University of Zurich^{UZH}

MAKERERE UNIVERSITY

Data Analysis with R:
Lecture Slides (all)

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2018

How to deal with missing values in R? (1/3)

- In R, missing values are represented by the symbol **NA** (not available).

- Impossible values (e. g., dividing by zero) are represented by the symbol **NaN** (not a number).

- Ask yourself why a **NA** and / or **NaN** occurs!

How to deal with missing values in R? (2/3)

- Testing for Missing Values

```r
vec1 <- c(1, 2, 3, NA)
is.na(vec1) # returns a vector (FALSE, FALSE, FALSE, TRUE)
# The TRUE indicates the position of the NA in vec1.
```

- Recoding Values to Missing

```r
# recode specific values (e. g. 0.001) to missing for variable x
# select rows where x is 0.001 and recode value in column x with NA
tmp.row <- which(dat$x == 0.001)
dat$x[tmp.row] <- NA
```

How to deal with missing values in R? (3/3)

- Excluding Missing Values from specific function calls

```
a <- c(1, 2, NA, 3)
mean(a) # returns NA
mean(a, na.rm=TRUE) # returns 2
```

- Check for complete cases with function
  complete.cases(...)

```
# list rows of data that have missing values
dat[!complete.cases(dat),]
subdat <- dat[complete.cases(dat),]
```

- Create new dataset without missing data with function
  na.omit(...)

```
new.dat <- na.omit(dat)
```

How to check your data for plausibility?

- Ask yourself what can go wrong?

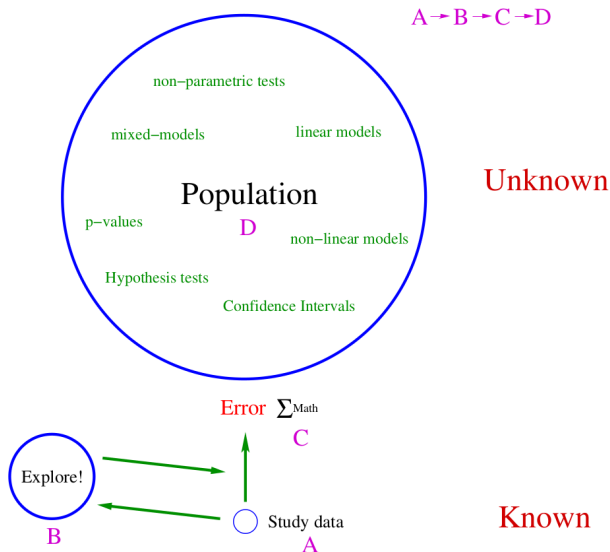- Implausible values?

- Impossible values?

- Logical errors?

Exercise 9A and 9B
Plausibility Checks & Missing Values

Exercise 10

Lecture Slides for Day 4

Basic Statistical Tests

Study data is collected for a purpose - to answer one or more specific scientific questions. The classical way to perform a formal statistical analyses of these data is to formulate these research questions into statistical **hypothesis tests**.

In this section we will go through a simple example in detail to highlight some of the important concepts - the general approach for more complex analyses is exactly same. *Note: the precise technical details are much less important than the concepts!*

Simple Example - One Population

After six weeks will the mean weight of a chicken be more than 250 grams?

There are 71 observations in `chickwts` from which to answer this question. This can be formulated into a statistical hypothesis test. A hypothesis test has two parts, the null hypothesis and the alternative hypothesis. This is typically written as follows:

$$H_0: \quad \mu \leq 250$$
$$H_A: \quad \mu > 250$$

where μ is the mean weight in the **population** of chickens from which the sample of 71 chickens was drawn. Remember - we know the mean weight in the sample of chickens is greater than 250 it is the **population** of chickens which we are interested in.

After six weeks will the mean weight of a chicken be at least 250 grams?

$$H_0: \quad \mu \leq 250$$
$$H_A: \quad \mu > 250$$

The null hypothesis (H_0) is the default situation, sometimes called the "state of nature". In a treatment-control trial, H_0 is typically that the effect of the treatment is not different from the control. In this example our default position is that the mean weight of chickens is ≤ 250. This is called a single-sided hypothesis test.

Simple Example - One Population

We now analyse the 71 observations to see whether there is evidence to **REJECT** the null hypothesis H_0, and if the null hypothesis is rejected then we can conclude that the available evidence supports the alternative hypothesis.

```r
t.test(chickwts$weight, mu = 250, alternative = "greater")
t.test(chickwts$weight, mu = 250, alternative = "less")
```

Note that hypothesis testing is concerned with finding evidence in support of the null hypothesis H_0 - the default situation - rather than evidence in favour of the alternative hypothesis.

One Sample t-test

For the chicken weights data an appropriate formal analyses is to use a **one-sample t-test**, why this test is appropriate will be discussed later. This analysis involves calculating a simple summary statistic - called a t-statistic - which we do entirely from the observed data.

$$T_{obs} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s the sample standard deviation and μ is the population mean in the null hypothesis which we wish to test for. We then look up the value of T_{obs} in a set of statistical tables/computer to see what the answer is to our research question.
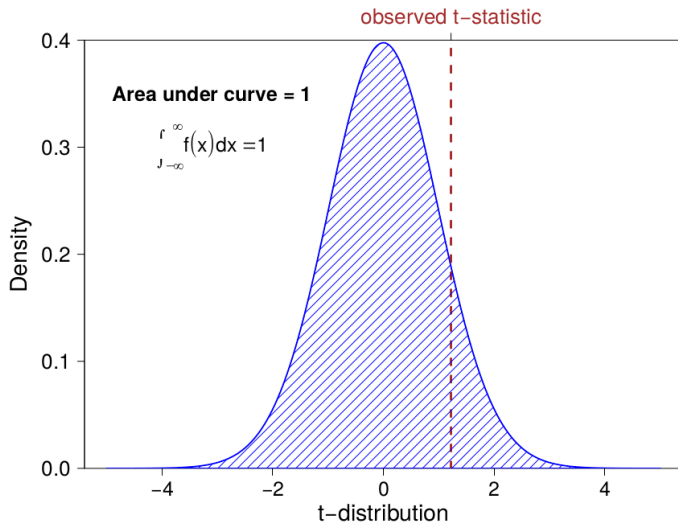
Important concept - sampling

Why is $T_{obs} = \frac{\bar{x} - \mu}{s/\sqrt{(n)}}$ called a t-statistic?

If another sample of 71 chickens from the same population were weighed then the values for $\bar{x}$ and s would be different, and hence the value for T_{obs}. If this was repeated many times and a histogram/Q-Q/P-P plot produced of the values for T_{obs} then this would follow the shape of a known distribution - **student-t probability distribution**. It is this piece of mathematics - knowing what the sampling distribution of T_{obs} is - which allows us to infer information about the population of chickens from which our original 71 chickens were sampled - without actually having to collect lots and lots of other samples of chickens! Mathematical theory is used to fill this data gap.

chickwts: t-test

$$T_{obs} = \frac{261.31 - 250}{78.07 / \sqrt{71}} = 1.22$$

Put the values for the sample mean and standard deviation into the t-statistic formula along with the $\mu = 250$. We now look up the value of this in a t-distribution reference table. All this calculation will be done for you in R but it is important to understand the general process as this is the same for hypothesis testing in other more complex analyses.

One Sample, one-sided, t-test

Important concept - p-values

- The result of a hypothesis test is usually communicated in the form of a *p*-value
- The interpretation of a *p*-value is of crucial importance - it is the *probability that the test statistic takes values at least as extreme as that observed **assuming that** H_0 **is true***
- Exactly what **at least as extreme as** refers to depends on the alternative hypothesis H_A.
- This may sound rather abstract but it is usually obvious in practice
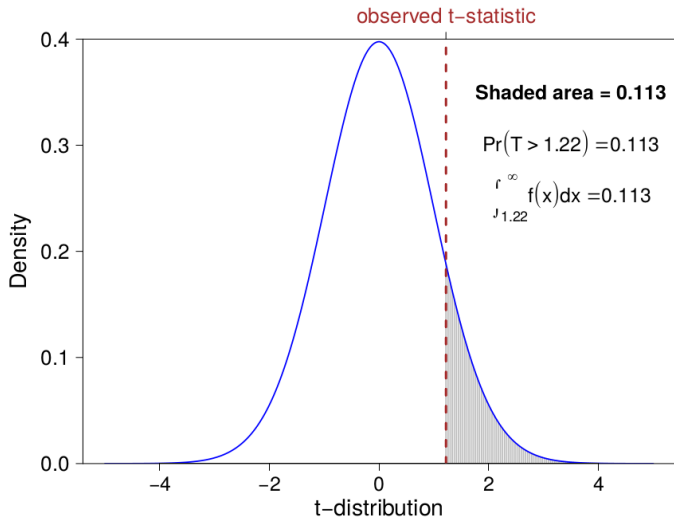
Simple Example - One Population

After six weeks will the mean weight of a chicken be at least 250 grams?

$$H_0: \quad \mu \leq 250$$
$$H_A: \quad \mu > 250$$

The alternative hypothesis is $\mu > 250$ so in this test **at least as extreme as** in the definition of a *p*-value is the probability of observing a t-statistic which is > 1.22 **assuming that H_0 is true** - this is why 250 was used for μ when calculating T_{obs}.

One Sample, one-sided, t-test

observed t−statistic

Shaded area = 0.113

$\Pr(T > 1.22) = 0.113$

$\int_{1.22}^{\infty} f(x)dx = 0.113$

Research Question

The purpose of this hypothesis test analysis is to answer a very specific scientific question:

After six weeks will the mean weight of a chicken be more than 250 grams?

So what is our answer?

The *p*-value for this hypothesis test is 0.113. Based on this value we can either **reject** H_0 and conclude that the mean weight of chickens in the population is likely to be greater than 250 grams or else we can **accept** H_0 where the mean chicken weight is less than 250 grams.

Research Question - be pragmatic with p-values

By convention a p-value of less than 0.05 is considered to provide **reasonable evidence** for rejecting H_0. A p-value of between 0.05 and 0.1 might be considered as **weak evidence** against H_0. Values of less that 0.01 are generally considered as **very strong evidence** for rejecting H_0. It is **always** best to provide a p-value in any analyses to let the reviewer/client see the strength of evidence rather than simply claiming statistically significant findings!

Communicating Results of Hypothesis Tests

Transparency is essential - the devil can be in the detail - which at the very least should comprise:

i. what hypothesis was being tested - be clear and precise
ii. what statistical test was used
iii. what the *p*-value is
iv. what the treatment effect is (more later).

This is particularly crucial if the analyses are to be given to someone *else* to then make a judgment on the scientific significance.

Two-sided Tests: One Population

After six weeks will the mean weight of a chicken be equal to 250 grams?

This is now a two sided hypothesis test:

$$H_0: \quad \mu \ = \ 250$$
$$H_A: \quad \mu \ \neq \ 250$$

This time the hypothesis test is asking how much evidence is there in our sample data to conclude that in the population of all chickens the mean weight is not equal to 250 grams.

Two-sided Tests: One Population

```r
# t.test(chickwts$weight, mu = 250, alternative = "two.sided")
t.test(chickwts$weight, mu = 250)

##
##  One Sample t-test
##
## data:  chickwts$weight
## t = 1.2206, df = 70, p-value = 0.2263
## alternative hypothesis: true mean is not equal to 250
## 95 percent confidence interval:
##  242.8301 279.7896
## sample estimates:
## mean of x
##  261.3099
```

Two-sided Tests

A two-sided test is similar to a one-sided test - the key difference is in what is now defined as **at least as extreme** in the definition of the *p*-value. This time the alternative hypothesis refers to observing a value of **either $\bar{x} > 250$ or $\bar{x} < -250$ assuming that H_0 is true**, which using the t-test approach is equivalent to the probability of observing $T_{obs} > 1.22$ or $T_{obs} < -1.22$ which we can again look up in reference tables.

One Sample, two-sided, t-test
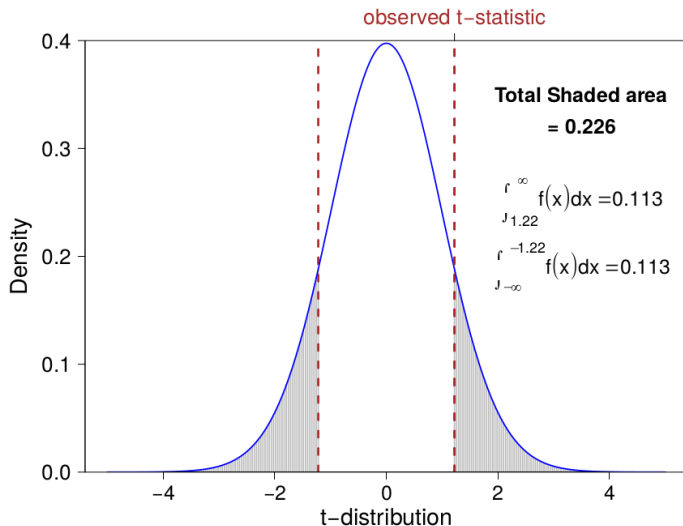
Two-sided Tests

- The two-sided t-test has a *p*-value which is exactly double the single sided test.
- Think! - Intuitively the *p*-value should be less for a single sided test as the research question you are asking is much narrower e.g. greater than 250 grams, as opposed to whether the mean chicken weight might be **either** less than 250 grams **or greater** than 250 grams.

  $\rightarrow$ You are using the same amount of information (71 observations) to answer a narrower research question and so all else being equal you should expect a "more powerful" analyses (e.g. a lower *p*-value all else being equal)

Exercise 11

Chi-square Test

There are two very commonly used statistical tests for testing dependence between two categorical variables: Chi-squared test & Fisher's exact test.

To test independence of rows and columns

Risk Factor	Disease		Total
	+	-	
+	a	b	a+b
-	c	d	c+d
Total	a+c	b+d	n = a+b+c+d

- Assumptions: a, b, c, d must have at least 5 observations!

$$\chi^2 = \frac{n * (a * d - b * c)^2}{(a + b) * (c + d) * (a + c) * (b + d)}$$

Test statistic chi-square

$$\chi^2 = \Sigma \frac{(O - E)^2}{E^2}$$

The Chi-square Distribution
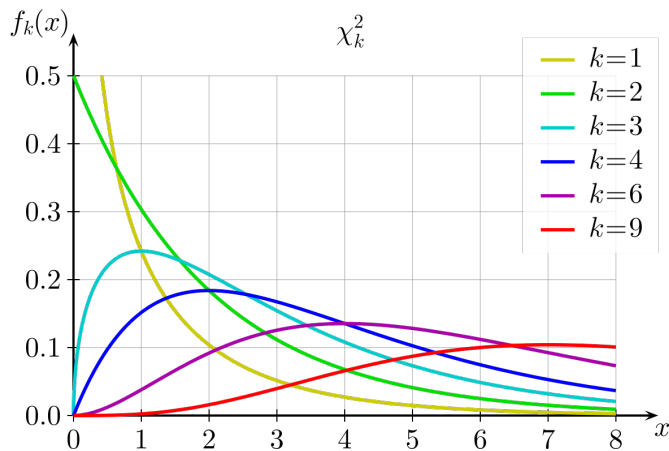
Exact Fisher Test: Permutation Test

	Success	Failure	Total
Therapy	7	2	9
New Therapy	2	8	10
Total	9	10	19

7	2
2	8

$P = 9!*10!*9!*10! / 19!*7!*2!*2!*8! = 0.01754$

8	1
1	9

$P = 9!*10!*9!*10! / 19!*8!*1!*1!*9! = 0.00097$

9	0
0	10

$P = 9!*10!*9!*10! / 19!*9!*0!*0!*10! = 0.00001$

For a one-sided test: p = 0.01754+0.00097+0.00001 = 0.01852

Exercise 12

Simulating in R

- Continuous variable

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

- Binary variable

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

- Count variable

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

- Other options

```
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)
```

- ...

- ctree(...) is a non-parametric class of regression trees embedding tree structured regression models into a well defined theory of conditional inference procedures.

- **Recursive partitioning algorithm:**
  - partition the observations by univariate splits in a recursive way
  - fit a constant model in each cell of the resulting partition
  - $\rightarrow$ uses an information measure of node impurity to select the variables showing the best split

Side remark: Conditional Inference Trees (2/2)

```
## Warning:  unable to access index for repository
https://cloud.r-project.org/src/contrib:
##  cannot open URL
'https://cloud.r-project.org/src/contrib/PACKAGES'
## Warning:  package 'partykit' is not available (for
R version 3.4.4)
## Warning:  unable to access index for repository
https://cloud.r-project.org/src/contrib:
##  cannot open URL
'https://cloud.r-project.org/src/contrib/PACKAGES'
## Warning:  package 'party' is not available (for R
version 3.4.4)
```

```r
library("partykit")
library("party")
data("ToothGrowth")
my.tooth.tree <- ctree(len ~ ., data = ToothGrowth)
plot(my.tooth.tree, tp_args = list(id = FALSE))
```

Plotting in R

Plotting in R

- Continuous data
  - Histogram
  - Boxplot

- Nominal / Ordinal data
  - Barplot
  - Mosaicplot
  - Scatterplots

Exercise 13A and 13B

Exercise 14

Lecture Slides for Day 5

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?

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

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

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i n_i (\bar{x}_i - \bar{x})^2 + \sum_i \sum_j (x_{ij} - \overline{x}_i)^2$$

Variance between groups

Variance within groups

$\bar{x}$ = 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

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

Fdensity : $\nu_1 = 2 \nu_2 = 15$

f		
f	3.682	0.05
f	9.27	0.00239
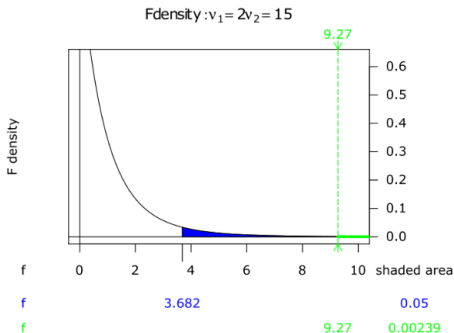
ANOVA in R with `chickwts`

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

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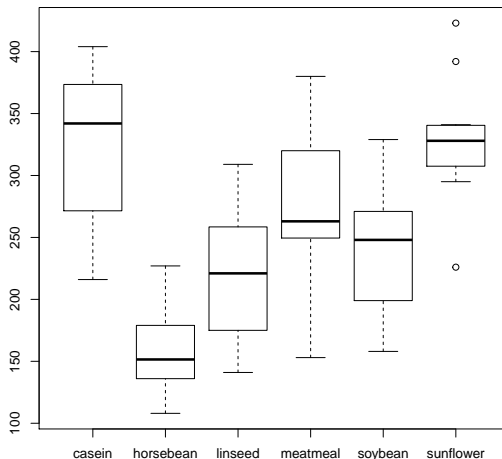
```
# aov.mod <- aov(chickwts$weight ~ chickwts$feed)
aov.mod <- aov(weight ~ feed, data = chickwts)
# What objects can we extract from a anova model?
objects(aov.mod)

## [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"

R

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


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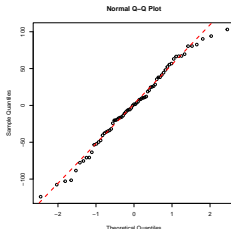
```r
# Shapiro-Wilk test (dependent on sample size --> limited use)
shapiro.test(aov.mod.resid)

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

```
# 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!

```r
# Plot fitted against residual values
plot(fitted.values(aov.mod), residuals(aov.mod))
```
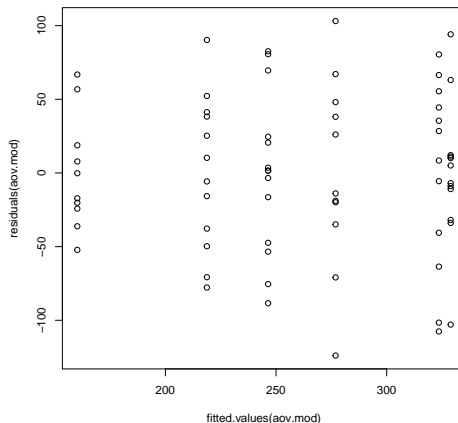
Checking model residuals: Variance homoscedasticity (3/3)
## Graphical interpretation is better!

```r
# 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)
```
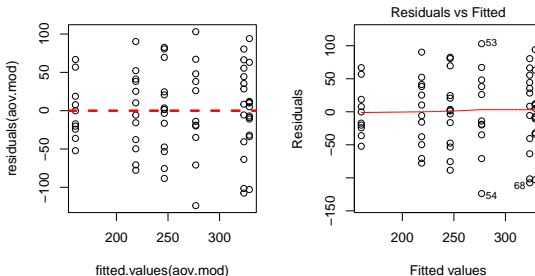
# Checking model residuals: Cook's distance

```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)
```


Cook's distance

Obs. number
aov(weight ~ feed)
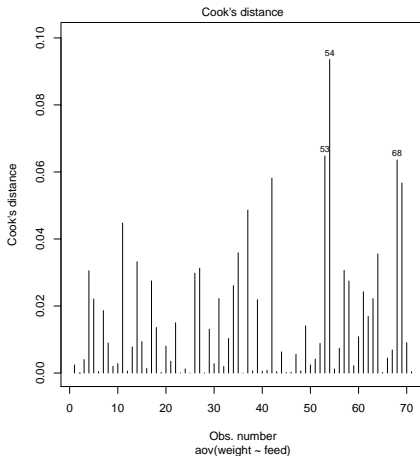
Checking for potential patterns

```r
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")
```
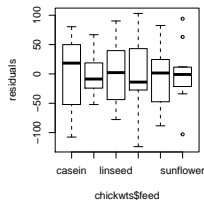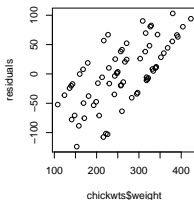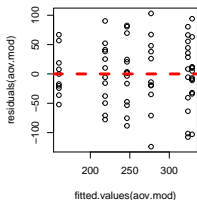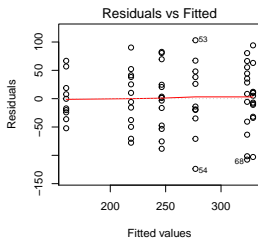
Plot anova objects

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

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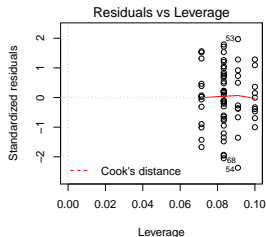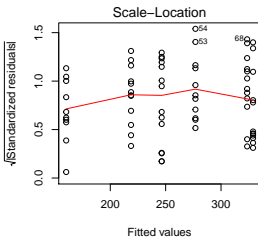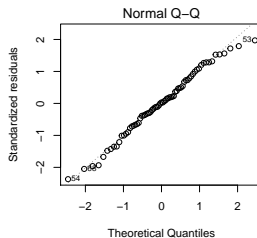
So far, we know there is difference between the `feed` types.
However, we do not yet know which `feed` type differ.
In principal, we could do multiple t-tests, BUT ... we would use our data several times. Classically, we choose an α-level of 5 %, in cases of multiple comparisons we are facing the familywise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

```r
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  0.16711
## soybean - casein == 0    -77.155     21.578  -3.576  0.00311 **
## 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)
```

```r
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  0.33196
## soybean - casein == 0       -77.155    21.578  -3.576  0.00835 **
## sunflower - casein == 0       5.333    22.393   0.238  0.99989
## linseed - horsebean == 0     58.550    23.485   2.493  0.14096
## meatmeal - horsebean == 0   116.709    23.966   4.870  < 0.001 ***
## soybean - horsebean == 0     86.229    22.710   3.797  0.00413 **
## sunflower - horsebean == 0  168.717    23.485   7.184  < 0.001 ***
## meatmeal - linseed == 0      58.159    22.896   2.540  0.12736
## soybean - linseed == 0       27.679    21.578   1.283  0.79297
## 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  0.22042
## sunflower - soybean == 0     82.488    21.578   3.823  0.00382 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
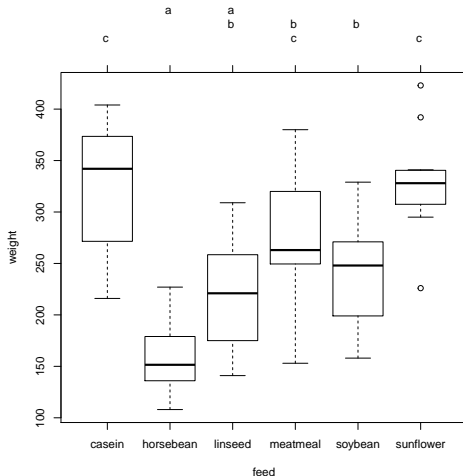
Multiple comparisons options in R: Tukey
Tukey Honest Significant Differences with homogeneous subgroups

```r
# 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

So far we have considered how to determine whether statistically significant differences exist between different "feed" groups (factors).

The explanatory variable has been categorical. We have not yet considered continuous explanatory variables.

ANOVA and linear regression are both a special cases of **linear models**.

Simple linear regression (1/3)

A **simple linear regression** fits a straight line through a set of data points. This straight line is fitted in a way that makes the sum of the squared residuals (the vertical distances between each data point and the fitted line) as small as possible.

Simple linear regression (2/3)

In a simple linear regression model for *n* data points
$(x_i, y_i), i = 1, ...n$ the following equation is used:

$$y_i = \alpha + \beta x_i + \epsilon,$$

where

- α: intercept or constant
- β: slope, regression coefficient (effect size)
- ϵ: error, residuals

Simple linear regression (3/3)

The goal is to find the equation of the straight line

$$f(x) = y = \alpha + \beta x,$$

which would provide a "best" fit for the data points (least square approach).

Assumptions of a linear regression

A **simple linear regression** is based on several assumptions which should be checked carefully.

- linearity of the relationship between the explanatory (independent) and the outcome (dependent) variable
- normality of the residuals
- independence
- constant variance (homoscedasticity)

Linear regression model in R

```r
data(water)
# mod.hard <- lm(mortality ~ hardness, data = water)
mod.hard <- lm(water$mortality ~ water$hardness)
summary(mod.hard)

##
## Call:
## lm(formula = water$mortality ~ water$hardness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -348.61 -114.52   -7.09  111.52  336.45
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1676.3556    29.2981  57.217  < 2e-16 ***
## water$hardness   -3.2261     0.4847  -6.656 1.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143 on 59 degrees of freedom
## Multiple R-squared:  0.4288,Adjusted R-squared:  0.4191
## F-statistic:  44.3 on 1 and 59 DF,  p-value: 1.033e-08
```

Checking linear model assumptions

```
par(mfrow=c(2,2))
plot(mod.hard)
```
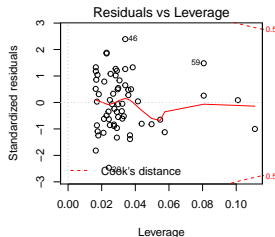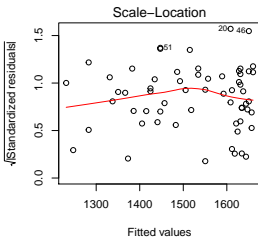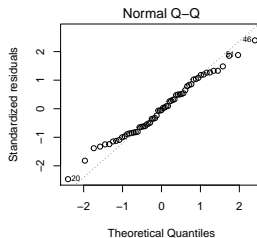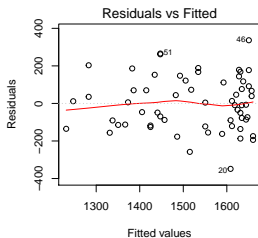
Linear model vs. t.test(...) in R

```
mod.loc <- lm(mortality ~ location, data = water)
coef(mod.loc)

##   (Intercept) locationSouth
##      1633.6000     -256.7923


t.test(water$mortality ~ water$location)

##
##  Welch Two Sample t-test
##
## data:  water$mortality by water$location
## t = 7.1427, df = 53.29, p-value = 2.584e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  184.6919 328.8928
## sample estimates:
## mean in group North mean in group South
##           1633.600           1376.808
```

The multiple linear model
An extension of the simple linear model (1/2)

If several explanatory variables are of interest, instead of performing multiple simple regressions, a **multiple linear regression** or **multivariable** approach is appropriate.

- the same assumptions as for simple linear regressions should be checked
- collinearity might be an issue (see vif(...) function from package usdm)
- model comparison (AIC) (...discussed later)

The multiple linear model
An extension of the simple linear model (2/2)

Simple linear regression model:

$$y_i = \alpha + \beta x_i + \epsilon$$

Multiple linear regression model:

$$y_i = \alpha + \beta_1 x1_i + \beta_2 x2_i + ... + \epsilon, \tag{1}$$
$$y_i = \alpha + \beta_1 x1_i * \beta_2 x2_i + ... + \epsilon,$$

where

- α: intercept or constant
- β_1: slope1, regression coefficient (effect size)
- β_2: slope2, regression coefficient (effect size)
- $\beta_1 * \beta_2$: interaction effect between x_1 and x_2
- ϵ: error, residuals

The multiple linear model
Interpretation of model coefficients

- α: intercept or constant
- β_1: slope1, regression coefficient (effect size)
- β_2: slope2, regression coefficient (effect size)
- $\beta_1 * \beta_2$: interaction effect between x_1 and x_2
- ϵ: error, residuals

- $\rightarrow \beta_1$ describes the number of units of a change in the outcome variable y as x_1 changes by one unit, x_2 being held constant.
- $\rightarrow \beta_2$ describes the number of units of a change in the outcome variable y as x_2 changes by one unit, x_1 being held constant.

The multiple linear model in R

```
mod.hard.loc <- lm(mortality ~ hardness + location, data = water)
summary(mod.hard.loc)

##
## Call:
## lm(formula = mortality ~ hardness + location, data = water)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -222.959 -77.281    7.143   90.751  307.836
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1695.4371    25.3285  66.938  < 2e-16 ***
## hardness        -2.0341     0.4829  -4.212 8.93e-05 ***
## locationSouth -176.7108    36.8913  -4.790 1.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.1 on 58 degrees of freedom
## Multiple R-squared:  0.5907,Adjusted R-squared:  0.5766
## F-statistic: 41.86 on 2 and 58 DF,  p-value: 5.601e-12
```

Model selection

- Multiple R^2: the larger, the better

- Akaike criterion (AIC): the smaller, the better

Nested models

Two models are nested if one of them is a particular case of the other one: the simpler model can be obtained by setting some coefficients of the more complex model to particular values.

```r
mod.hard <- lm(mortality ~ hardness, data = water)              # mod 1
mod.loc <- lm(mortality ~ location, data = water)              # mod 2
mod.hard.loc <- lm(mortality ~ hardness + location, data = water) # mod 3
```

Among the 3 above models ...

- which ones are nested?

- which ones are not nested?

R^2 is a measure of fit quality:

$$R^2 = \frac{\text{SSRegression}}{\text{SSTotal}}$$
$$R^2 = 1 - \frac{\text{SSError}}{\text{SSTotal}}$$

CAREFUL:

- The SSError always decreases as more predictors are added to the model.
- R^2 always increases and can be artificially large.

Akaike criterion (AIC)

- Model fit (R^2) always improves with model complexity. We would like to strike a good balance between **model fit** and **model simplicity**.

- AIC combines a measure of model fit with a measure of model complexity: The smaller, the better.

For a given data set and a given model:

$$AIC = -2 \cdot \log(L) + 2p$$

L stands for the likelihood. p stands for the number of parameters in the models (penalizes complex models).

Model selection strategy for AIC

- Consider a number of candidate models.
  (They need not be nested.)

- Calculate their AIC.

- Choose the model(s) with the smallest AIC.

$\rightarrow$ **CAREFUL**: The absolute value of AIC is meaningless. The relative AIC values, between models, is meaningful.

Model selection with AIC in R

```
mod1 <- lm(mortality ~ hardness, data = water)
mod2 <- lm(mortality ~ location, data = water)
mod3 <- lm(mortality ~ hardness + location, data = water)
AIC(mod1, mod2, mod3)

##      df      AIC
## mod1  3 782.5692
## mod2  3 778.5186
## mod3  4 764.2366


round(AIC(mod1, mod2, mod3), 2)

##      df    AIC
## mod1  3 782.57
## mod2  3 778.52
## mod3  4 764.24
```
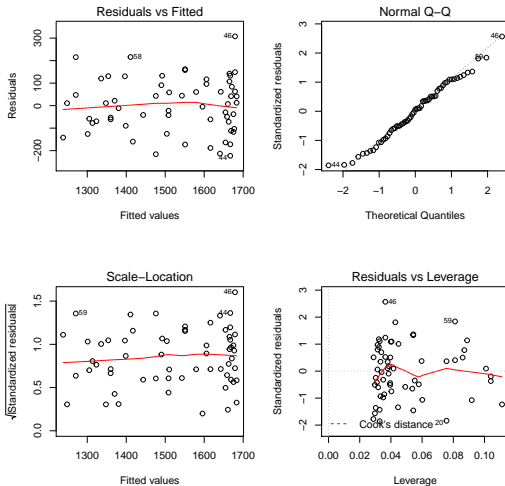
mod.hard.loc is the best!

```
mod3 <- lm(mortality ~ hardness + location, data = water)
par(mfrow=c(2,2))
plot(mod3)
```

Exercise 16

Exercise 17