the distributions of the x_{i2}'s and the ε_i's are skewed, the differences

$$x_{i2} - x_{i+n/2,2} \qquad \text{and} \qquad \varepsilon_i - \varepsilon_{i+n/2}$$

will typically be symmetrically distributed.

Blocking

Originally one divided a field into *blocks* to account for possible trends in soil fertility.

Today the term ”block” is used when the observations are grouped according to the levels of one factor, which is *not* the factor of main interest (treatment)

Example: Production of penicillin (B&S, p. 61)

- Want to compare treatments
- Raw material consisting of various mixtures
- The mixtures have effect on the response y

Possible strategies:

1. Randomize without taking the blocks into account
2. Randomize within each block

ANOVA-table from 2. strategy

Source	SS	*df*	MS	F	p
Treatmant (A)	SS_A	$r-1$	MS_A	F_A	p_A
Block (B)	SS_B	$c-1$	MS_B	F_B	p_B
Residual	SS_{res}	$n-c-r+1$	MS_{res}		
Total	SS_{tot}	$n-1$			

This is the same as for a two-way ANOVA without replicates.

If there is a substantial block effect, strategy 2 it to be preferred to strategy 1.

Summary of multi-factorial designs

$$y_i = a_i + \beta x_i + \varepsilon_i$$

To take into account a covariate x_i reduces the variance of the residuals and increases the significance (except a possible loss of df).

If x_i is categorical it can be used for blocking

For balanced block designs the sum of squares can be uniquely decomposed

Paroles:

1. Block what is possible

2. Randomize the rest

One by one variation

- Keep levels for factors B, C, D, etc. constant
- Vary level of factor A = treatment

Alternative:

- Vary all factors simultaneously

Advantages with the alternative

- Can analyze the effect of all factors in one design
- Can discover interactions
- Less residual variance