Exercises for Day 2

Used datasets and R scripts can be downloaded in a ZIP archive from the Absalon *page (Applied Statistics) or from*

> https://datalab.science.ku.dk/english/course/smb-1/

Exercise 2.1: *Wilcoxon and T-tests: Syntax possibilities.*

In this exercise we will work with data example 2 used on course Day 1 in Exercise 1.2 and Exercise 1.4. Recall that this example was concerned with tenderness of pork meat from two pH-groups (high and low pH) after two chilling methods (tunnel and fast chilling). In the organization (tidy data on wide format) of the dataset in the text file `dataExample2.txt` there are four variables:

<p align="center">Pork, pH.group, Tunnel, Fast</p>

The objective of this exercise is to pay attention to the various syntaxes that may be used in the basic test functions like `t.test()` and `wilcox.test()`. First we will do this "theoretically", i.e. *without* opening R or RStudio. Instead you should simply imagine that the dataset is available in a data frame called `example2` containing the four variables listed above. Now suppose that you for the porks that have been tunnel chilled want to compare the tenderness in the low pH-group against the tenderness in the high pH-group. Which of the following 7 calls do this?, and why/why not?

```
t.test(example2$Tunnel,example2$Fast,paired=T)
with(example2,wilcox.test(Tunnel,Fast,paired=T))
with(example2,wilcox.test(Tunnel~pH.group))
t.test(Tunnel~pH.group,data=example2,var.equal=T)
t.test(Tunnel~pH.group,data=example2[-3,])
with(example2,t.test(Tunnel[pH.group=="low"],Tunnel[pH.group=="high"]))
wilcox.test(Tunnel~pH.group,data=example2,paired=T)
```

Two of the calls do another analysis. Which ones?, and what analysis do they do? Finally, there is one call making a nonsense test. Which one, and why does this test not make sense?

See next page for hints and remarks!

Hints and remarks:

- The idea of `with(my.data,my.expression)` is that the variables inside the data frame '`my.data`' are available when computing '`my.expression`'. E.g., in the second of the 7 calls the variables '`Tunnel`' and '`Fast`' inside the data frame '`example2`' are available to the Wilcoxon test.

  Thus, the function `with()` can be used just like a data-option, when the latter is not available.

- Both `t.test()` and `wilcox.test()` accept two different input syntaxes called the *"Default S3 method"* and *"S3 method for class 'formula'"* on the help pages (see `?t.test`).

- For two sample tests the *"Default S3 method"* requires the samples to be given in two vectors. If we want these vectors to be taken from a data frame one possibility is to use the `with()` function as described above.

- One advantage of the *"S3 method for class 'formula'"* is that the data frame may be specified in the data-option (see the 4'th, 5'th and 7'th call above..., by the way what happens in the 5'th call?).

- For more complicated functions like `lm()`, `glm()`, and `nlme::lme()` only the *formula method* makes sense.

- To get an improved feeling of the syntax you, of course, are welcome to open **RStudio** and try the code.

Exercise 2.2: *Two by two table: method consideration and understanding data organization.*

A study was made of 65 patients who had received or were receiving sodium aurothiomalate as a treatment for rheumatoid arthritis. The aim was to examine the possibility that toxicity to sodium aurothiomalate (SA) might be linked to impaired sulphoxidation capacity. The results were:

Impaired sulphoxidation	Major adverse reaction (toxicity)		Total
	Yes	No	
Yes	30	9	39
No	7	19	26
Total	37	28	65

The authors wrote: "The incidence of impaired sulphoxidation in patients showing SA toxicity (30/37, 81.0%) was significantly greater than in the group without adverse reaction (9/28, 32.1%) ($X^2 = 27.6, P < 0.001$). Similarly, the incidence of toxicity was significantly increased in those with impaired sulphoxidation (30/39, 76.9%) compared to those with extensive sulphoxidation (7/26, 26.9%) ($X^2 = 36.2, P < 0.001$)."

- Why can't both of the above chi squared tests be correct?

- Carry out a chi squared test of the data in the table and compare your answer with the two results in the above paragraph.

  Remark: This may be done using either `chisq.test()` or `prop.test()`, which will give the same result. If you don't want the Yates continuity correction then add the option `correct=F`.

- Compute a 95% confidence interval for the difference between the incidences of toxicity in the group with impaired sulphoxidation and the group with extensive sulphoxidation.

  Help: Here the `prop.test()` function is helpful.

- Additional questions related to structure of the dataset:

  - What are the variables in this study?
  - How many observations have been made?

In a call to `chisq.test()` the observations are provided in a matrix, i.e.

```
chisq.test(matrix(c(30,7,9,19),2,2))
```

However, in their original laboratory diary the people who made this study probably had the data organized in 65 rows somewhat like this:

Patient	Impaired sulphoxidation	Adverse reaction
1	yes	no
2	no	no
3	yes	yes
4	yes	no
⋮	⋮	⋮
64	yes	no
65	no	yes

Do you agree? Suppose data actually are given like this in a data frame called `arthitis`. Please try to decipher the following code:

```
chisq.test(table(arthritis))
```

(References: Altman, *Practical Statistics for Medical Research*, exercise 10.5, and Ayesh, R., Mitchell, S.C., Waring, R.H., et al. (1987): Sodium aurothiomalate toxicity and sulphoxidation capacity in rheumatoid arthritic patients. *Br. J. Rheumatol.*, **26**, 197–201.)

Exercise 2.3: *Analysis of a two-way table.*

The effect of dramanine as a remedy against seasickness was studied in an experiment on soldiers who crossed the Atlantic in a transportation of troops. Among 64 persons susceptible to seasickness a group of 34 soldiers were given dramanine while the remaining 30 soldiers received a placebo. The results were:

	Response		
Treatment	seasick	not seasick	Total
dramanine	3	31	34
placebo	12	18	30
Total	15	49	64

Analyse the data to see whether dramanine has an effect on seasickness. Try both `chisq.test()`, `prop.test()` and `chisq.test.simulate()` to do the analysis, and compare the outputs. What is the same, and what is different? (Remember to specify the `conditioning`-option in `chisq.test.simulate()` from the LabApplStat-package.)

A confidence interval for the probability of seasickness in the placebo group may be found using the R code:

```
prop.test(12,30)
```

Find a confidence interval for the probability of seasickness in the dramanine group. Is the effect of dramanine positive or negative?

Now suppose that the data is available in the text-file `dramanine.txt`, and not in the above table! Read the dataset into R using

```
read.table("dramanine.txt",header=T)
```

(alternatively use the *Import Dataset* menu), and analyse the data.

(The data are from Chinn, H.I et al. (1950): Prophylaxis of motion sickness: evaluation of some drugs in seasickness. U.S. Air Force School of Aviation Medicine Project 21-32-014, Rep. 4.)

Exercise 2.4: *A case-control study.*

In a study the relation between Hodgkin's disease and the presence of tonsillectomy was investigated. 85 Hodgkin's patients had a sibling of the same sex who was free of the disease and whose age was within 5 years of the patient's. The proportion of tonsillectomy in the Hodgkin's and the control group (i.e. the siblings) was presented by the investigators in the following table:

	Tonsillectomy	No tonsillectomy	Total
Hodgkins	41	44	85
Control	33	52	85

Apparently the following R call says that there is no relation between Hodgkin's disease and tonsillectomy (p=0.2789). However, this analysis is *wrong!* But why?

```
> prop.test(matrix(c(41,33,44,52),2,2))

2-sample test for equality of proportions with continuity correction

data:  matrix(c(41, 33, 44, 52), 2, 2)
X-squared = 1.1726, df = 1, p-value = 0.2789
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06603241  0.25426771
sample estimates:
   prop 1    prop 2
0.4823529 0.3882353
```

To perform the correct analysis we need more information than given in the above table, e.g. that in 37 of the patient-sibling pairs neither the patient nor the sibling had tonsillectomy. Use this additional information to perform the correct analysis.

Exercise 2.5: *Power calculations.*

As discussed in the lectures and in the two papers by Sterne & Smith (2001), and by Gelman & Carlin (2014) there is risk of both *Type I error* (false positives) and of *Type S* and *Type M error* (sign and magnitude) when the *power* (i.e., the probability of rejecting the null hypothesis given some

hypothesized effect size) of a study is too low. The classical rule of thumb is that the power should be a least 80%.

R provides some functions (both standard functions, and functions in the package PWR) for doing power calculations in simple situations. The purpose of this exercise is to try some of these functions:

- Exercise 2.3 is about the effect of the drug dramanine as a remedy against seasickness. Suppose that we believe that dramanine reduces the risk of seasickness by 75%, e.g. from 50% (with placebo) to 12.5% (with dramanine). Then the required sample size needed in order to have power=80%, say, may be found using the following R code:

  ```
  > power.prop.test(power=0.8,p1=0.50,p2=0.125)
  ```

  Try this! How many soldiers are needed to have sufficient statistical power in the experiment?

  Read more about the function from the help pages `?power.prob.test`, and find the required sample size under different scenarios (power=80%, 90%, and with different proportions of seasickness in the placebo group, e.g. 30%, 40% and 50%).

- Exercise 1.5 is about the effect of two drugs (E and N) on the treatment of high blood pressure. The study was done as a cross-over in order to reduce the biological variation between patients. We hypothesize that the difference between the effects of drug E and drug N is 8 mmHg, and that the biological variation within patients (as quantified by `sd(E_diff_N)`) has standard deviation given by 15 mmHg. What is the power with sample size $n = 19$?

  Hint: Use the function `power.t.test()` with options `delta=8`, `sd=15` and `type="one.sample"`. See `?power.t.test` for further details.

  Suppose that the investigators had expected the biological variation within patients to be smaller, e.g. `sd(E_diff_N) = 10`. Would this increase or decrease the power of the study?

Exercise 2.6: *Beauty and Sex ratios.*

In this exercise we redo the power calculation presented in the first data example in *Gelman & Carlin (2014)*. Please read the first column on page 5 in the paper for an introduction to the data example.

To perform the retrospective power analysis suggested by Gelman & Carlin two things are needed, namely *a hypothesized effect size* and *a standard error*[1]:

- Based on biological knowledge a hypothesized effect size, i.e. change of sex ratio, of 0.001, 0.003 or 0.01 is suggested.

- From the information from the paper *Kanazawa (2007)* Gelman and Carlin conclude that the standard error on the change is sex ratio is 0.033.[2]

The R function `retrodesign()`, which is available by running the R script `retrodesign.R` from the zip-file `day2.zip`, can be used to do the retrospective power analysis. This function takes two main arguments:

- A=hypothesized effect size. I have recoded the function given on page 9 in the paper such that it is possible to give a vector of possible effect sizes.

- SE=standard error.

Use this function to redo the analysis discussed on the second column of page 5 in the paper. Do you agree with the remarks made by Gelman and Carlin?

In the second last paragraph of the data example on beauty and sex ratio a traditional power calculation is made. So let's us also try this: What is the sample size needed to have power=80% when comparing proportion of girls between *attractive* and *unattractive* parents?

- Answer this question using `power.prop.test()` assuming that the proportion of girls with unattractive parents is 0.49, and with the 3 different hypothesized effect sizes given above.

Remark: The function `retrodesign()` is a slight generalization of the R code given in the appendix of Gelman & Carlin (2014). The generalization allows for several effect sizes to be considered in a single call.

---

[1]Not to be confused with the standard deviation given as input to the power calculations done for the hypertension example considered in Exercise 2.5.

[2]If you will do such a retrospective power analysis on your own data, then you often will be able to directly read off the standard error of the relevant parameter estimate from the R output.

Exercise 2.7: *Discretization of continuous variables.*

As you know the analyses of continuous data often assume underlying normal distributions. The purpose of this exercise is to convey the idea to discretize continuous variables into categories, which then may be analysed using the methods presented on this course day.

To illustrate this idea I have used one of the classical datasets, which contains joint observations of parents heights (a weighted and corrected average of fathers and mothers heights) and their sons heights (measured in inches). This dataset has historical relevance since it made Galton invent *regression analysis*[3]. The dataset is available in the text-file `Galton.txt`. Read the dataset into an R data frame called `Galton`, and make a scatter plot of data using the R code:

```
plot(child.ht~parent.ht,data=Galton,main="Galton's classical dataset")
```

Investigate whether the height measurements are normally distributed.

Hint: You may e.g. use the `qqnorm()` function.

Doing this you should conclude that the height measurements are nice normally distributed, so probably there is no need for a discretization here. But we will do it anyway to exemplify the idea of categorizing continuous data. Let us say that a man is *small* if he is less than 68 inches tall, and *tall* otherwise. The following R code makes the cross tabulation of small/tall vs. father/son:

```
with(Galton,table(parent.tall=(parent.ht>68),child.tall=(child.ht>68)))
```

Analyse the resulting 2-by-2 table to see if there is a relationship between the fathers and their sons heights.

End of exercises.

---

[3]I'll explain why it is called regression analysis on course Day 4.