Stats summary

Milena Trabert

September 2020 (Semester 1)

Contents

1 Lab 1

1.1 The p-value, type I and type II errors

Definition. *p-vlaue:* probability to observe your effect size (e.g. the measured difference between sample means), or an even larger effect size, given that the null-hypothesis is true.

General task: Test whether there is a difference between males and females (use normal distributed numbers to simulate data). In this case, the data is generated in a way that H_0 is true, there is no difference between males and females. The average score is 30 and the standard deviation is 10.

```
males <- rnorm(10, 30, 10)
females <- rnorm(10, 30, 10)

hist(males)
hist(females)
```

The Data fulfills the requirements for the t-test:

- The data is normally distributed
- each sample is taken at random from its respective population
- the variances are the same (for an independent sample test)

```
t.test(males, females, var.equal = TRUE)

'  Two Sample t-test

data:  males and females
t = 0.23661, df = 18, p-value = 0.8156
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.499698 10.657185
sample estimates:
mean of x mean of y
 32.19754 31.11880 '
```

The p-value shows that there is no significant difference between the samples ($0.8 \gg 0.05$).
The t value gives the effect size and the direction of the effect (positive or negative, in this case the mean of the males is greater since it was entered first into the function and the value is positive).

1.2 Multiple testing

Does the result change if we test a lot of groups of 10 men and 10 women? Repeat the experiment 1000 times (`ttestmf()`) and see how many significant results there are.

```
P_value <- 1:(1000)
ttestmf <- function(numpeople){
  for(i in 1:1000){
    males <- rnorm(numpeople, 30, 10)
    females <- rnorm(numpeople, 30, 10)
    P_value[i] <- t.test(males, females, var.equal = TRUE)\$p.value
  }
  sign.tests <-ifelse(P_value < 0.05, 1,0)
  return(sign.tests)
}

t_tasks <- 1:100
for (i in 1:100){
  t_tasks[i]<-sum(ttestmf(1000))
}
```

For a significance threshold of 0.05 there should be about 50 out of 1000 significant results in each run of `ttestmf()`, which is the case.

Some experiments in `t_tasks` show a significant difference in mean, but this might be due to the large number of experiments. For this reason we need to control for multiple testing using the function `p.adjust()` with the bonferroni correction. This correction multiplies all p-values with the number of tests made, so finding significant tests is a lot harder to find significant result. The false discovery rate `fdr` can also be used instead of bonferroni.

```
P_value <- 1:(100)
for(i in 1:100){
  males <- rnorm(numpeople, 30, 10)
  females <- rnorm(numpeople, 30, 10)
  P_value[i] <- t.test(males, females, var.equal = TRUE)$p.value
}

p.adjust(P_value, "bonferroni") #hist() can be used to visualize the results
```

Multiple testing will discard results that would be significant on their own (if there wouldn't have been multiple tests).

2 Lab2

2.1 One-way ANOVA

Analysis of variances: compare weight loss between three different diets

```
setwd('set/working/directory')
read.delim("diet.txt") -> diet
diet$weight.loss <- diet$pre.weight -diet$weight6weeks
boxplot(weight.loss ~ Diet, data = diet, col = 'light gray', ylab = 'Weight loss (kg)', xlab =
    'Diet type')
```

Boxplot to understand the spread of the data, no practical use or actual results.

There are two basic ways to perform ANOVA in R:

```
mod1a <- aov(weight.loss ~ Diet, data = diet)
mod1b <- lm(weight.loss ~ Diet, data = diet)
par(mfrow = c(2,2))
plot(mod1a)
plot(mod1b)
```

```
hist(resid(mod1a))
```

The dependent response variable weight loss is related to the independent explanatory variable Diet. (R has built in diagnostics when the response variable is continuous and normally distributed). The first plot is the QQ-plot, if the residuals follow the diagonal, they are normally distributed.

```
summary(mod1a) #aov()
'           Df Sum Sq Mean Sq F value Pr(>F)
Diet         2   60.5  30.264   5.383 0.0066 **
Residuals   73  410.4   5.622    '

summary(mod1b) #lm()
'Residuals:
    Min      1Q  Median      3Q     Max
-5.3680 -1.4420 0.1759 1.6519 5.7000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.3000     0.4840   6.818 2.26e-09 ***
DietB       -0.0320     0.6776  -0.047  0.96246
DietC        1.8481     0.6652   2.778  0.00694 **

Residual standard error: 2.371 on 73 degrees of freedom
Multiple R-squared: 0.1285, Adjusted R-squared: 0.1047
F-statistic: 5.383 on 2 and 73 DF, p-value: 0.006596'
```

In chapter 19 in *McKillup* the process of choosing a test is outlined.