Chapter 15

STAT 701 – Semhar Michael

Simultaneous Inference and Multiple Comparisons

```
library(HSAUR3)
```

```
## Loading required package: tools
```

```
library(coin)
```

```
## Loading required package: survival
```

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser

## The following object is masked from 'package:HSAUR3':
##
##     birds
```

```
library(sandwich)
library(lme4)
```

```
## Loading required package: Matrix
```

Simultaneous Inference and Multiple Comparisons

- Multiplicity is an intrinsic problem of any simultaneous inference
- If each of k, say, null hypotheses is tested at nominal level α on the same data set, the overall type I error rate can be substantially larger than α
  - *i.e* the probability of at least one erroneous rejection is larger than α for $k \geq 2$
- Simultaneous inference procedures adjust for multiplicity and thus ensure that the overall type I error remains below the pre-specified significance level α

Multiple comparison

- The term *multiple comparison* procedure refers to simultaneous inference
  - *i.e.* simultaneous tests or confidence intervals, where the main interest is in comparing characteristics of different groups represented by a nominal factor

alpha data

Various studies have linked alcohol dependence phenotypes to chromosome 4. One candidate gene is NACP (non-amyloid component of plaques), coding for alpha synuclein. Bonsch et al. (2005) found longer alleles of NACP-REP1 in alcohol-dependent patients and report that the allele lengths show some association with levels of expressed alpha synuclein mRNA in alcohol-dependent subjects.
Allele length is measured as a sum score built from additive dinucleotide repeat length and categorized into three groups: short (0-4, n = 24), intermediate (5-9, n = 58), and long (10-12, n = 15).

alpha data

Here, we are interested in comparing the distribution of the expression level of alpha synuclein mRNA in three groups of subjects defined by the allele length. A global F-test in an ANOVA model answers the question if there is any difference in the distribution of the expression levels among allele length groups but additional effort is needed to identify the nature of these differences. Multiple comparison procedures, *i.e.*, tests and confidence intervals for pairwise comparisons of allele length groups, may lead to additional insight into the dependence of expression levels and allele length.

alpha data

```
data(alpha)
head(alpha)
```

```
##         alength elevel
## 1         short   1.43
## 2 intermediate  -1.90
## 3 intermediate   1.55
## 4 intermediate   3.27
## 5 intermediate   0.30
## 6 intermediate   1.90
```

```
#?alpha
```

alpha data

```
summary(alpha)
```

```
##          alength        elevel
##  short       :24   Min.   :-2.830
##  intermediate:58   1st Qu.: 1.470
##  long        :15   Median : 2.770
##                    Mean   : 2.341
##                    3rd Qu.: 3.370
##                    Max.   : 5.800
```

```
tapply(alpha$elevel, alpha$alength, mean)
```

```
##        short intermediate         long
##     1.897917     2.332069     3.086667
```

```
tapply(alpha$elevel, alpha$alength, sd)
```

```
##        short intermediate         long
##    1.8548268    1.5808245    0.9738632
```

Multiple comparisons

- In chapter 5 we conducted multiple comparisons where multiple differences of **mean** rat weights were compared for all combinations of the mother rat's genotype.
  - Here, we used Tukeys honest significant difference (Tukey HSD) multiple comparison test
- Other multiple comparison procedures include
  - Dunnet: many-to-one comparison -*e.g.* treatments against control

Multiple Comparisons

- Here, we follow a slightly more general approach allowing for null hypotheses on arbitrary model parameters, not only mean differences.
- Each individual null hypothesis is specified through a linear combination of elemental model parameters and we allow for k of such null hypotheses to be tested simultaneously, regardless of the number of elemental model parameters p.
- More precisely, we assume that our model contains fixed but unknown p-dimensional elemental parameters θ.
- We are primarily interested in linear functions $\theta := \mathbf{K}\theta$ of the parameter vector θ as specified through the constant $k \times p$ matrix $\mathbf{K}$.

Simultaneous inference

- Consider a linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq} + \epsilon_i$$

- Suppose we are interested in inference about the parameters $\beta_1, \beta_q, \beta_2 - \beta_1$
- In chapter 6 can answer this separately but not all questions together.
- Formulate the three inference as a linear combination of the parameter vector $\theta = (\beta_0, \beta_1, \ldots, \beta_q)$
- If we let $q = 3$ we get

$$K = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{pmatrix}$$

- The global null hypothesis is

$$H_0 : \vartheta := K\theta^T = m$$

Simultaneous inference

- The global hypothesis H_0 is classically tested using an F-test
  - in linear and ANOVA models (see Chapter 5 and Chapter 6).
- Such a test procedure gives only the answer $\vartheta_j \neq m_j$ for at least one j
- but doesn't tell us which subset of our null hypotheses actually can be rejected.
- Here, we are mainly interested in which of the k partial hypotheses $H_0^j : \vartheta_j = m_j$ for $j = 1, \ldots, k$ are actually false.
- A simultaneous inference procedure gives us information about which of these k hypotheses can be rejected in light of the data.

Simultaneous inference

The estimated elemental parameters $\hat{\theta}$ are normally distributed in classical linear models and consequently, the estimated parameters of interest $\hat{\vartheta} = K\hat{\theta}$ share this property. It can be shown that the t-statistics

$$\left(\frac{\hat{\theta}_1 - m_1}{se(\hat{\theta}_1)}, \ldots, \frac{\hat{\theta}_k - m_k}{se(\hat{\theta}_k)} \right)$$

follow jointly multivariate k-dimentional t-distribution.

- The key aspect of simultaneous inference procedures is to take these joint distributions and thus the correlation among the estimated parameters of interest into account when constructing p-values and confidence intervals.

alpha data - *Genetic components of Alcoholisim*

Mean and standard deviation by levels of the factor

```
head(alpha, n = 3)
```

```
##         alength elevel
## 1         short   1.43
## 2 intermediate  -1.90
## 3 intermediate   1.55
```

```
tapply(alpha$elevel, alpha$alength, mean)
```

```
##         short intermediate          long
##      1.897917     2.332069     3.086667
```

```
tapply(alpha$elevel, alpha$alength, sd)
```

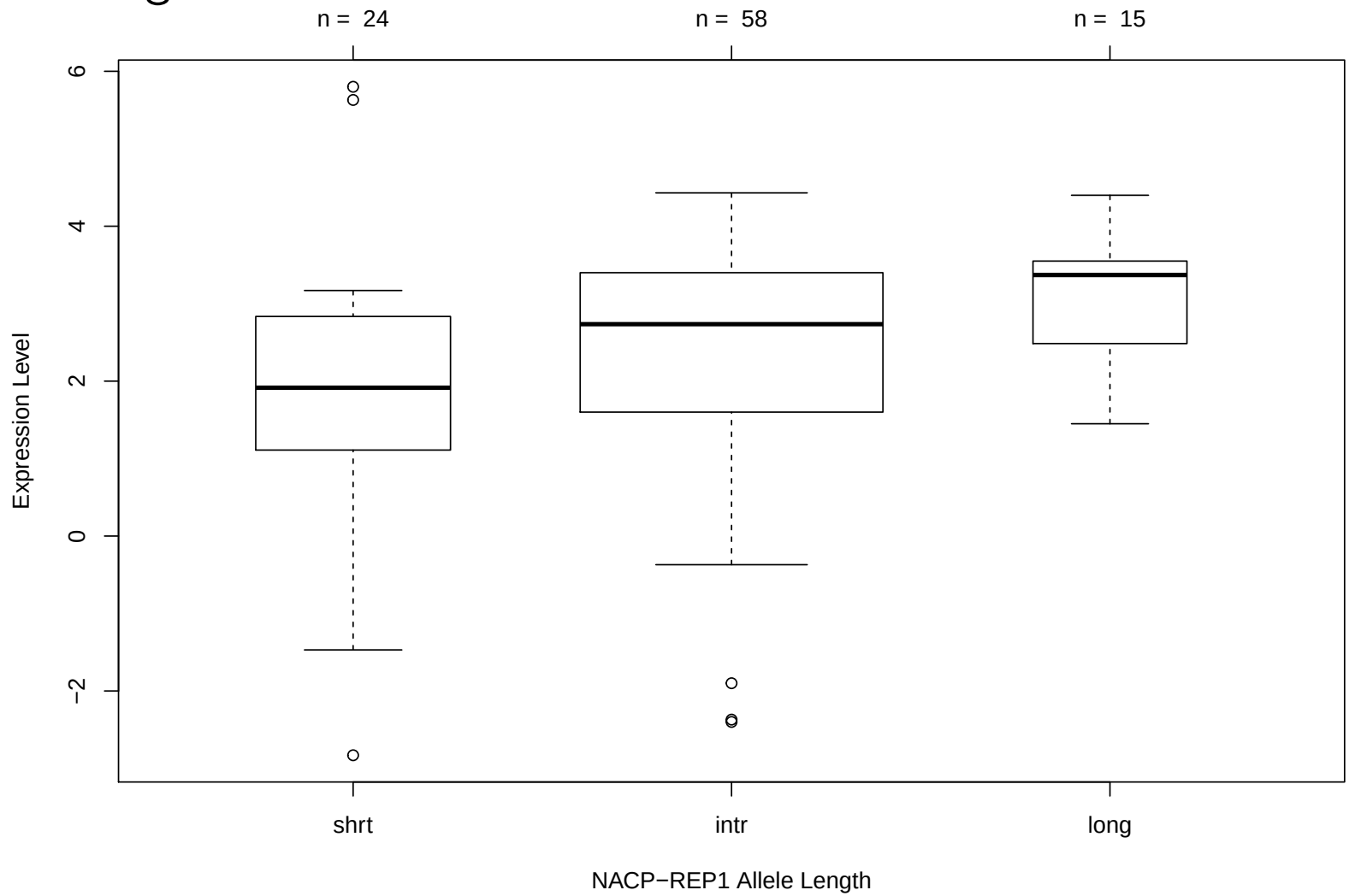
```
##         short intermediate          long
##     1.8548268    1.5808245    0.9738632
```

alpha data - *Genetic components of Alcoholisim*

```
n <- table(alpha$alength)
levels(alpha$alength) <- abbreviate(levels(alpha$alength),
plot(elevel ~ alength, data = alpha, varwidth = TRUE,
 ylab = "Expression Level",  xlab = "NACP-REP1 Allele Lengt
axis(3, at = 1:3, labels = paste("n = ", n))
```

alpha data - *Genetic components of Alcoholisim*

- Observe increasing expression levels of alpha synuclein mRNA for longer NACP-REP1 alleles.



alpha data - *Genetic components of Alcoholisim*

The model - simple one way ANOVA

$$y_{ij} = \mu + \gamma_i + \epsilon_{ij}$$

where

- $\epsilon_{ij} \sim N(0, \sigma^2)$
- $j \in \{short, intermediate, long\}$, and
- $i = 1, \ldots, n_j$.
- $\mu + \gamma_{short}$,$\mu + \gamma_{intermediate}$,$\mu + \gamma_{long}$ can be interpreted as the mean expression levels in the corresponding groups.
- This model is overparameterized (see chap 5)
- Solution: treatment contrast
  - $\theta = (\mu, \gamma_{intermediate} - \gamma_{short}, , \gamma_{long} - \gamma_{short})$ (default in R)
  - Equivalent to the restriction $\gamma_{short} = 0$

alpha data - *Genetic components of Alcoholisim*

The model - simple one way ANOVA - contrast

- For comparison among our three groups choose K as such

$$K_{Tukey} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}$$

- Then

$\vartheta_{Tukey} = K_{Tukey}\theta =$
$(\gamma_{intermediate} - \gamma_{short},, \gamma_{I}ong - \gamma_{s}hort, \gamma_{long} - \gamma_{intermediate})$

alpha data - *Genetic components of Alcoholisim*

- In R: *glht* (generalized linear hypothesis) function from the **multcomp** package
  - takes the *aov* (or lm/glm) object and
  - a description of the K-matrix

```
amod <- aov(elevel ~ alength, data = alpha)
amod_glht <- glht(amod, linfct = mcp(alength = "Tukey"))
```

alpha data - *Genetic components of Alcoholisim*

```
summary(amod)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## alength       2  13.06   6.528   2.613 0.0786 .
## Residuals    94 234.85   2.498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

alpha data - *Genetic components of Alcoholisim*

```
amod_glht$linfct
```

```
##                (Intercept) alengthintr alengthlong
## intr - shrt              0           1           0
## long - shrt              0           0           1
## long - intr              0          -1           1
## attr(,"type")
## [1] "Tukey"
```

alpha data - *Genetic components of Alcoholisim*

```
coef(amod_glht)
```

```
## intr - shrt long - shrt long - intr
##   0.4341523   1.1887500   0.7545977
```

```
vcov(amod_glht)
```

```
##                intr - shrt long - shrt long - intr
## intr - shrt   0.14717604   0.1041001 -0.04307591
## long - shrt   0.10410012   0.2706603  0.16656020
## long - intr -0.04307591   0.1665602  0.20963611
```

alpha data - *Genetic components of Alcoholisim*

```
#Simultaneous confidence intervals
confint(amod_glht)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = elevel ~ alength, data = alpha)
##
## Quantile = 2.3716
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                     Estimate lwr      upr
## intr - shrt == 0  0.43415 -0.47568  1.34398
## long - shrt == 0  1.18875 -0.04507  2.42257
## long - intr == 0  0.75460 -0.33126  1.84046
```

alpha data - *Genetic components of Alcoholisim*

```
#$p$-values
summary(amod_glht)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = elevel ~ alength, data = alpha)
##
## Linear Hypotheses:
##                     Estimate Std. Error t value Pr(>|t|)
## intr - shrt == 0   0.4342     0.3836   1.132   0.4924
## long - shrt == 0   1.1888     0.5203   2.285   0.0614 .
## long - intr == 0   0.7546     0.4579   1.648   0.2270
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
## (Adjusted p values reported -- single-step method)
```

alpha data - *Genetic components of Alcoholisim*

homogeneity of variances?

```
amod_glht_sw <- glht(amod, linfct = mcp(alength = "Tukey"),
                     vcov = sandwich)
summary(amod_glht_sw)
```

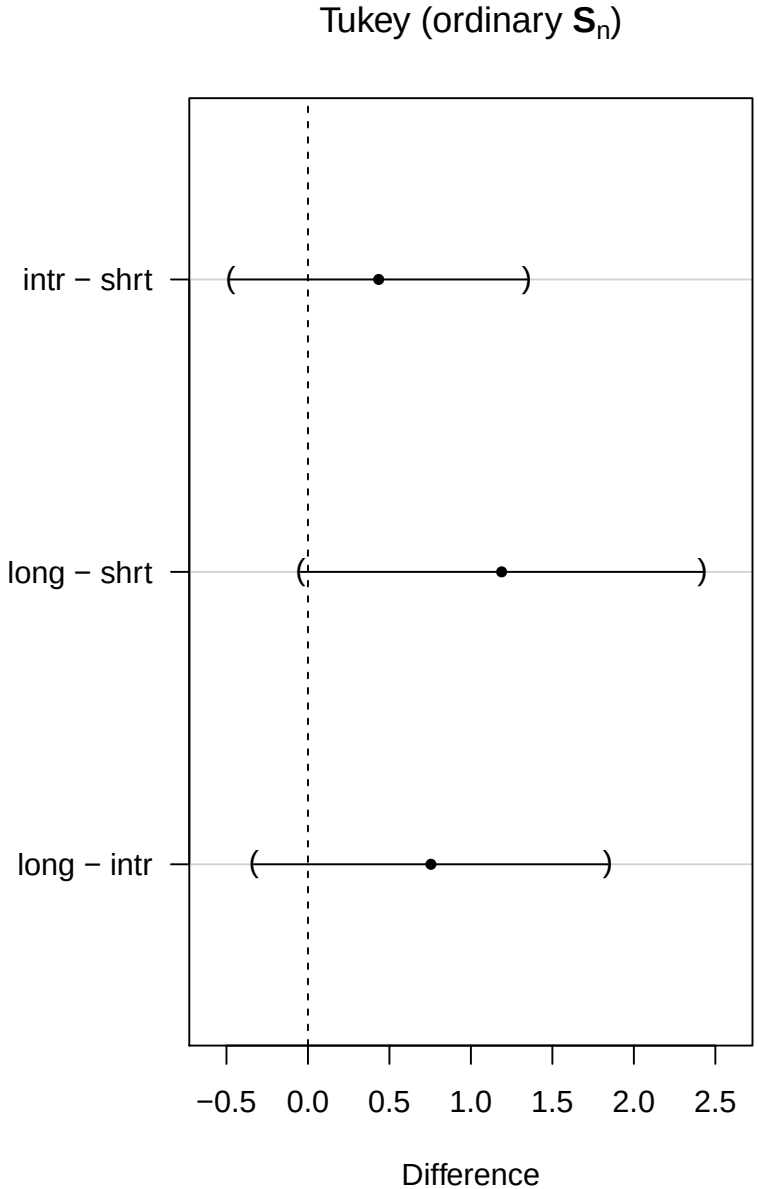
```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = elevel ~ alength, data = alpha)
##
## Linear Hypotheses:
##                     Estimate Std. Error t value Pr(>|t|)
## intr - shrt == 0   0.4342     0.4239   1.024   0.5594
## long - shrt == 0   1.1888     0.4432   2.682   0.0227 *
## long - intr == 0   0.7546     0.3184   2.370   0.0502 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

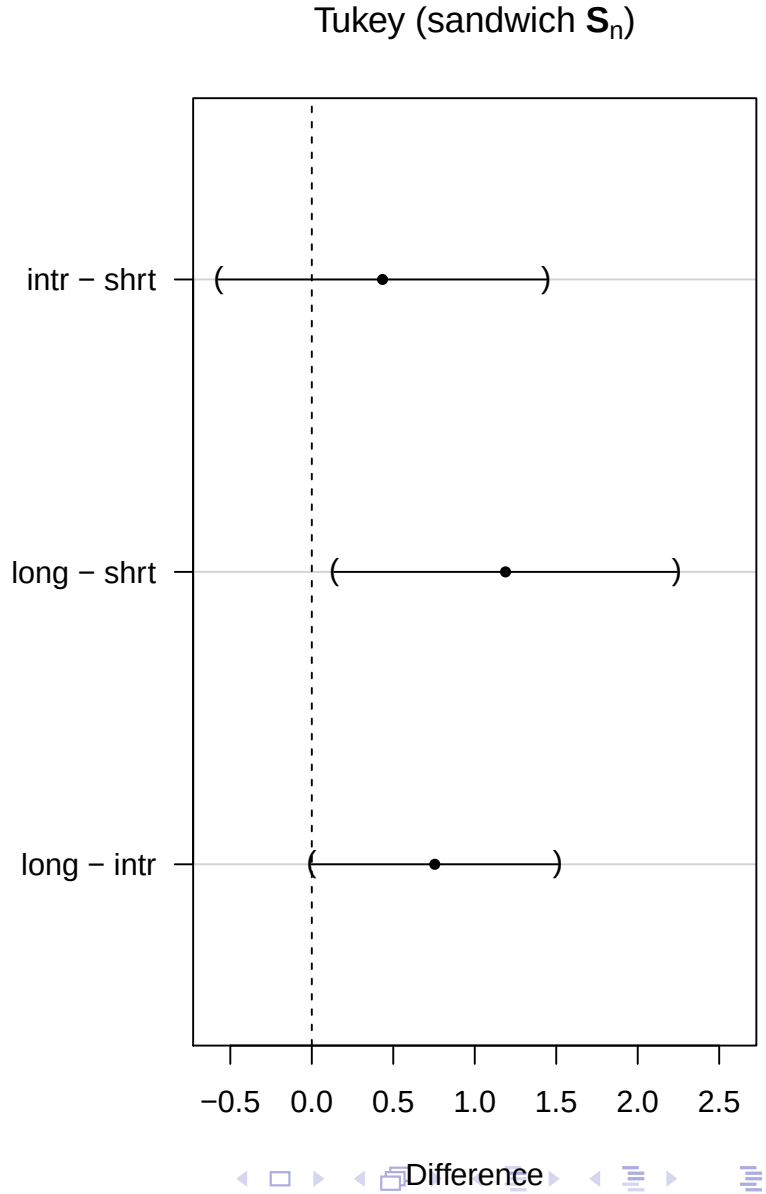
alpha data - *Genetic components of Alcoholisim*

```
par(mai = par("mai") * c(1, 2.1, 1, 0.5))
layout(matrix(1:2, ncol = 2))
ci1 <- confint(glht(amod,
                    linfct = mcp(alength ="Tukey")))
ci2 <- confint(glht(amod,
      linfct = mcp(alength = "Tukey"),
      vcov = sandwich))

ox <- expression(paste("Tukey (ordinary ",
                       bold(S)[n], ")"))
sx <- expression(paste("Tukey (sandwich ",
                       bold(S)[n], ")"))

plot(ci1, xlim = c(-0.6, 2.6), main = ox,
  xlab = "Difference", ylim = c(0.5, 3.5))
plot(ci2, xlim = c(-0.6, 2.6), main = sx,
  xlab = "Difference", ylim = c(0.5, 3.5))
```

alpha data - *Genetic components of Alcoholisim*

Simultaneous confidence intervals for the alpha data based on the ordinary covariance matrix (left) and a sandwich estimator (right).





alpha data - *Genetic components of Alcoholisim*

- We studied all pairwise differences in expression levels for three groups of subjects defined by allele length
- Overall there seem to be different expression levels for short and long alleles but no difference between these two groups and the intermediate group

trees513 data - Deer browsing data

- The survey takes place in all 756 game management districts ('Hegegemeinschaften') in Bavaria (data from 2006)
- The data of 2700 trees include the species and a binary variable indicating whether or not the tree suffered from damage caused by deer browsing
- For each of 36 points on a predefined lattice laid out over the observation area, 15 small trees are investigated on each of 5 plots located on a 100m transect line
- Thus, the observations aren't independent of each other and this spatial structure has to be taken into account for our analysis
- Our main target is to estimate the probability of suffering from roe deer browsing for all tree species simultaneously

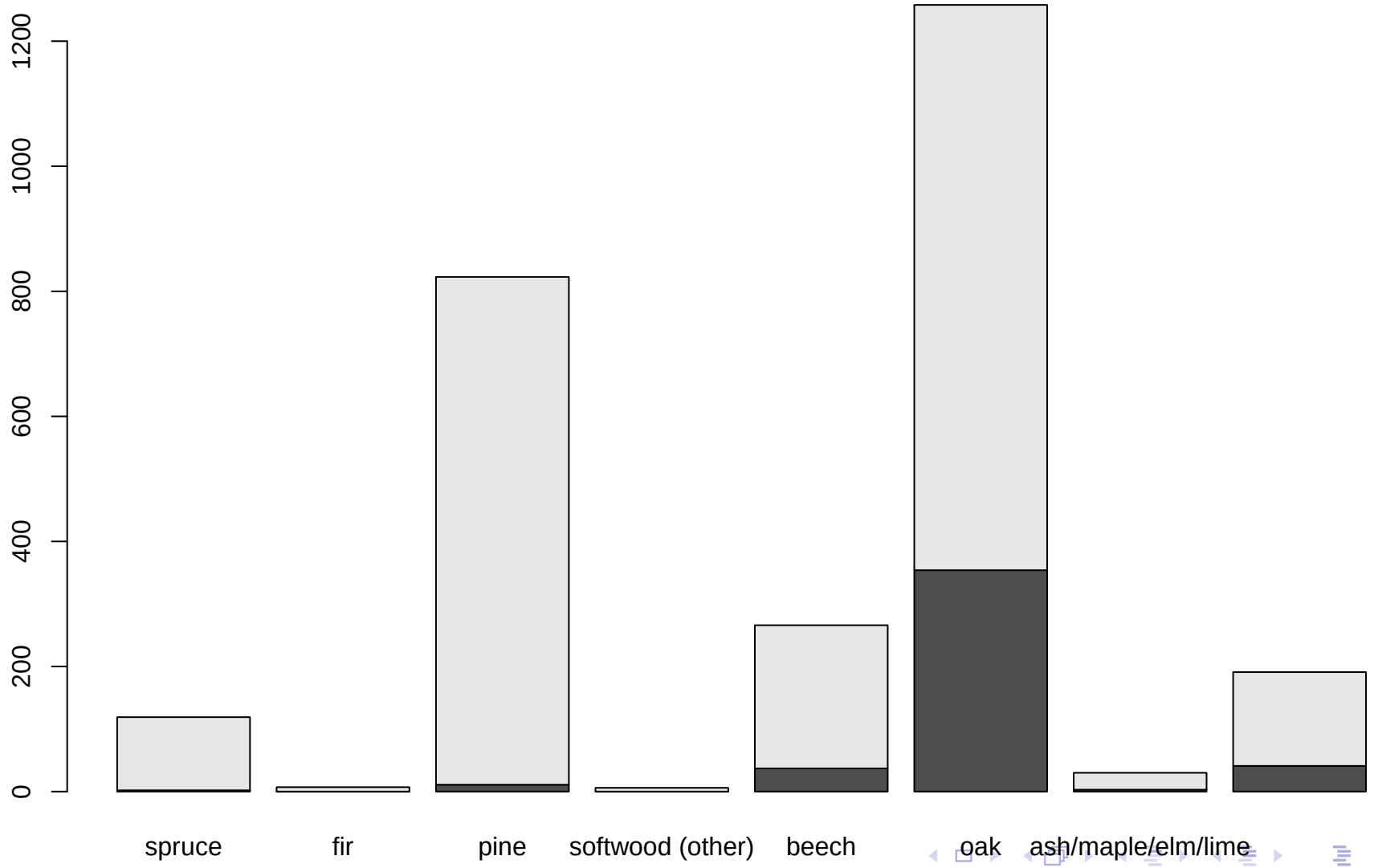
trees513 data - Deer browsing data

```
data("trees513")
head(trees513, n = 20)
```

```
##    damage           species lattice plot
## 1     yes               oak       1  1_1
## 2      no              pine       1  1_1
## 3      no               oak       1  1_1
## 4      no              pine       1  1_1
## 5      no              pine       1  1_1
## 6      no              pine       1  1_1
## 7     yes               oak       1  1_1
## 8      no hardwood (other)       1  1_1
## 9      no               oak       1  1_1
## 10     no hardwood (other)       1  1_1
## 11     no               oak       1  1_1
## 12     no              pine       1  1_1
## 13     no              pine       1  1_1
## 14    yes               oak       1  1_1
## 15     no               oak       1  1_1
## 16     no              pine       1  1_2
```

trees513 data - Deer browsing data

```
tn <- table(trees513$damage, trees513$species)
barplot(tn)
```



trees513 data - Deer browsing data

- Since we have to take the spatial structure of the deer browsing data into account, we cannot simply use a logistic regression model
- One possibility is to apply a mixed logistic regression model (using package **lme4**)

trees513 data - Deer browsing data

```
mmod <- glmer(damage ~ species - 1 + (1 | lattice / plot),
 data = trees513, family = binomial())
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl
## $checkConv, : unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl
## $checkConv, : Model failed to converge: degenerate Hess
## eigenvalues
```

Warning!

trees513 data - Deer browsing data

```
trees513.2 <- subset(trees513,
  species %in% c("spruce", "pine",
  "beech", "oak", "hardwood (other)"))
trees513.3 <- droplevels(trees513.2)
mmod <- glmer(damage ~ species - 1 +
                  (1 | lattice / plot),
 data = trees513.3, family = binomial())
K <- diag(length(fixef(mmod)))
K
```

```
##       [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    1
```

trees513 data - Deer browsing data

```
colnames(K) <- rownames(K) <-
 paste(gsub("species", "", names(fixef(mmod))),
 "(", table(trees513.3$species), ")", sep = "")
K
```

```
##                          spruce(119) pine(823) beech(266) o
## spruce(119)                        1          0          0
## pine(823)                          0          1          0
## beech(266)                         0          0          1
## oak(1258)                          0          0          0
## hardwood (other)(191)              0          0          0
##                          hardwood (other)(191)
## spruce(119)                                  0
## pine(823)                                    0
## beech(266)                                   0
## oak(1258)                                    0
## hardwood (other)(191)                        1
```

trees513 data - Deer browsing data

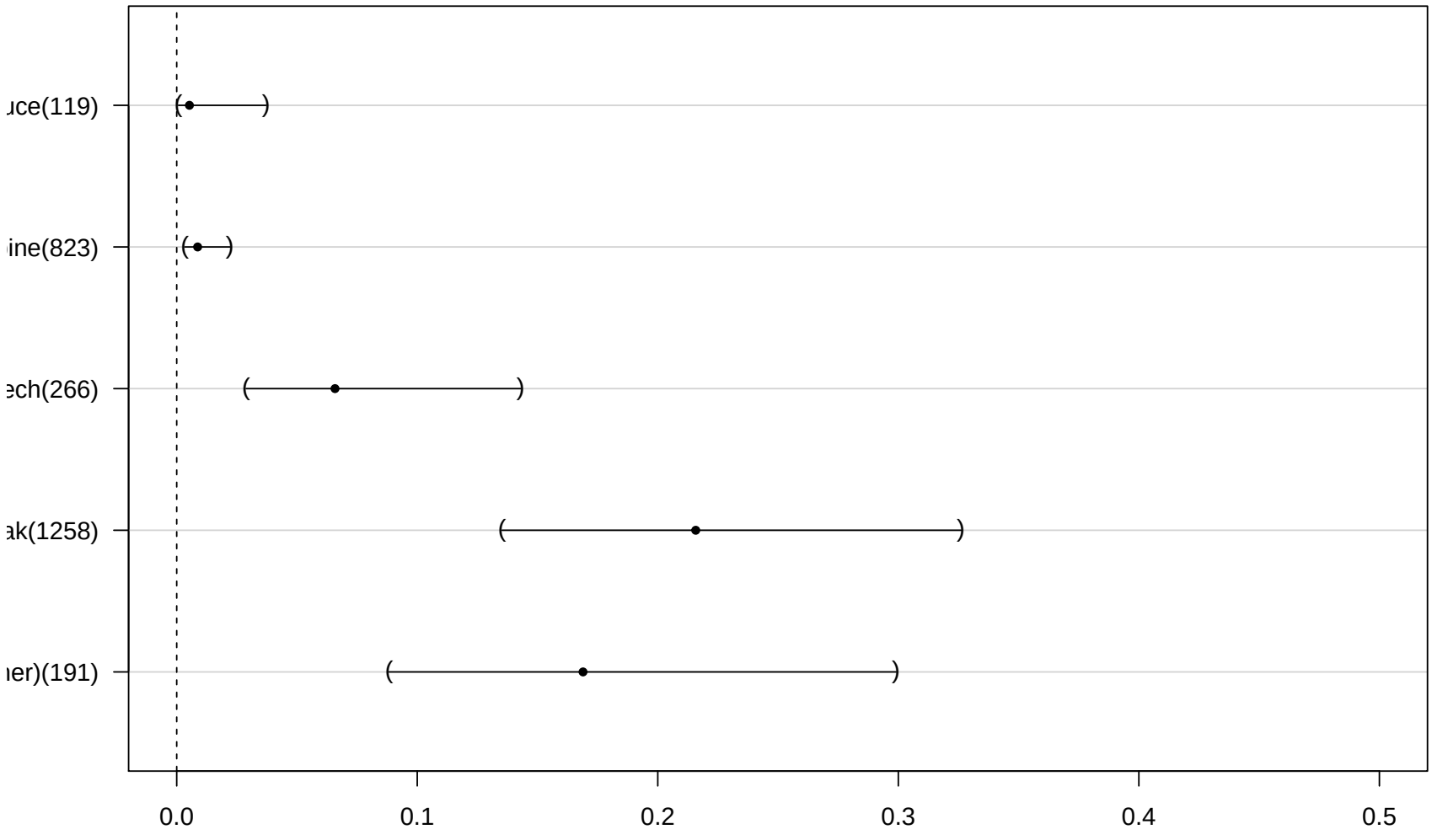
- Based on K, we first compute simultaneous confidence intervals for $K\theta$ and transform these into probabilities
- Note that $(1 + \exp(-\hat{\vartheta}))^{-1}$ is the vector of estimated probabilities; simultaneous confidence intervals can be transformed to the probability scale in the same way

trees513 data - Deer browsing data

```
ci <- confint(glht(mmod, linfct = K))
ci$confint <- 1 - binomial()$linkinv(ci$confint)
ci$confint[,2:3] <- ci$confint[,3:2]
```

trees513 data - Deer browsing data

```
plot(ci, xlab = "Probability of Damage Caused by Browsing",
     xlim = c(0, 0.5),
     main = "", ylim = c(0.5, 5.5))
```



trees513 data - Deer browsing data

- Browsing is more frequent in hardwood but especially small oak trees are severely at risk.
- Consequently, the local authorities increased the number of roe deers to be harvested in the following years.
- For a number of tree species, the simultaneous confidence intervals for the probability of browsing damage show that there is rather precise information about browsing damage for spruce and pine with more variability for the broad-leaf species. For oak, more than 0.14% of the trees are damaged.

Summary

Multiple comparisons in linear models have been in use for a long time. The multcomp package extends much of the theory to a broad class of parametric and semi-parametric statistical models, which allows for a unified treatment of multiple comparisons and other simultaneous inference procedures in generalised linear models, mixed models, models for censored data, robust models, etc. Honest decisions based on simultaneous inference procedures maintaining a pre-specified familywise error rate (at least asymptotically) can be derived from almost all classical and modern statistical models. The technical details and more examples can be found in Hothorn et al. (2008a) and the package vignettes of package multcomp (Hothorn et al., 2009a).