Statistical Computing
Master Data Science
Winter Semester 2019/20

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 12 — Hypothesis tests 2

Exercise 1 Blood pressure for diabetic men

The systolic blood pressure for healthy men between 35 and 44 years has an expected value of 127.2 mm Hg and a standard deviation of 7.82 mm Hg. The data file `systolic.txt` (Moodle) contains the systolic blood pressure for diabetic men between 35 and 44 years.

Researchers are interested to see if the expected systolic blood pressure for diabetic men is greater that for healthy men in this age group.

(a) Import the data using `read.csv2`.

(b) Find the sample size, mean and standard deviation for these data.

(c) Obtain a box plot of the data and compare the plot with the expected value for healthy men.

(d) Write down the null and alternative hypothesis appropriate to the researchers' question.

(e) Use the function `t.test` to carry out this test. You will need to specify the null value `mu=?`. Depending on your alternative hypothesis you will also need to specify `alternative=` as one of `"two.sided"`, `"greater"` or `"less"`.

(f) From the output, determine the test statistic and the p-value.

(g) What is your conclusion?

Exercise 2

The small data set `Medication.csv` contains the time in minutes for blood to clot after patients had been given one of two types of blood coagulation drug type. Type A is the new drug (test group) and Type B is an existing drug (control group). The plan is to test whether the new drug gives a faster mean coagulation time μ_A than the control group μ_B.

(a) Formulate the two hypothesis for a one-sided two-sample t-test.

$$H_0 : \mu_A \; ? \; \mu_B \qquad \text{vs} \qquad \mu_A \; ? \; \mu_B$$

(b) Load in the data

```
> medication<-read.csv(file="Medication.csv")
```

(c) Answer the following questions:

(i) How many patients had each type of drug?

(ii) What is the mean of all coagulation times?

(iii) What is the standard deviation of all coagulation times?

(iv) What is the sample mean coagulation time for type A and for type B?

(v) Produce a box plot of `Time` depending on `Type`

(d) Looking at the box plot, what would you expect from the hypothesis test?

(e) Use the function `t.test` to carry out this test. Assume that the (population) variance is the same in the two types using the argument `var.equal=TRUE` (the default is to assume that te variances are different). You will also need to specify `alternative=` as one of `"two.sided"`, `"greater"` or `"less"`.

(f) What are your conclusions?

Exercise 3

Following up the last exercise, before reporting your conclusions to the researchers, it is good practice to check that the variance assumption is sensible.

We do this using another hypothesis test, an F**-test for the equality of two variances**.

$$H_0 : \sigma_A = \sigma_B \qquad \text{vs} \qquad H_1 : \sigma_A \neq \sigma_B$$

An alternative formulation for the null and alternative hypotheses is

$$H_0 : \frac{\sigma_A}{\sigma_B} = 1 \qquad \text{vs} \qquad H_1 : \frac{\sigma_A}{\sigma_B} \neq 1.$$

In other words the test is to see if the ration of the two population variances is equal to 1. This leads to the test statistic, $f_{\text{stat}} = s_A^2/s_B^2$, the ratio of the two sample variances.

The name for the test and test statistic comes from the so called f-distribution. The details of this distribution is not needed for this course, but the density function visually looks similar to that of the χ^2 distribution.

The R command for a variance test is:

```
var.test(x~group,data=dataframe)
```

Use this function to check that the variance assumption in the medication data set is acceptable.

Exercise 4 Chi-squared test of independence

This subject is well explained in the website:

`https://www.spss-tutorials.com/chi-square-independence-test/`

Read through this website before carrying out the test on these data using R.

In part (a) you will work through the calculations step by step to better understand the method. In part (b) you will carry out the test directly using an R command.

(a) Note that we do not need the original data, which consists of two variables, for 300 individuals. We can work directly on the cross table of joint frequencies given in the website, which is small enough to enter the joint frequencies by hand. Create the data using

```
> M<-matrix(c(???),nrow=4,ncol=5,byrow = T)
```

Type in the joint frequencies. The `byrow=TRUE` argument tells R that you will type in the data for each row rather than by each column. So that we can easily read the table, we can add row and column names.

```
> row.names(M)<-c("Never married","Married","Divorced","Widowed")
> colnames(M)<-c("<=Middle school","High school","Bachelor's",
+ "Master's","PhD+")
```

Check the data is correct and that the sample size is n=300.

```
> M
> sum(M)
```

(i) The easiest way to calculate the *expected frequencies* is using the outer product function `outer(x,y)`. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors `outer(x,y)` creates a matrix, with i,j-th element equal to $x_i y_j$. We require

$$e_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n}$$

The row frequencies $h_{i\cdot}$ are obtained using `apply(M,1,sum)`.

Complete the following command to get a matrix of the expected frequencies `E`, and check your answers with those in the website.

```
> E<-outer(apply(M,1,sum),???)/???
```

(ii) The rest is straightforward. Get the matrix of χ^2-elements $\dfrac{(h_{ij} - e_{ij})^2}{e_{ij}}$

```
> Chi<-(?-?)^2/?
> Chi
```

and add up all the elements in this matrix. This sum is the test statistic χ^2_{stat}.

(iii) We compare the test statistic with the $(1-\alpha)$-quantile from the χ^2_k distribution.

```
> qchisq(1-alpha,k)
```

You need to provide the appropriate value for `alpha` and `k`, the so called degrees of freedom.

(iv) Do you accept or reject the null hypothesis?

(b) Usually we don't need to go through all the steps above

(i) Everything is calculated using the function `chisq.test()`, and just printing the output gives us the test statistic, degrees of freedom and the p-value.

```
> output<-chisq.test(M)
> output
```

(ii) Type `qchisq(1-pvalue,k)` using appropriate numbers for `pvalue` and `k`. The value obtained should match your the test statistic χ^2_{stat} from part (a-ii). The definition of the p-value is the value of α which corresponds to the test statistic being on the boundary of the critical region.

(iii) The expected frequencies can be obtained using

```
> output$expected
```

The χ^2-elements are the square of the (relative) residuals

```
> output$residuals^2
```

NB: the function `chisq.test()` accepts the data in two forms: either as a matrix of joint frequencies, or as two vectors (such as two variables in a data frame).