# Course 02402 Introduction to Statistics Lecture 11:

# Twoway Analysis of Variance, ANOVA

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

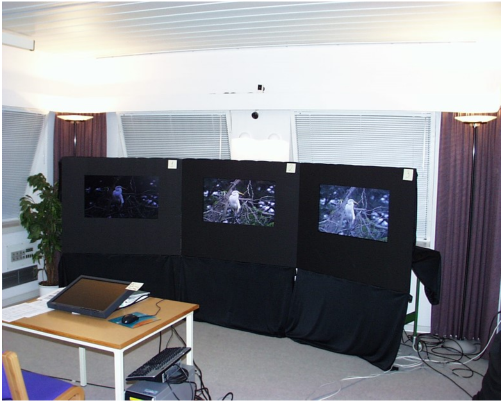
# Agenda

1. Intro: Small example and TV-data from B&O
2. Model
3. Computation - decomposition and the ANOVA table
4. Hypothesis test (F-test)
5. Post hoc analysis
6. Model control
7. A complete example - from the book

# Agenda

# TV set development at Bang & Olufsen

*Sound and image quality is measured by th human perceptual instrument:*



We developed a tool that is used by B&O to ANOVA (among other things) PanelCheck (*Show Panelcheck programme with TV data*)

# Bang & Olufsen data in R:

```
##  # Getting the Bang and Olufsen data from the lmerTest-package:
library(lmerTest) # (Developed by us)
data(TVbo)

# Each of 8 assessors scored each of 12 combinations 2 times
# Let's look at only a single picture and one of the two reps:
# And let us look at the sharpness
TVbosubset <- subset(TVbo,Picture==1 & Repeat==1)[,c(1, 2, 9)]

sharp <- matrix(TVbosubset$Sharpness, nrow=8, byrow=T)
colnames(sharp) <- c("TV3", "TV2", "TV1")
rownames(sharp) <- c("Person 1", "Person 2", "Person 3",
                     "Person 4", "Person 5", "Person 6",
                     "Person 7", "Person 8")

library(xtable)
xtable(sharp)
```

# Bang & Olufsen data in R:

| | TV3 | TV2 | TV1 |
|---|---|---|---|
| Person 1 | 9.30 | 4.70 | 6.60 |
| Person 2 | 10.20 | 7.00 | 8.80 |
| Person 3 | 11.50 | 9.50 | 8.00 |
| Person 4 | 11.90 | 6.60 | 8.20 |
| Person 5 | 10.70 | 4.20 | 5.40 |
| Person 6 | 10.90 | 9.10 | 7.10 |
| Person 7 | 8.50 | 5.00 | 6.30 |
| Person 8 | 12.60 | 8.90 | 10.70 |

# Twoway ANOVA - example

- Same data as for oneway, but now we know that the experiment was split in blocks

| | Group A | Group B | Group C |
|---|---|---|---|
| Block 1 | 2.8 | 5.5 | 5.8 |
| Block 2 | 3.6 | 6.3 | 8.3 |
| Block 3 | 3.4 | 6.1 | 6.9 |
| Block 4 | 2.3 | 5.7 | 6.1 |

  - hence three *Groups* on four *blocks*
  - or three *treatments* on four *persons*
  - or three *varieties* on four *fields* (hence blocks)
  - or similarly

# Twoway ANOVA - example

- Same data as for oneway, but now we know that the experiment was split in blocks

| | Group A | Group B | Group C |
|---|---|---|---|
| Block 1 | 2.8 | 5.5 | 5.8 |
| Block 2 | 3.6 | 6.3 | 8.3 |
| Block 3 | 3.4 | 6.1 | 6.9 |
| Block 4 | 2.3 | 5.7 | 6.1 |

  - hence three *Groups* on four *blocks*
  - or three *treatments* on four *persons*
  - or three *varieties* on four *fields* (hence blocks)
  - or similarly

- *oneway* vs. *twoway* ANOVA

- *Completely randomized design* vs. *Randomized block design*

# Twoway ANOVA - example

- Same data as for oneway, but now we know that the experiment was split in blocks

| | Group A | Group B | Group C |
|---|---|---|---|
| Block 1 | 2.8 | 5.5 | 5.8 |
| Block 2 | 3.6 | 6.3 | 8.3 |
| Block 3 | 3.4 | 6.1 | 6.9 |
| Block 4 | 2.3 | 5.7 | 6.1 |

# Twoway ANOVA - example

- Same data as for oneway, but now we know that the experiment was split in blocks

| | Group A | Group B | Group C |
|---|---|---|---|
| Block 1 | 2.8 | 5.5 | 5.8 |
| Block 2 | 3.6 | 6.3 | 8.3 |
| Block 3 | 3.4 | 6.1 | 6.9 |
| Block 4 | 2.3 | 5.7 | 6.1 |

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

- ANOVA can be used if the observations in each group are (approximately) normally distributed - OR if $n_i$s are large enough (CLT)

# The toy data in R

```
###################################
## Input data and plot

## Observations
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)

## treatments (Groups, varieties)
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

## blocks (persons, fields)
block <- factor(c(1, 2, 3, 4,
                  1, 2, 3, 4,
                  1, 2, 3, 4))

## for later formulas
(k <- length(unique(treatm)))
(l <- length(unique(block)))

## Plots
par(mfrow=c(1,2))

## Plot histogramms by treatments
plot(treatm, y, xlab="Treatments", ylab="y")
## Plot histogrammer by  blocks
plot(block, y, xlab="Blocks", ylab="y")
```

# Agenda

1. Intro: Small example and TV-data from B&O
2. Model
3. Computation - decomposition and the ANOVA table
4. Hypothesis test (F-test)
5. Post hoc analysis
6. Model control
7. A complete example - from the book

# Twoway ANOVA, model

- Express a model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

where the deviations

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

  - $\mu$ is the overall mean
  - $\alpha_i$ is the effect of treatment $i$
  - $\beta_j$ is the level for Block $i$
  - there are $k$ treatments and $l$ blocks
  - $j$ indicates the observations in the groups, from 1 to $n_i$ for treatment $i$

# Estimates of parameters in the model

- We can compute the estimates of the parameters ($\hat{\mu}$ and $\hat{\alpha}_i$, and $\hat{\beta}_j$)

$$\hat{\mu} = \bar{y} = \frac{1}{k \cdot l} \sum_{i=1}^{k} \sum_{j=1}^{l} y_{ij}$$

$$\hat{\alpha}_i = \left( \frac{1}{l} \sum_{j=1}^{l} y_{ij} \right) - \hat{\mu}$$

$$\hat{\beta}_j = \left( \frac{1}{k} \sum_{i=1}^{k} y_{ij} \right) - \hat{\mu}$$

```
####################################
## Compute estimates of parameters in the model

## Sample mean
(muHat <- mean(y))
## Sample mean for each treatment
(alphaHat <- tapply(y, treatm, mean) - muHat)
## Sample mean for each  Block
(betaHat <- tapply(y, block, mean) - muHat)
```

# Agenda

1. Intro: Small example and TV-data from B&O
2. Model
3. Computation - decomposition and the ANOVA table
4. Hypothesis test (F-test)
5. Post hoc analysis
6. Model control
7. A complete example - from the book

# Twoway ANOVA, decomposition and the ANOVA table, Theorem 8.20

- With the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- the total variation in the data can be decomposed:

$$SST = SS(Tr) + SS(Bl) + SSE$$

# Twoway ANOVA, decomposition and the ANOVA table, Theorem 8.20

- With the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- the total variation in the data can be decomposed:

$$SST = SS(Tr) + SS(Bl) + SSE$$

- 'twoway' refers to the fact that there are two factors in the experiment(Two "ways" of the data table)
- 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") (same as for oneway)

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij} - \hat{\mu})^2$$

# Formulas for sums of squares

- Total sum of squares ("the total variance") (same as for oneway)

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{l} (y_{ij} - \hat{\mu})^2$$

- treatment sum of squares ("Variance explained by the treatment part of the model")

$$SS(Tr) = l \cdot \sum_{i=1}^{k} \hat{\alpha}_i^2$$

# Formulas for sums of squares

- Sum of squares for blocks (persons) ("Variance explained by the block part of the model")

$$SS(Bl) = k \cdot \sum_{j=1}^{l} \hat{\beta}_j^2$$

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

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{l} (y_{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu})^2$$

# Agenda

1. Intro: Small example and TV-data from B&O
2. Model
3. Computation - decomposition and the ANOVA table
4. Hypothesis test (F-test)
5. Post hoc analysis
6. Model control
7. A complete example - from the book

# Twoway ANOVA: hypothesis of no effect of treatment, Theorem 8.22

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

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- So we can express the hypothesis:

$$\begin{aligned} H_{0,Tr}: &\quad \alpha_i = 0 \quad \text{for all } i \\ H_{1,Tr}: &\quad \alpha_i \neq 0 \quad \text{for at least one } i \end{aligned}$$

# Twoway ANOVA: hypothesis of no effect of treatment, Theorem 8.22

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

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- So we can express the hypothesis:

$$H_{0,Tr}: \quad \alpha_i = 0 \quad \text{for all } i$$
$$H_{1,Tr}: \quad \alpha_i \neq 0 \quad \text{for at least one } i$$

- Under $H_{0,Tr}$ the following is true:

$$F_{Tr} = \frac{SS(Tr)/(k-1)}{SSE/((k-1)(l-1))}$$

is F-distributed with $k-1$ and $(k-1)(l-1)$ degrees of freedom

# Twoway ANOVA: hypothesis of no effect of persons (blocks), Theorem 8.22

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

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- So we can express the hypothesis

$$\begin{aligned} H_{0,Bl}: &\quad \beta_i = 0 \quad \text{for all } i \\ H_{1,Bl}: &\quad \beta_i \neq 0 \quad \text{for at least one } i \end{aligned}$$

# Twoway ANOVA: hypothesis of no effect of persons (blocks), Theorem 8.22

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

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- So we can express the hypothesis

$$\begin{aligned} H_{0,Bl}: &\quad \beta_i = 0 \quad \text{for all } i \\ H_{1,Bl}: &\quad \beta_i \neq 0 \quad \text{for at least one } i \end{aligned}$$

- Under $H_{0,Bl}$ the following is true:

$$F_{Bl} = \frac{SS(Bl)/(l-1)}{SSE/((k-1)(l-1))}$$

follows an F-distribution with $l-1$ and $(k-1)(l-1)$ degrees of freedom

# F-distribution and treatments hypothesis

```
##################################
## Plot the F distribution and see the critical value for treatments

## Remember, this is "under H0" (that is we compute as if H0 is true):
## Sequence for plot
xseq <- seq(0, 10, by=0.1)
## Plot the density of the  F distribution
plot(xseq, df(xseq, df1=k-1, df2=(k-1)*(l-1)), type="l")
##The critical value for significance level 5 %
cr <- qf(0.95, df1=k-1, df2=(k-1)*(l-1))
## Mark it in the plot
abline(v=cr, col="red")
## The value of the  test statistic
  (Ftr <- (SSTr/(k-1)) / (SSE/((k-1)*(l-1))))
## The p-value hence is:
(1 - pf(Ftr, df1=k-1, df2=(k-1)*(l-1)))
```

# F-distribution and blocks hypothesis

```
##################################
## Plot the F distribution  and see the critical value

## Remember, this is "under H0" (that is we compute as if H0 is true):
## Sequence for plot
xseq <- seq(0, 10, by=0.1)
## Plot the density of the  F distribution
plot(xseq, df(xseq, df1=l-1, df2=(k-1)*(l-1)), type="l")
##The critical value for significance level 5 %
cr <- qf(0.95, df1=l-1, df2=(k-1)*(l-1))
## Mark it in the plot
abline(v=cr, col="red")
## The value of the  test statistic
(Fbl <- (SSBl/(l-1)) / (SSE/((k-1)*(l-1))))
## The p-value hence is:
(1 - pf(Fbl, df1=l-1, df2=(k-1)*(l-1)))
```

# The twoway ANOVA table

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

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

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## treatm     2  30.79   15.40   74.40 5.8e-05 ***
## block      3   3.95    1.32    6.37   0.027 *
## Residuals  6   1.24    0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Agenda

# Post hoc confidence interval

- As for oneway (Use methods 8.9 and 8.10) substitute $(n-k)$ degrees of freedom with $(k-1)(l-1)$ (and use MSE from twoway).
- Can be done with either treatments or blocks

# Post hoc confidence interval

- As for oneway (Use methods 8.9 and 8.10) substitute $(n-k)$ degrees of freedom with $(k-1)(l-1)$ (and use MSE from twoway).
- Can be done with either treatments or blocks
- A single pre-planned confidence interval for the difference between treatment $i$ and $j$ is found as:

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

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

# Post hoc confidence interval

- As for oneway (Use methods 8.9 and 8.10) substitute $(n-k)$ degrees of freedom with $(k-1)(l-1)$ (and use MSE from twoway).
- Can be done with either treatments or blocks
- A single pre-planned confidence interval for the difference between treatment $i$ and $j$ is found as:

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

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

- 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

- A single pre-planned level $\alpha$ hypothesis tests:

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

is carried out as:

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{2}$$

and:

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

where the $t$-distribution with $(k-1)(l-1)$ degrees of freedom is used.

# Post hoc pairwise hypothesis test

- A single pre-planned level $\alpha$ hypothesis tests:

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

is carried out as:

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{2}$$

and:

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

where the $t$-distribution with $(k-1)(l-1)$ degrees of freedom is used.

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

# Agenda

# Variance homogeneity

Look at box-plot to check whether the variability seems different for the groups

```
################################
## Check assumption of  homogeneous variance



## Save the  fit
fit <- lm(y ~ treatm + block)
## Box plot
par(mfrow=c(1,2))
plot(treatm, fit$residuals, y, xlab="Treatment")
## Box plot
plot(block, fit$residuals, xlab="Block")
```

# Normal assumption

Look at qq-normal plot

```
################################
## Check the assumption  of  normality of residuals

## qq-normal plot of residuals
qqnorm(fit$residuals)
qqline(fit$residuals)

## Or with a  Wally plot
require(MESS)
qqwrap <- function(x, y, ...) {qqnorm(y, main="",...);
  qqline(y)}
## Can we see a deviating qq-norm plot?
wallyplot(fit$residuals, FUN = qqwrap)
```

# Agenda

# A complete example - from the book

Introduction to Statistics perbb
Agendas eNotes Course Material Podcast Forum Quiz Admin Logout
ærktøj Dokumentegenskaber...

### Example 8.26 Car tires

In a study of 3 different types of tires ("treatment") effect on the fuel economy, drives of 1000 km in 4 different cars ("blocks") were carried out. The results are listed in the following table in km/l.

| | Car 1 | Car 2 | Car 3 | Car 4 | *Mean* |
|---|---|---|---|---|---|
| Tire 1 | 22.5 | 24.3 | 24.9 | 22.4 | 22.525 |
| Tire 2 | 21.5 | 21.3 | 23.9 | 18.4 | 21.275 |
| Tire 3 | 22.2 | 21.9 | 21.7 | 17.9 | 20.925 |
| *Mean* | 21.400 | 22.167 | 23.167 | 19.567 | 21.575 |

Let us analyse these data with a two-way ANOVA model, but first some explorative plotting:

```
## Collecting the data in a data frame
```

# Agenda

1. Intro: Small example and TV-data from B&O
2. Model
3. Computation - decomposition and the ANOVA table
4. Hypothesis test (F-test)
5. Post hoc analysis
6. Model control
7. A complete example - from the book