Analysis of variance (ANOVA)

(STATS201.stats 201 30: Statistical inference)

Anna Helga Jónsdóttir
Sigrún Helga Lund

December 15, 2012

Introduction

- In lecture 110 we discussed inference on the mean of a population (μ).
- In the former part of lecture 120 we discussed inference on the difference of the mean of two populations ($\mu_1 - \mu_2$).
- In the latter part of lecture 120 we discussed inference on paired measurements (μ_D).
- Now we will discuss a method that we can apply to compare the means of two or more populations. The method is called analysis of variance, or ANOVA.

Analysis of variance

- Analysis of variance is one of the most commonly used statistical methods. There are several variants of it that can be used in a vast number of various different cases.

- We will only look at one variant of the method that is called *one-sided ANOVA*.

- It is applied to data that contain samples from two or more populations and it is common to speak of groups when discussing the samples.

- The method compares the variability of the measurements within the groups on one hand and between them on the other hand.

- ANOVA assumes that the samples are random samples, that they are sampled from populations with a normal distribution and that

Conducting hypothesis tests

Conducting hypothesis tests

1 Decide which hypothesis test is appropriate for our measurements.

2 Decide the α-level.

3 Propose a null hypothesis and decide the direction of the test (one- or two-sided).

4 Calculate the test statistic for the hypothesis test.

5a See whether the test statistic falls within the rejection interval.

5b Look at the p-value of the test statistic.

6 Draw conclusions.

$_{1.} = 8.14$

Average change group 2: $_{2.} = 6.28$

Average change group 3: $_{3.} = 13.01$ Th

question is, do the drug decrease the blood pressure equally or not?

The data

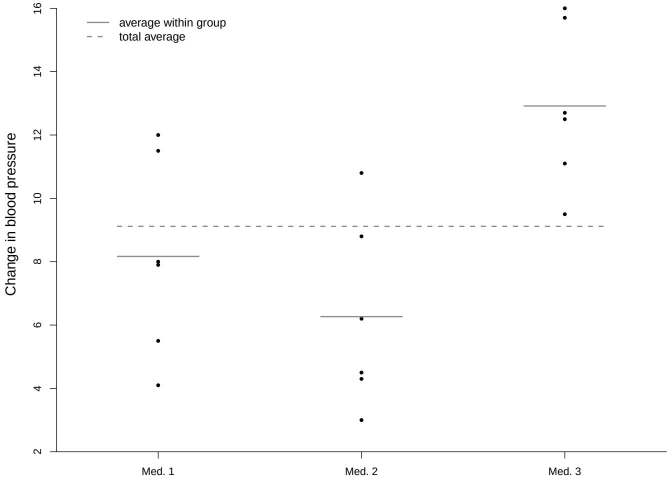

Figure: Data for ANOVA

Syntax

Syntax in ANOVA

The following syntax is common in textbooks and papers discussing ANOVA.

- y_{ij} : i denotes the number of the group and j denotes the number of a measurement within a group. y_{ij} is the j-th measurement in group i.
- a : We denote the number of groups with a.
- n_i : We denote the number of measurements in group i with n_i.
- N : The total number of measurements is denoted with N.

$$N = n_1 + n_2 + ... + n_a.$$

- $\bar{y}_{i.}$: $\bar{y}_{i.}$ denotes the mean of group i

$$\bar{y}_{i.} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}.$$

- $\bar{y}_{..}$: $\bar{y}_{..}$ denotes the overall mean of all measurements (in all groups).

$$\bar{y}_{..} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}}{N}.$$

Sums of squares

- We need to calculate three sums of squares, and are they denoted with SS_T, SS_{Tr} and SS_E.
- SS_T is the total sums of squares and is a measure of the total variation of the measurements.
- SS_{Tr} is a measure of the variation between groups (or treatments), that is, how much to the means of the groups vary.
- SS_E is a measure of the variability within groups (or treatments) and is therefore a measure of the error. It shows how much the measurements deviate from the mean of the group.

Sums of squares

Sums of squares in one sided ANOVA

The Sums of squares are calculated with

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$$

$$SS_{Tr} = \sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

The total variation can be divided into the variation between groups on one hand and the variation within groups on the other hand or

$$SS_T = SS_{Tr} + SS_E.$$

Sums of squares - graphically

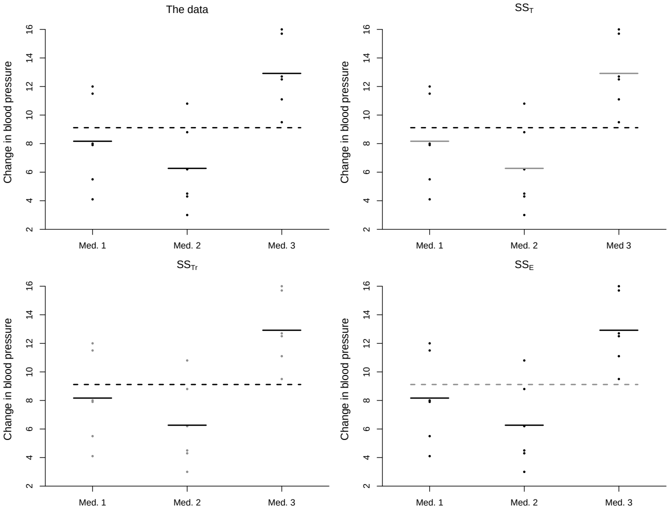

Figure: Sums of squares

ANOVA table

- It is common to visualize the sums of squares in a so-called *ANOVA table*.
- The table consist of three columns and three lines.
- The first column contains the sums of squares, the next one contains the number of *degrees of freedom*. The first column contains the sums of squares, the second one contains the number of *degrees of freedom* for each sum of squares and the third column contains so-called mean sum of squares.
- Mean sum of squares is calculated by dividing the corresponding sum of squares with the number of corresponding degrees of freedom (in the same line).

ANOVA table

Sums of squares	Degrees of freedom	Mean sum of squares
SS_{Tr}	$a-1$	$MS_{Tr} = \frac{SS_{Tr}}{a-1}$
SS_E	$N-a$	$MS_E = \frac{SS_E}{N-a}$
SS_T	$N-1$	

Hypothesis testing with ANOVA

Hypothesis testing with ANOVA

The hypothesis we want to test is generally

$$H_0 : \mu_1 = \mu_2 = ... = \mu_a$$

against the alternative hypothesis

$$H_1 : \text{At least one of the means differs from the other means.}$$

The test statistic is

$$F = \frac{SS_{Tr}/(a-1)}{SS_E/(N-a)} = \frac{MS_{Tr}}{MS_E}.$$

If the null hypothesis is true, the test statistic follows the F-distribution with $a-1$ and $N-a$ degrees of freedom, or $F \sim F_{(a-1,N-a)}$, where a is the number of groups and N is the total number of measurements.
H_0 is rejected if F $> F_{1-\alpha,(a-1,N-a)}$

Hypothesis testing with one-sided ANOVA

- The alternative hypothesis is that at least one of the means differs from the others, it is therefore then only information we receive if the null hypothesis is rejected.
- We do not know which of the means differs from the others or if they are potentially all different.
- Further analysis needs to be done in order to find that out. A common test i Tukey's test, but they will not be covered in this lecture.