Short note on multiple hypothesis testing
TMA4267 Linear Statistical Models (V2017)

Kari K. Halle, Øyvind Bakke and Mette Langaas

March 15, 2017

March 15, 2017: Corrections made to formulas with R_j in Section 5, compared to version from March 13.
May 2, 2017: Corrections made to Figure 2.

This note gives a short introduction to central elements of the topic of *multiple hypothesis testing*. If you, after you have read this note, want to know more about multiple testing the article Goeman and Solari (2014) is an excellent read.

1 Single hypothesis testing

Consider a linear regression model, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the vector of regression coefficients, $\boldsymbol{\beta}$ has length p, the vector of univariate responses $\boldsymbol{Y}$ has length n and the design matrix has dimension $n \times p$ and rank p. Let β_j denote an element of $\boldsymbol{\beta}$, and assume that we are interesting in testing a two-sided hypothesis about no linear association between the response and the jth covariate

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0.$$

Two types of errors are possible, type I error and type II error. A type I error would be to reject H_0 when H_0 is true, that is concluding that there is a linear association between the response and the predictor where there is no such association. This is called a *false positive* finding.

A type II error would be to fail to reject H_0 when the alternative hypothesis H_1 is true, that is not detecting that there is a linear association between the response and the covariate. This is called a *false negative* finding.

The two types of errors for a single hypothesis test can be presented in Table 1.

	Not reject H_0	Reject H_0
H_0 true	Correct	Type I error
H_0 false	Type II error	Correct

Table 1: Single hypothesis testing set-up.

A *p-value* $p(X)$ is a test statistic satisfying $0 \leq p(\boldsymbol{Y}) \leq 1$ for every vector of observations $\boldsymbol{Y}$. Small values give evidence that H_1 is true. In single hypothesis testing, if the p-value is less

than the chosen significance level (chosen upper limit for the probability of committing a type I error), then we reject the null hypothesis, H_0. The chosen significance level is often referred to as α.

A p-value is *valid* if

$$P(p(\boldsymbol{Y}) \leq \alpha) \leq \alpha$$

for all α, $0 \leq \alpha \leq 1$, whenever H_0 is true (Casella and Berger, 2001, p. 397), that is, if the p-value is valid, rejection on the basis of the p-value ensures that the probability of type I error does not exceed α. An example of a p-value that is not valid is given in Appendix A.1.

If $P(p(\boldsymbol{Y}) \leq \alpha) = \alpha$ for all α, $0 \leq \alpha \leq 1$, the p-value is called an *exact* p-value. Observe that exact p-values are uniformly distributed when the null hypothesis is true, see Appendix A.2 for a short proof. This is a fact that is often misunderstood by users of p-values. The incorrect urban myth is that p-values from true null hypotheses are close to one, when the correct fact is that all values in intervals of the same length are equally probable (which is a property of the uniform distribution).

2 From single to multiple hypothesis testing

In many situations we are not interested in testing only one hypothesis, but instead m hypotheses.

- In a regression setting m might be the number of covariates in the regression model, and we would test $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ for all $j = 1, \ldots, m$.
- If we have a linear regression with one categorical covariate with k levels, called a one-way analysis of variance model, we might first want to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against the alternative hypothesis, H_1, that the means of at least two of the k levels are different from each other. If the null hypothesis is rejected we might want to continue to test which of all possible pairs of the means that are different – giving $m = \binom{k}{2}$ hypothesis tests, or compare the mean of all levels to a common reference level μ_1, giving $m = k-1$ hypothesis tests.

Let us assume that we perform m hypothesis tests, giving m p-values and then choose a cut-off on the p-values at some value α_{loc} (called a local significance level) to decide if we want to reject each null hypothesis. We then reject the null hypotheses where the p-value is smaller than α_{loc}, and this leads to rejection of R hypotheses. Benjamini and Hochberg (1995) summarized the (unknown) result of the m hypotheses tests as shown in Table 2. Note that in this table only the total number of hypotheses, m, and the number of rejected null hypotheses, R, are observed, and all other quantities are unknown (unless we for some reason know which null hypotheses are true or not).

In the table we assume that out of the m hypotheses tested, the (unknown) number of true null hypotheses is m_0. Further, V represents the number of type I errors (false positive findings) and T the number of type II errors (false negative findings). The last two quantities represent correct actions, with U true null hypotheses that are not rejected and S false null hypotheses that are rejected.

It is possible to define a generalization of the type I error rate to include the results of m tests, and in the next sections, we will present two different overall type I error rates which are used

	Not reject H_0	Reject H_0	Total
H_0 true	U	V	m_0
H_0 false	T	S	$m - m_0$
Total	$m - R$	R	m

Table 2: Multiple testing set-up

in multiple testing problems, the familywise error rate (FWER), which gives the probability of at least one false positive result among m hypotheses tested and the false discovery rate (FDR), which gives the expected proportion of false positive results among the m hypotheses tested.

But, first, why is this needed?

Consider a multiple testing problem where $m = 20$ covariates are tested for association with a response based on a method producing exact p-values. First we look at the expected number of false positive findings, $\mathrm{E}(V)$. If we use $\alpha_{\mathrm{loc}} = 0.05$ as the significance level cut-off for each individual test, the probability of falsely rejecting each null hypothesis is controlled at 5%, that is, given that the null hypothesis is true the probability that the p-value is smaller than $\alpha_{\mathrm{loc}} = 0.05$ is 0.05. So, for each of the m hypotheses the probability of falsely rejecting the null hypothesis is 0.05 (because we are working with exact p-values). If we assume that all null hypotheses are true, then the expected number of false positive findings is $m \cdot \alpha_{\mathrm{loc}} = 20 \cdot 0.05 = 1$. We expect to get 1 false positive finding when we use significance cut-off $\alpha_{\mathrm{loc}} = 0.05$ and perform $m = 20$ tests. It that good enough for us?

Further, we look at the probability of at least one false positive result among the m tests. Also here we assume that all null hypotheses are true, and now we also assume that the p-values from the 20 hypothesis tests are independent. Let the p-value for the ith test be denoted P_i. The probability of at least one false positive result among the $m = 20$ tests is

$$\begin{aligned} P(V > 0) &= 1 - P(V = 0) = 1 - P(P_1 > \alpha_{\mathrm{loc}} \cap P_2 > \alpha_{\mathrm{loc}} \cap \cdots \cap P_{20} > \alpha_{\mathrm{loc}}) = \\ &= 1 - P(P_1 > \alpha_{\mathrm{loc}}) \cdot P(P_2 > \alpha_{\mathrm{loc}}) \cdots P(P_{20} > \alpha_{\mathrm{loc}}) = 1 - (1 - 0.05)^{20} \\ &= 0.64, \end{aligned}$$

that is, for a multiple testing problem with $m = 20$ tests, the probability of obtaining at least one false positive result is 64%.

This example illustrates the importance of choosing the local significance cut-off α_{loc} lower than what we usually do for single hypothesis testing when performing many hypotheses tests.

3 Familywise error rate

The familywise error rate (FWER) is defined as *the probability of one or more false positive findings*

$$\mathrm{FWER} = P(V > 0)$$

where V is the number of false positive findings among the m hypotheses tested. The number of false positive findings is not known in a real life situation, but still we may find a cut-off on the p-value, called α_{loc}, that gives an upper limit to (controls) the FWER.

Raw and adjusted p-values Multiple testing methods can be described in terms of the raw or adjusted p-values. Raw p-values, p_j, are the lowest nominal level to reject the null hypothesis. The adjusted p-value, $\tilde{p}_j$, is the nominal level of the multiple (simultaneous) test procedure at which $H_{0j}, j = 1, \ldots, m$ is just rejected, given the values of all test statistics involved. The adjusted p-values can be defined as (Westfall and Young, 1993, p. 11)

$$\tilde{p}_j = \inf\{\alpha \mid H_{0j} \text{ is rejected at FWER level } \alpha\}.$$

In a multiple testing problem where all adjusted p-value below α are rejected, the overall type I error rate (for example FWER) will be controlled at level α.

Strong and weak control Methods for multiple testing correction can give strong or weak control of the overall type I error rate. Strong control means control of the overall type I error rate under any combination of true and false hypotheses. Weak control means control of the type I error rate only under the complete null hypothesis. Strong control of the overall type I error rate implies weak control of the type I error rate (Goeman and Solari, 2014). In many situations weak control implies strong control.

4 False discovery rate

The false discovery rate (FDR) is defined as

$$\text{FDR} = \begin{cases} \text{E}(\frac{V}{R}) & \text{for } R > 0 \\ 0 & \text{else} \end{cases}$$

(Benjamini and Hochberg, 1995) where V (the unknown number of false positive findings) and R (the number of rejected null hypotheses) are given in Table 2.

FWER vs FDR If all null hypotheses are true, the FWER and FDR are equal (Benjamini and Hochberg, 1995) and if only a subset of the null hypotheses are true, the FDR is less than the FWER for a given data set. The FWER is the probability of obtaining at least one false positive result, while the FDR is the expected proportion of false positives among the rejected null hypotheses.

5 Methods for control of the familywise error rate

We say that a method *controls* the FWER at some level α when FWER$\leq \alpha$. The wording *corrects for* multiple testing is also used to mean the same as *controls for*.

Single-step methods controls for multiple testing by estimating one local significance level, α_{loc}, which is used as a cut-off to detect significance for each individual test (Westfall and Young, 1993, p. 43). On the other hand, step-up and step-down methods use different cut-offs for a ranking of the tests based on the p-values. We will only consider single-step methods in this note.

We have m hypothesis tests and corresponding p-values. Let us define the event R_j,

$$\begin{aligned} R_j =& \text{the } j\text{th null hypothesis is rejected} \\ =& \text{ the } p\text{-value for the } j\text{th hypothesis test is below } \alpha_{\text{loc}}. \end{aligned}$$

Let $\bar{R}_j$ be the complementary event, i.e. that the jth null hypothesis is not rejected.

Now, we assume that all the m null-hypotheses are true. Then the familywise error rate can written as

$$\text{FWER} = P(R_1 \cup R_2 \cup \cdots \cup R_m) = 1 - P(\bar{R}_1 \cap \bar{R}_2 \cap \cdots \cap \bar{R}_m) \tag{1}$$

The local significance level, α_{loc}, which controls the FWER at level α is found by solving Equation (1) for α_{loc} using FWER $= \alpha$. But, where is α_{loc} in Equation (1)?

We assume that we perform m two-sided tests based on continuous test statistics $T_1, T_2, \ldots, T_m$ and let $f(t_1, t_2, \ldots, t_m)$ be the joint distribution of the m test-statistics. Then the FWER can be written as an m dimensional integral, and α_{loc} found by solving

$$\text{FWER} = 1 - \int_{-c}^{c} \int_{-c}^{c} \cdots \int_{-c}^{c} f(t_1, t_2, \ldots, t_m) dt_1 \cdots dt_m = \alpha \tag{2}$$

where $1 - \alpha_{\text{loc}} = P(-c \leq T_j \leq c)$ for all T_j. That is, to control the FWER in an optimal way we need to know the joint distribution of the test statistics and be able to solve an m-dimensional integral. This is in many cases not possible, and that is why methods that control the FWER without specifying the joint distribution is most often used. We will now look at two such methods.

5.1 The Bonferroni method

The Bonferroni method is valid for all types of dependence structures between the test statistics. Using Boole's inequality (the probability of a union of events is smaller than or equal to the sum of the probability of each of the events):

$$\alpha = \text{FWER} = P(R_1 \cup \cdots \cup R_m) \leq \sum_{j=1}^{m} P(R_j) = \sum_{j=1}^{m} \alpha_{\text{loc}} = m\alpha_{\text{loc}} \tag{3}$$

and the local significance level is $\alpha_{\text{loc}} = \frac{\alpha}{m}$ for the Bonferroni method. In Equation (3) the equality is if all events are disjoint, that is, perfectly negatively associated hypotheses.

The Bonferroni method gives strong control of the FWER (Goeman and Solari, 2014), but is known to be conservative when the tests are dependent. *Conservative* means that it is possible to get a higher value for α_{loc} that controls the FWER error rate by modelling the dependency structure between the tests.

5.2 The Šidák method

If the tests are assumed independent we find the Šidák correction

$$\text{FWER} = 1 - P(\bar{R}_1 \cap \cdots \cap \bar{R}_m) = 1 - \prod_{j=1}^{m} P(\bar{R}_j) = 1 - \prod_{j=1}^{m} (1 - \alpha_{\text{loc}})$$

which when we solve FWER=α gives the local significance level for the Šidák method

$$\alpha_{\text{loc}} = 1 - (1 - \alpha)^{1/m}. \tag{4}$$

The method of Šidák (1967) controls the FWER when an inequality called the Šidák inequality is satisfied, see Šidák (1967). It is beyond the scope of this note to elaborate on this, but as we have seen, independence of tests is sufficient for the Šidák method to control the FWER.

With the same explanation as for the Bonferroni method the Šidák method is also conservative. It can be shown that α_{loc} based on the Bonferroni method will always be smaller than α_{loc} based on the Šidák method, and as m increases the ratio between the α_{loc} from Šidák divided by the α_{loc} from Bonferroni goes to the limit $-\ln(1-\alpha)/\alpha$, which is 1.026 when $\alpha = 0.05$.

6 Examples

6.1 Illustration: the p-value is a random variable

We assume that in a normal population of females the systolic blood pressure is normally distributed with mean 120 mmHg and standard deviation 10 mmHg. We would like to study a population of females with a specific disease, and want to investigate if the mean systolic blood pressure in this population is larger than 120 mmHg. We assume that the standard deviation of the systolic blood pressure in the population with the specific disease is known to be 10 mmHg (just to make this simple). The null hypothesis we want to study is one-sided,

$$H_0 : \mu = 120 \text{ vs. } H_1 : \mu > 120$$

where μ is the unknown mean for the systolic blood pressure in the population of females with a specific disease.

We draw a random sample of size $n = 100$ from the population with the specific disease, and measure the systolic blood pressure, define random variables $X_1, X_2, \ldots, X_n$. As test statistic we use $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and when the null hypothesis is true then $\bar{X} \sim N(120, 1)$ (since $\text{SD}(\bar{X}) = \frac{10}{\sqrt{n}} = \frac{10}{10} = 1$). Let us assume that in our sample we observe $\bar{x} = 122$ mmHg. This will give a p-value of $P(\bar{X} > 122) = 1 - \Phi(122 - 120) = 1 - \Phi(2) = 0.023$.

How can we interpret this p-value? Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value. Should I conclude that $\mu > 120$? Yes, if we choose significance level higher than 0.02. But, we should also report a (two-sided) confidence interval for μ: Here $[120.04, 123.96]$.

But now, what if we have the possibility to repeat the experiment. Assume that we repeat the experiment three times, and that we get

1. $\bar{x}$=120.9, p-value=0.18.
2. $\bar{x}$ =118.9, p-value=0.86.
3. $\bar{x}$=121.2, p-value=0.12.

So, the p-value is also a random variable, and for each time we draw a new sample we get an observed version of the p-value. Let us assume that we do this first 100 times, and then 10 000 times. Histograms of p-values are presented in Figure 1.

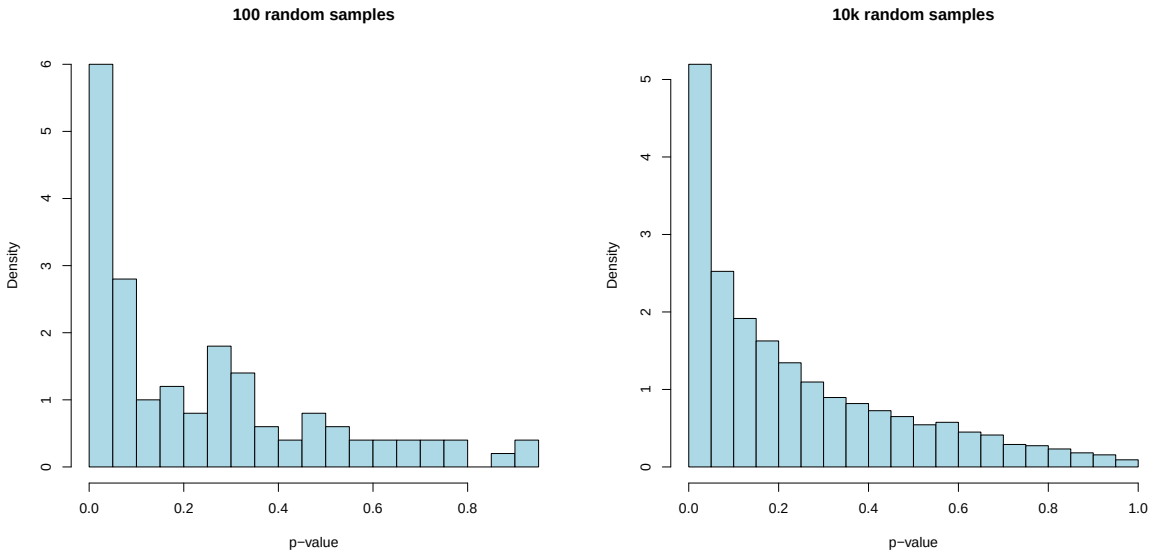

Figure 1: Histogram – and smoothed histogram – of p-values from the blood pressure example in Section 6.1.

6.2 Simulation: distribution of p-values from true null hypotheses

We continue with the blood pressure example. Now, we assume that the mean of the population of females with a specific disease is $\mu = 120$, that is, the null hypothesis is true. Again, we collect a random sample of size 100 and define the p-value to be $P(\bar{X} > \bar{x})$ for an observed value for the random sample of $\bar{x}$. What is then the distribution of the p-value? See the upper part of Figure 2 for R-code for drawing samples and plotting p-values, and the left panel of Figure 3 for the histogram of p-values when the experiment is repeated 10 000 times.

As we saw in Section 1, with proof in Appendix A.2, the p-values from the true null hypotheses is uniformly distributed when the p-value is exact, and in this example we have exact p-values.

6.3 Simulation: distribution of p-values from false null hypotheses

We continue with the blood pressure example, but now we assume that the null hypothesis is fals, and that the mean of the population of females with a specific disease is $\mu = 122$. Again, we collect a random sample of size 100 and define the p-value to be $P(\bar{X} > \bar{x})$ for an observed value for the random sample of $\bar{x}$. What is then the distribution of the p-value? See the lower part of Figure 2 for R-code for drawing samples and plotting p-values, and the right panel of Figure 3 for the histogram of p-values when the experiment is repeated 10 000 times.

There is no general result for the distribution of p-values from false null hypotheses. In our example we may change the true value from $\mu = 122$ to another value under the alternative hypothesis to see the effect on the distribution of the p-values.

```
sigma=10
n=100
sn=sigma/sqrt(n)
# H0 true
mu0=120
set.seed(123)
pval=NULL
for(i in 1:10000)
{xbar=mean(rnorm(n,mu0,sigma))
pval=c(pval,1-pnorm((xbar-mu0)/sn))
}
hist(pval,nclass=20,main="10000 random samples",
     xlab="p-value",prob=TRUE,col="lightblue")
# H1 true and mu=122
mu1=122
set.seed(123)
pval=NULL
for(i in 1:10000)
{xbar=mean(rnorm(n,mu1,sigma))
pval=c(pval,1-pnorm((xbar-mu0)/sn))
}
hist(pval,nclass=20,main="10000 random samples",
     xlab="p-value",prob=TRUE,col="lightblue")
```

Figure 2: The blood pressure example. R-code for simulating data when the null hypothesis ($\mu = 120$) is true, and when the null hypothesis is false and $\mu = 122$. The histograms are based on 10 000 drawings of samples of size 100.

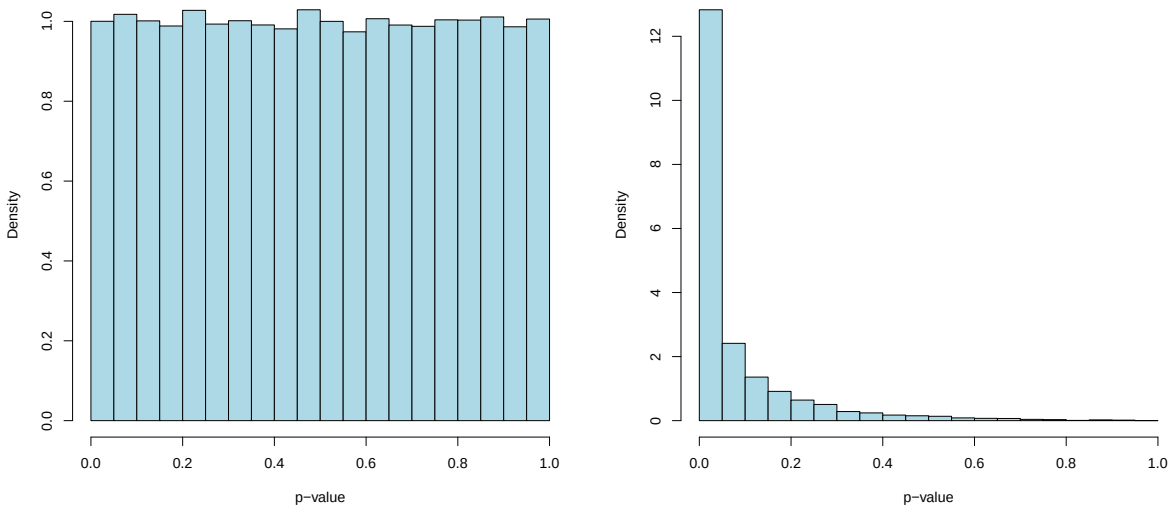

Figure 3: Histogram of p-values for the blood pressure example. Left panel is for $\mu = 120$ (null hypothesis is true in Section 6.2), and the right panel is for $\mu = 122$ (null hypothesis is false in Section 6.3).

6.4 Diabetes progression data

In a medical study the aim was to explain the ethology of diabetes progression, see Efron et al. (2004). Data was collected from $n = 442$ diabetes patients, and from each patient the following measurements are available:

- `age` (in years) at baseline
- `sex` (0=female and 1=male) at baseline
- body mass index (`bmi`) at baseline
- mean arterial blood pressure (`map`) at baseline
- six blood serum measurements: total cholesterol (`tc`), ldl cholesterol (`ldl`), hdl cholesterol (`hdl`), `tch`, `ltg`, glucose `glu`, all at baseline,
- a quantitative measurement of disease progression one year after baseline (`prog`)

All measurements except `sex` are continuous.

A multiple linear regression model is fitted to the data set with `prog` as response and all the other measurements as covariates, and is presented in Figure 4. Let $p = 11$ be the number of regression parameters in the model (intercept and all the covariates).

First, observe from Figure 4 that the regression is found to be significant, and that we then want to test the following hypothesis test for each of the regression parameters (we are not interested in testing the intercept):

$$H_1 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$$

for $j = 1, \ldots, 10$, so $m = 10$ hypothesis tests.

What should then be the cut-off used for calling each regression parameter significant if we want to control the FWER at level $\alpha = 0.1$?

First we look at the Bonferroni method, which gives $\alpha_{\text{loc}} = \alpha/m = 0.1/10 = 0.01$. Using this cut-off is always valid, and then we reject $R = 4$ hypotheses (sex, bmi, map, and ltg).

Then, is it possible to give a higher value of α_{loc} while still controlling the FWER at level 0.1? It is possible to try an asymptotic version of the integral presented in (1). In the linear regression we assume that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and that $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. This leads to $\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2)$. The test statistic for testing the hypotheses (as reported in Figure 4) is the t-statistic

$$T_j = \frac{\beta_j - 0}{\sqrt{c_{jj}\hat{\sigma}^2}}$$

where c_{jj} is the element (j, j) in $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\sigma}^2 = \frac{1}{n-p}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. The vector of test statistics will have a multivariate t-distribution, but asymptotically – and we have $n = 442$ – the vector of test statistics will follow a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix equal to the estimated correlation matrix of $\hat{\boldsymbol{\beta}}$. This integral can be calculated using the `mvtnorm` library with function `pmvnorm` and method Genz-Bretz. Then α_{loc} can be found by solving FWER=0.1 using a bi-sectioning algorithm. For our data set this gave $\alpha_{\text{loc}} = 0.0127$, which then takes into account dependency structure between the test statistics through the joint distribution of the test statistics.

```
>ds=read.csv("https://web.stanford.edu/~hastie/CASI_files/DATA/diabetes.csv",
sep=",")
>apply(ds,2,summary)
            age    sex   bmi    map    tc    ldl   hdl  tch   ltg    glu  prog
Min.    19.00 0.0000 18.00  62.00  97.0  41.60 22.00 2.00 1.410  58.00  25.0
1st Qu. 38.25 0.0000 23.20  84.00 164.2  96.05 40.25 3.00 1.860  83.25  87.0
Median  50.00 0.0000 25.70  93.00 186.0 113.00 48.00 4.00 2.005  91.00 140.5
Mean    48.52 0.4683 26.38  94.65 189.1 115.40 49.79 4.07 2.016  91.26 152.1
3rd Qu. 59.00 1.0000 29.28 105.00 209.8 134.50 57.75 5.00 2.170  98.00 211.5
Max.    79.00 1.0000 42.20 133.00 301.0 242.40 99.00 9.09 2.650 124.00 346.0
>full=lm(prog~.,data=ds)
>summary(full)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -356.64395   67.01983  -5.321 1.66e-07 ***
age           -0.03529    0.21705  -0.163 0.870910
sex          -22.79233    5.83657  -3.905 0.000109 ***
bmi            5.59548    0.71746   7.799 4.75e-14 ***
map            1.11589    0.22526   4.954 1.05e-06 ***
tc            -1.08286    0.57294  -1.890 0.059428 .
ldl            0.73914    0.53032   1.394 0.164108
hdl            0.36783    0.78274   0.470 0.638648
tch            6.54048    5.95956   1.097 0.273045
ltg          157.17606   36.04811   4.360 1.63e-05 ***
glu            0.28148    0.27332   1.030 0.303661
---
Residual standard error: 54.16 on 431 degrees of freedom
Multiple R-squared:  0.5176,        Adjusted R-squared:  0.5065
F-statistic: 46.25 on 10 and 431 DF,  p-value: < 2.2e-16
```

Figure 4: R code and print-out for fitting the diabetes progression data.

```
mm=model.matrix(full)
covB=solve(t(mm)%*%mm)
dd=diag(1/sqrt(diag(covB)))
corrB=dd%*%covB%*%dd

FWERGB=function(a,R)
{
  c=qnorm(1-a/2)
  m=dim(R)[1]
  return(1-pmvnorm(rep(-c,m),rep(c,m),sigma=R,
                   alg=GenzBretz(maxpts=1e6,abseps=1e-6))[[1]])
}
library(mvtnorm)
thislower=0.1/10
thisupper=0.1/5
alphalocm=uniroot(function(a)FWERGB(a,corrB)-.1,
                  c(thislower,thisupper),tol=1e-8)$root
alphalocm
[1] 0.01265296
```

Figure 5: The diabetes progression data. R-code for finding the local significance level α_{loc} based on integration of the asymptotic joint distribution of the test statistics.

Please observe, that with our p-values in Figure 4 all values of α_{loc} less than 0.059 and larger than 0.0001 gives the same result – rejection of the four hypotheses for sex, bmi, map and ltg.

A Appendix

A.1 Example of a p-value that is not valid

Assume that X_1 is binomially distributed with parameters $n_1 = 5$ and $p_1 = 0.5$, and that X_2 is binomially distributed with parameters $n_2 = 5$ and $p_2 = 0.5$. Also assume that X_1 and X_2 are independent. Let us test

$$H_0 : p_1 = p_2 \text{ vs. } H_1 : p_1 \neq p_2$$

using the test statistic

$$T(X_1, X_2) = \frac{\frac{X_1}{5} - \frac{X_2}{5}}{\sqrt{\frac{X_1+X_2}{10}(1 - \frac{X_1+X_2}{10})(\frac{1}{5} + \frac{1}{5})}}$$

(as we used in the first course in statistics). We assume that $T(X_1, X_2)$ is approximately standard normally distributed, and the p-value of the sample point (x_1, x_2) can be calculated as

$$p(x_1, x_2) = 2P(Z \geq |T(x_1, x_2)|)$$

where Z is standard normally distributed.

In our case the p-value $p(x_1, x_2)$ is less than 0.05 for 10 sample points:

$$(x_1, x_2) \in \mathcal{S} = \{(0, 3), (0, 4), (0, 5), (1, 5), (2, 5), (3, 0), (4, 0), (5, 0), (5, 1)\}$$

We get

$$P(p(X_1, X_2) \leq 0.05) = \sum_{(x_1,x_2)\in\mathcal{S}} \binom{5}{x_1} 0.5^{x_1} 0.5^{5-x_1} \binom{5}{x_2} 0.5^{x_2} 0.5^{5-x_2} = 0.0605$$

Thus, this p-value is not valid. The normal approximation should not be used to define p-values for such small samples.

A.2 Uniformity of exact p-values

Result:
Assume we have an hypothesis test situation and W is an exact p-value based on a continuous test-statistic. Then the p-values W from true null hypotheses are uniformly distributed.

Proof:
Assume that large values of the test statistic T leads to rejection of the null hypothesis, and that a value t of the test statistic T corresponds to a value w of the p-value W. This means that $P(T \geq t) = P(W \leq w)$. On the other hand the p-value is $P(W \leq w) = P(T \geq t) = w$ when H_0 is true.

This means that $P(W \leq w) = w$ when H_0 is true. If W is a continuous random variable taking values from 0 to 1, the the p-value W must be uniformly distributed over the interval from 0 to 1.

References

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Rociety. Series B (Methodological)*, 289–300.

Casella, G. and R. L. Berger (2001). *Statistical Inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*, 407–499.

Goeman, J. J. and A. Solari (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine 33*, 1946–1978.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*(318), 626–633.

Westfall, P. H. and S. S. Young (1993). *Resampling-Based Multiple Testing*. John Wiley and Sons, Inc.