University of Zurich^UZH

MAKERERE UNIVERSITY

Data Analysis with R:
Day 5 - Slides

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2018

Simulating in R

- Continuous variable

```r
n <- 100
rnorm(n, mean = 0, sd = 1)       # Normal distribution
```

- Binary variable

```r
rbinom(n, size = 1, prob = 0.4)  # Binomial distribution
```

- Count variable

```r
rpois(n, lambda = 7)             # Poisson distribution
```

- Other options

```r
seq(from = 0, to = 100, by = 1)             # ID
sample(3:30, size = n, replace = TRUE)      # herd size
c(rep(c(1, 2, 3), times = 33), 1)
c(rep(c(1, 2, 3), each = 33), 1)
```

- ...

Plotting in R

Plotting in R

- Continuous data
  - Histogram
  - Boxplot
- Nominal / Ordinal data
  - Barplot
  - Mosaicplot
  - Scatterplots

Exercise 13A and 13B

Exercise 14

Overview: ANOVA and linear models

- Introduction to ANOVA
- How to run an ANOVA in R
- Checking model assumptions in R
- Multiple comparisons options in R
- ANOVA as a special case of a linear model
- The simple linear regression model
- The multiple linear regression model
- Model selection: R^2 and AIC
- Two-way Interactions in R
- Confounding

Hypothesis Testing - One Way ANOVA

We have seen how to perform hypothesis tests when comparing two populations using the two sample t-test. In many analyses we may have multiple populations - for example suppose we have a treatment which has a number of different levels high/medium/low/placebo, or equivalently a number of different treatments. What then is the hypothesis we wish to test?

Is there a difference in the effect of the treatment?

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A &: \quad \text{at least one pair of } \mu_1, \ldots, \mu_k \quad \text{are different} \end{aligned}$$

where $\mu_1, \ldots, \mu_k$ denote the mean effect of treatment levels 1 through k.

One Way ANOVA

Analysis of variance, **ANOVA**, to analyze differences between group means. The observed variance in the outcome variable is partitioned into components attributable to different sources of variation.

ANOVA estimates three sample variances (sum of squares)

- a total variance based on all observation deviations from the grand mean
- a variance based on the group mean deviations from the grand mean
- an (error) variance based on all the observation deviations from their group mean

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \underset{\substack{\uparrow \\ \text{Variance} \\ \text{between groups}}}{\sum_i n_i (\bar{x}_i - \bar{x})^2} + \underset{\substack{\uparrow \\ \text{Variance} \\ \text{within groups}}}{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}$$

$\bar{x}$= grand mean

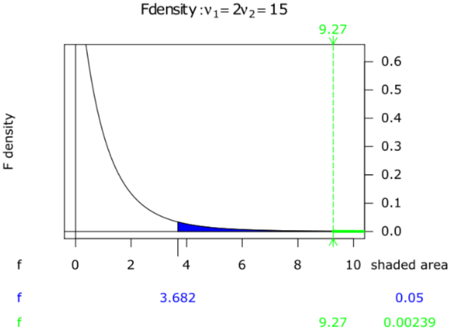
$\bar{x}_i$= group mean

F-test / F-distribution

An **F-test**, a statistical test, in which the test statistic has an F-distribution under the null hypothesis is used to assess statistical significance in an ANOVA.

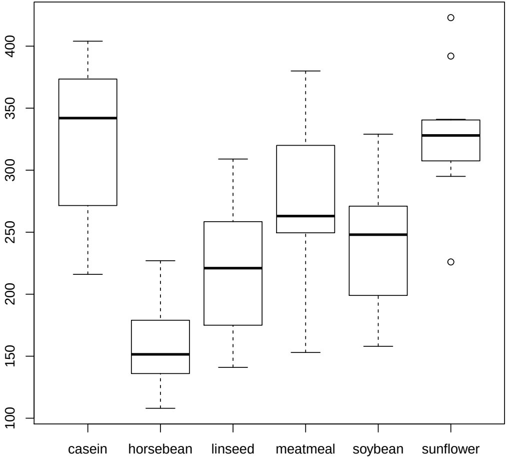
- degrees of freedom

$$F = \frac{\textit{variance between groups}}{\textit{variance within groups}}$$



ANOVA in R with `chickwts`

```
data(chickwts)
boxplot(chickwts$weight ~ chickwts$feed)
```



ANOVA in R with `chickwts`

```
# aov.mod <- aov(chickwts$weight ~ chickwts$feed)
aov.mod <- aov(weight ~ feed, data = chickwts)
# What objects can we extract from a anova model?
objects(aov.mod)

##  [1] "assign"        "call"          "coefficients"  "contrasts"
##  [5] "df.residual"   "effects"       "fitted.values" "model"
##  [9] "qr"            "rank"          "residuals"     "terms"
## [13] "xlevels"

#
summary(aov.mod)

##             Df Sum Sq Mean Sq F value   Pr(>F)
## feed         5 231129   46226   15.37 5.94e-10 ***
## Residuals   65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overview: Checking model assumptions

- mean(residuals) = 0
- Residuals are normally distributed (qqnorm, qqplot)
- Variance homoscedasticity (Bartlett & Levene's Test)
- Cook's distance: Influential data points
- Any pattern(s)?

Checking model residuals: mean(residuals) = 0

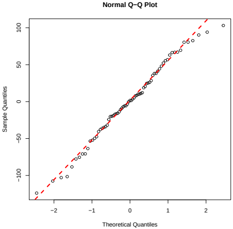
”Unexplained rest of the model”

```
# What are residuals?
chickwts$residuals <- residuals(aov.mod)
tapply(chickwts$weight, chickwts$feed, mean)

##    casein horsebean   linseed  meatmeal   soybean sunflower
##  323.5833  160.2000  218.7500  276.9091  246.4286  328.9167

chickwts[c(1:3),]

##   weight      feed residuals
## 1    179 horsebean      18.8
## 2    160 horsebean      -0.2
## 3    136 horsebean     -24.2

# Save residuals to an objects and check mean of residuals
aov.mod.resid <- residuals(aov.mod)
mean(aov.mod.resid)

## [1] 7.573045e-16
```

Checking model residuals: Residuals normally distributed

”Unexplained rest of the model”

```
par(mfrow=c(1,1))
qqnorm(aov.mod.resid)
qqline(aov.mod.resid, col = "red", lwd = 3, lty = 2)
```



```
# Shapiro-Wilk test (dependent on sample size --> limited use)
shapiro.test(aov.mod.resid)


##
##  Shapiro-Wilk normality test
##
## data:  aov.mod.resid
## W = 0.98616, p-value = 0.6272
```

Checking model residuals: Variance homoscedasticity (1/3)

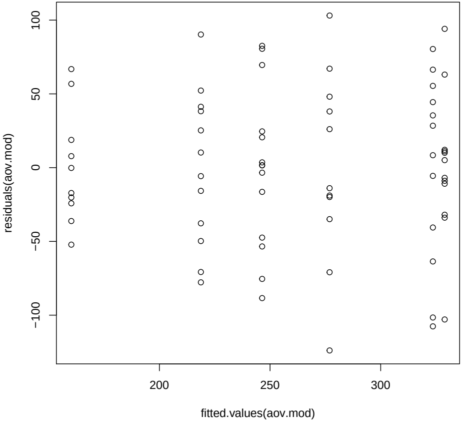
Hypothesis tests

```
# Bartlett Test
bartlett.test(chickwts$weight ~ chickwts$feed)

##
##  Bartlett test of homogeneity of variances
##
## data:  chickwts$weight by chickwts$feed
## Bartlett's K-squared = 3.2597, df = 5, p-value = 0.66

# Levene's Test
# library("Rcmdr")
# levene.test(chickwts$weight ~ chickwts$feed)
```

Checking model residuals: Variance homoscedasticity (2/3)

Graphical interpretation is better!

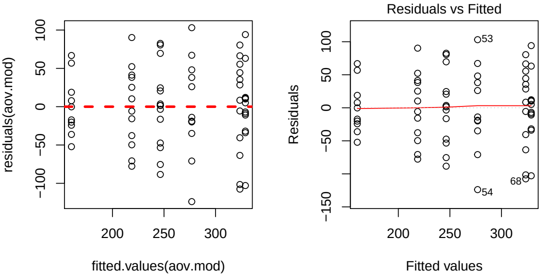
```
# Plot fitted against residual values
plot(fitted.values(aov.mod), residuals(aov.mod))
```



Checking model residuals: Variance homoscedasticity (3/3)

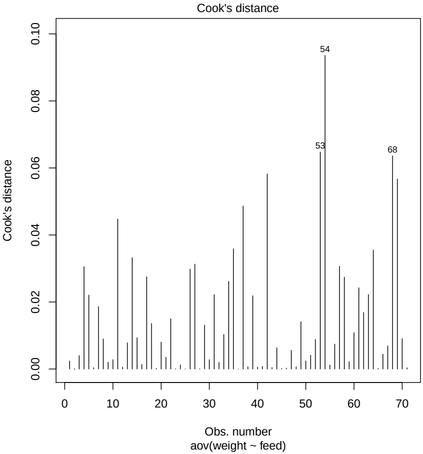
Graphical interpretation is better!

```
# Plot fitted against residual values
par(mfrow=c(1,2), pty="s", mar = c(10, 4, 1, 2))
plot(fitted.values(aov.mod), residuals(aov.mod))
abline(h = 0, col = "red", lwd = 3, lty = 2)
plot(aov.mod, which=1)
```



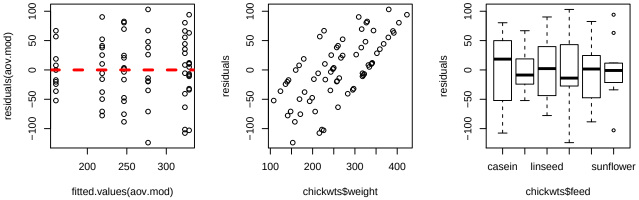
Checking model residuals: Cook's distance

```r
# Plot fitted against residual values
# Cut-off at 3
par(mfrow=c(1,1), pty="s", mar = c(5, 4, 4, 2))
plot(aov.mod, which=4)
```



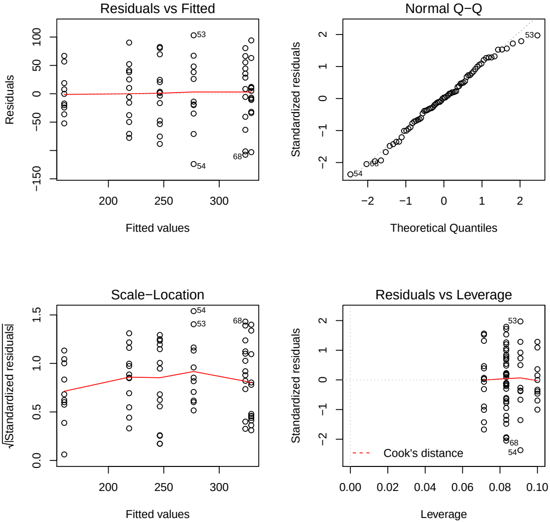
Checking for potential patterns

```
par(mfrow=c(1,3), pty="s", mar = c(10, 4, 1, 2))
plot(fitted.values(aov.mod), residuals(aov.mod))
abline(h = 0, col = "red", lwd = 3, lty = 2)
# Plot residuals against variables from the model
plot(chickwts$weight, residuals(aov.mod), ylab = "residuals")
plot(chickwts$feed, residuals(aov.mod),
     xlab = "chickwts$feed", ylab = "residuals")
```



Plot anova objects

```r
par(mfrow=c(2, 2))
plot(aov.mod)
```



Multiple comparisons options in R

So far, we know there is difference between the `feed` types. However, we do not yet know which `feed` type differ.
In principal, we could do multiple t-tests, BUT ... we would use our data several times. Classically, we choose an α-level of 5 %, in cases of multiple comparisons we are facing the familywise error rate:

$$FWE \leq 1 - (1 - \alpha)^n,$$

where α = 5 % and n = number of comparisons.

```
alpha = 0.05
1 - (1 - alpha)^1   # n = 1     # [1] 0.05
1 - (1 - alpha)^5   # n = 5     # [1] 0.2262191
1 - (1 - alpha)^10  # n = 10    # [1] 0.4012631
```

Multiple comparisons options in R

Hence, we are better of using one of the following procedures to adjust for multiple comparisons:

- **Bonferroni**: *p*-value correction by testing each individual hypothesis at a significance level of $\frac{\alpha}{m}$, where α is the desired overall α level and m is the number of hypotheses.
- **Dunnett**: Multiple comparisons of each group to a reference.
- **Tukey Honest Significant Differences** (Homogeneous subgroups): Multiple comparisons of all possible combinations.
- ...

Multiple comparisons options in R: Bonferroni

```
aov.mod <- aov(weight ~ feed, data = chickwts)
pairwise.t.test(chickwts$weight, chickwts$feed, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  chickwts$weight and chickwts$feed
##
##           casein  horsebean linseed meatmeal soybean
## horsebean 2.1e-09 -         -       -        -
## linseed   1.5e-05 0.01522   -       -        -
## meatmeal  0.04557 7.5e-06   0.01348 -        -
## soybean   0.00067 0.00032   0.20414 0.17255  -
## sunflower 0.81249 8.2e-10   6.2e-06 0.02644  0.00030
##
## P value adjustment method: none
```

```
pairwise.t.test(chickwts$weight, chickwts$feed, p.adj = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  chickwts$weight and chickwts$feed
##
##           casein  horsebean linseed meatmeal soybean
## horsebean 3.1e-08 -         -       -        -
## linseed   0.00022 0.22833   -       -        -
## meatmeal  0.68350 0.00011   0.20218 -        -
## soybean   0.00998 0.00487   1.00000 1.00000  -
## sunflower 1.00000 1.2e-08   9.3e-05 0.39653  0.00447
##
## P value adjustment method: bonferroni
```

Multiple comparisons options in R: Dunnett

```
library("multcomp")
# compares always to baseline levels (here: casein) --> saves degrees of freedom
dunnett <- glht(aov.mod, linfct = mcp(feed = "Dunnett"))
summary(dunnett)


##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = weight ~ feed, data = chickwts)
##
## Linear Hypotheses:
##                         Estimate Std. Error t value Pr(>|t|)
## horsebean - casein == 0 -163.383     23.485  -6.957  < 0.001 ***
## linseed - casein == 0   -104.833     22.393  -4.682  < 0.001 ***
## meatmeal - casein == 0   -46.674     22.896  -2.039  0.16717
## soybean - casein == 0    -77.155     21.578  -3.576  0.00304 **
## sunflower - casein == 0    5.333     22.393   0.238  0.99945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Multiple comparisons options in R: Tukey

Tukey Honest Significant Differences

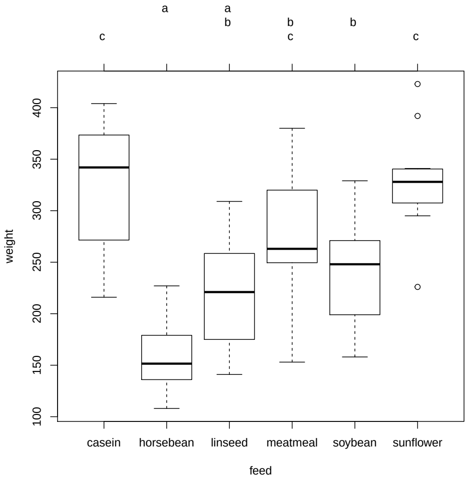
```
library("multcomp")
# compares all factor levels
tukey <- glht(aov.mod, linfct = mcp(feed = "Tukey"))
summary(tukey)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = weight ~ feed, data = chickwts)
##
## Linear Hypotheses:
##                             Estimate Std. Error t value Pr(>|t|)
## horsebean - casein == 0    -163.383     23.485  -6.957  < 0.001 ***
## linseed - casein == 0      -104.833     22.393  -4.682  < 0.001 ***
## meatmeal - casein == 0      -46.674     22.896  -2.039  0.33201
## soybean - casein == 0       -77.155     21.578  -3.576  0.00831 **
## sunflower - casein == 0       5.333     22.393   0.238  0.99989
## linseed - horsebean == 0     58.550     23.485   2.493  0.14100
## meatmeal - horsebean == 0   116.709     23.966   4.870  < 0.001 ***
## soybean - horsebean == 0     86.229     22.710   3.797  0.00417 **
## sunflower - horsebean == 0  168.717     23.485   7.184  < 0.001 ***
## meatmeal - linseed == 0      58.159     22.896   2.540  0.12740
## soybean - linseed == 0       27.679     21.578   1.283  0.79295
## sunflower - linseed == 0    110.167     22.393   4.920  < 0.001 ***
## soybean - meatmeal == 0     -30.481     22.100  -1.379  0.73877
## sunflower - meatmeal == 0    52.008     22.896   2.271  0.22044
## sunflower - soybean == 0     82.488     21.578   3.823  0.00393 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Multiple comparisons options in R: Tukey

Tukey Honest Significant Differences with homogeneous subgroups

```
# summary(tukey)           # standard display
tukey.cld <- cld(tukey)    # letter-based display
# the cld(...) function sets up a compact letter display of all pair-wise comparisons
par(mfrow=c(1,1), mar=c(5, 4, 5, 2))
plot(tukey.cld)
```



Exercise 15