02323 Introduction to Statistics

Lecture 11: One-way Analysis of Variance, ANOVA

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

1. Intro: Small example and TV-data from B&O
2. Model and hypothesis
3. Computation - decomposition and the ANOVA table
4. Hypothesis test (F-test)
5. Within-group variability and relation to the 2-sample t-test
6. Post hoc analysis
7. Model control / model validation
8. A complete example - from the book

Overview

One-way ANOVA - simple example

Group A	Group B	Group C
2.8	5.5	5.8
3.6	6.3	8.3
3.4	6.1	6.9
2.3	5.7	6.1

Is there a difference (in means) between the groups A, B and C?

Analysis of variance (ANOVA) can be used for the analysis, if the observations in each group can be assumed to be normally distributed.

TV set development at Bang & Olufsen

Sound and image quality measured by the human perceptual instrument.

Bang & Olufsen data in R

```
# Get the B&O data from the lmerTest-package
library(lmerTest)

## Warning: pakke 'lmerTest' blev bygget under R version 4.1.3
## Warning: pakke 'lme4' blev bygget under R version 4.1.3

data(TVbo)
head(TVbo) # First rows of the data

# Define factor identifying the 12 TV set and picture combinations
TVbo$TVPic <- factor(TVbo$TVset:TVbo$Picture)

# Each of 8 assessors scored each of the 12 combinations twice.
# Average the two replicates for each assessor and combination of
# TV set and picture
library(doBy)
TVbonoise <- summaryBy(Noise ~ Assessor + TVPic, data = TVbo,
                       keep.names = T)

# One-way ANOVA of the noise (not the correct analysis!)
anova(lm(Noise ~ TVPic, data = TVbonoise))
```

One-way ANOVA - simple example in R

```
# Input data
y <- c(2.8, 3.6, 3.4, 2.3,
       5.5, 6.3, 6.1, 5.7,
       5.8, 8.3, 6.9, 6.1)

## Define treatment groups
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

## Plot data by treatment groups
par(mfrow = c(1,2))
plot(y ~ as.numeric(treatm), xlab = "Treatment", ylab = "y")
boxplot(y ~ treatm, xlab = "Treatment", ylab = "y")
```

Overview

One-way ANOVA, model

- The model may be formulated as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}\,,$$

  where the ε_{ij} are assumed to be independent and identically distributed (i.i.d.) with

$$\varepsilon_{ij} \sim N(0, \sigma^2)\,.$$

- μ: overall mean.
- α_i: effect of group (treatment) i.
- Y_{ij}: jth measurement in group i (j runs from 1 to n_i).

One-way ANOVA, hypothesis

- We want to compare the (more than 2) means $\mu + \alpha_i$ in the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)\,.$$

- The hypothesis may be formulated as

$$\begin{aligned} H_0 &: \quad \alpha_i = 0 \quad \text{for all } i \\ H_1 &: \quad \alpha_i \neq 0 \quad \text{for at least one } i \end{aligned}$$

Overview

One-way ANOVA, decomposition and the ANOVA table

- With the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

  the total variation in the data can be decomposed:

$$SST = SS(Tr) + SSE\,.$$

- 'One-way' refers to the fact that there is only one factor in the experiment on k levels.
- The method is called analysis of variance, because the testing is carried out by comparing certain variances.

Formulas for sums of squares

- Total sum of squares ("the total variance")

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

Formulas for sums of squares

- Total sum of squares ("the total variance")

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$$

- The sum of squares for the residuals ("residual variance after model fit")

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

Formulas for sums of squares

- Total sum of squares ("the total variance")

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

- The sum of squares for the residuals ("residual variance after model fit")

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

- Sum of squares of treatment ("variance explained by the model")

$$SS(Tr) = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

The ANOVA table

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares
Treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$
Total	$n-1$	SST	

```
# One-way ANOVA using anova() and lm()
anova(lm(y ~ treatm))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## treatm     2   30.8   15.40    26.7 0.00017 ***
## Residuals  9    5.2    0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overview

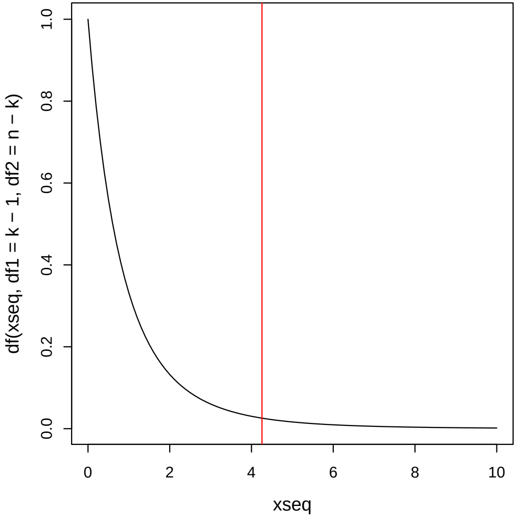
One-way ANOVA, F-test

- We have: (Theorem 8.2)
$$SST = SS(Tr) + SSE$$
- and we can find the test statistic
$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{MS(Tr)}{MSE}$$
where
  - k is the number of levels of the factor,
  - n is the total number of observations.
- Choose the significance level α, and compute the test statistic F.
- Compare the test statistic to the relevant quantile of the F-distribution:
$$F \sim F_{\alpha}(k-1, n-k) \text{ (Theorem 8.6)}$$

The F-distribution and the F-test

```r
# Remember, this is "under H0" (i.e. we compute as if H0 is true)

# Number of groups
k <- 3

# Total number of observations
n <- 12

# Sequence for plot
xseq <- seq(0, 10, by = 0.1)

# Plot density of the F-distribution
plot(xseq, df(xseq, df1 = k-1, df2 = n-k), type = "l")

# Plot critical value for significance level 5%
cr <- qf(0.95, df1 = k-1, df2 = n-k)
abline(v = cr, col = "red")
```

An F-distribution with a critical value



The ANOVA table

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p-value
treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{obs}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{obs}})$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

```
anova(lm(y ~ treatm))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## treatm     2   30.8   15.40    26.7 0.00017 ***
## Residuals  9    5.2    0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-way ANOVA F-test “by hand”

```
k <- 3; n <- 12  # Number of groups k, total number of observations n

# Total variation, SST
(SST <- sum( (y - mean(y))^2 ))

# Residual variance after model fit, SSE
y1 <- y[1:4]; y2 <- y[5:8]; y3 <- y[9:12]

(SSE <- sum( (y1 - mean(y1))^2 ) +
        sum( (y2 - mean(y2))^2 ) +
        sum( (y3 - mean(y3))^2 ))

# Variance explained by the model, SS(Tr)
(SSTr <- SST - SSE)

# Test statistic
(Fobs <- (SSTr/(k-1)) / (SSE/(n-k)))

# P-value
(1 - pf(Fobs, df1 = k-1, df2 = n-k))
```

Overview

Within-group variability and relation to the 2-sample t-test (Theorem 8.4)

The residual sum of squares, SSE, divided by $n-k$, also called residual mean square, $MSE = SSE/(n-k)$, is the average within-group variability:

$$MSE = \frac{SSE}{n-k} = \frac{(n_1-1)s_1^2 + \cdots + (n_k-1)s_k^2}{n-k} \tag{1}$$

$$s_i^2 = \frac{1}{n_i-1}\sum_{i=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

ONLY when $k=2$: (cf. Method 3.52)

$$MSE = s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n-2}$$

$$F_{\text{obs}} = t_{\text{obs}}^2$$

where t_{obs} is the pooled t-test statistic from Methods 3.52 and 3.53.

Overview

Post hoc confidence interval - Method 8.9

- A single pre-planned confidence interval for the difference between treatment i and j is found as:

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{\frac{SSE}{n-k}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{2}$$

  where $t_{1-\alpha/2}$ is based on the t-distribution with $n-k$ degrees of freedom.

- Note the fewer degrees of freedom as more unknowns are estimated in the computation of $MSE = SSE/(n-k) = s_p^2$ (i.e. pooled variance estimate)

- If all $M = k(k-1)/2$ combinations of pairwise confidence intervals are found use the formula M times, but each time with $\alpha_{\text{Bonferroni}} = \alpha/M$.

Post hoc pairwise hypothesis test- Method 8.10

- A single pre-planned level α hypothesis test:

$$H_0:\ \mu_i = \mu_j,\ \ H_1:\ \mu_i \neq \mu_j$$

is carried out as:

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{3}$$

and

$$p - \text{value} = 2P(t > |t_{\text{obs}}|)$$

where the t-distribution with $n-k$ degrees of freedom is used.

- If all $M = k(k-1)/2$ combinations of pairwise hypothesis tests are carried out use the approach M times, but each time with significance level $\alpha_{\text{Bonferroni}} = \alpha/M$.

Overview

Variance homogeneity

Look at a box plot to check whether the variability seems different across the groups.

```
# Check assumption of homogeneous variance using, e.g.,
# a box plot.
plot(treatm, y)
```

Normal assumption

Look at a normal QQ-plot of the residuals

```
# Check normality of residuals using a normal QQ-plot
fit1 <- lm(y ~ treatm)
qqnorm(fit1$residuals)
qqline(fit1$residuals)
```

Overview

A complete example - from the book

Introduction to Statistics

Agendas eNotes Course Material Podcast Forum Quiz Admin

Dokumentegenskaber...

8.2.5 A complete worked through example: plastic types for lamps

Example 8.17 Plastic types for lamps

On a lamp two plastic screens are to be mounted. It is essential that these plastic screens have a good impact strength. Therefore an experiment is carried out for 5 different types of plastic. 6 samples in each plastic type are tested. The strengths of these items are determined. The following measurement data was found (strength in kJ/m^2):

Type of plastic				
I	II	III	IV	V
44.6	52.8	53.1	51.5	48.2
50.5	58.3	50.0	53.7	40.8
46.3	55.4	54.4	50.5	44.5
48.5	57.4	55.3	54.4	43.9
45.2	58.1	50.6	47.5	45.9
52.3	54.6	53.4	47.8	42.5

Overview

1. Intro: Small example and TV-data from B&O
2. Model and hypothesis
3. Computation - decomposition and the ANOVA table
4. Hypothesis test (F-test)
5. Within-group variability and relation to the 2-sample t-test
6. Post hoc analysis
7. Model control / model validation
8. A complete example - from the book