university of groningen

faculty of behavioural and social sciences

Statistics II & Statistiek 2

Module code: PSBE2-07 & PSBA2-07

2019-2020

Material for Practicals

Contents

Purpose of the practicals

The aim of the practicals of Statistics 2 is:

1. To increase the insight in the theory of statistical methods.
2. To be able to apply statistical methods to empirical data.
3. To be able to decide what statistical method to use in order to answer a particular question.
4. To be able to work with the statistical software independently.

Learning to use R will be regarded as an extra goal – only for those students who are interested in this – and is explicitly not a course requirement. There will be practical groups for students who want to work withR. In these groups, practical assistants are able to help with R and JASPR questions. In all other groups, practical assistants are able to help with JASP-related questions. In case you have R-related questions you can use Nestor's discussion forum.
Please read the 'course information' to find out about the practical requirement.

As preparation for each week's practical, read the corresponding sections detailed in the reading material. Work on the computer exercises from the reader. There will be a computer provided for everyone. You are allowed and advised to work together with fellow students on the questions.

Answers to the questions will be published on Nestor *after* the practicals. Short worked-out solutions will be published. If you have follow-up questions regarding a practical exercise, you can ask your practical assistant or use the discussion board on Nestor.

All data sets referred to in the practical manual can be found under 'Course Documents' on Nestor.

At various places, a link to a scientific paper behind a certain data set is given. These links serve as a service to those who are interested in finding out more about the study. The referred papers are **not** part of the exam material.

On the next few pages, you will find a 'JASP Sheet' and an 'R Sheet' containing quick hints on how to use them.

R quick guide

The default option for the Statistics 2 practical is to work using JASP. Students aspiring a career in academia are encouraged to do the exercises using R. For questions and queries, please use Nestor's discussion board.
The aim of this quick guide is to provide a *quick* guide as to where to find the options in R. You can consult this guide to find help if you do not know how to solve practical exercises. For more elaborate information on using R, google and the library are your friends.

1 Beginning R-users

R can be downloaded from http://www.r-project.org/. R is open source, which means it is free, also after you graduate. The R-interface is fundamentally different from that of software like Excel and JASP and will take some time to get used to. To make things easier for yourself, you are encouraged to use an IDE like R-studio (http://www.rstudio.com/, also open source).
In order to understand the basics of R, please work your way through some R-tutorial[1].
When you need information on a command, type `?command`.

2 Installing R packages

R contains many useful built-in functions, for instance `mean()`. Sometimes though we want to use more specific functions. To get these functions we can install packages using `install.packages("packageName")`. After a package has been installed on your computer you have downloaded the functions contained in it to your computer. Now you can use these functions by loading the package with `load(packageName)`. Consequently, you have to install a package only once on your computer, but you will have to `load()` the package in every R session where you want to use one of its functions. You can see an example below where we install package `gdata` to use its function `read.xls()`.

3 Create a project

To import data and organize your files it is useful to create projects. Go to:

`File > New Project > New Directory > Empty Project`

Enter a name for the project, for instance "Rpracticals" then press `Browse` and coordinate to the folder where you want your project to be saved, press `Create project`. This creates a "mother" directory. From here on, all R script files are by default saved in this folder, and all datasets are by default read directly from this folder, so it is useful to download and unzip the Datafiles folder from Nestor in this folder.

4 Importing data files

R can only open plain text files (ASCII, such as CSV-files) directly. For all other filetypes, additional packages need to be installed.

[1] By googling, you will find a plethora of tutorials, such as http://cran.r-project.org/doc/manuals/r-release/R-intro.html, https://www.nceas.ucsb.edu/files/scicomp/Dloads/RProgramming/BestFirstRTutorial.pdf and http://www.cyclismo.org/tutorial/R/.

CSV files For CSV-files, use the command `read.csv`.

- `Data <- read.csv(file.choose(), header = TRUE)`

Excel files For Excel files, there are two main options:

- Open the file in Excel and save it there as CSV file.
- Install a package. There are many packages that can do this. Using packages requires three steps:
  1. Install the package on your machine: `install.packages("gdata")`. You only need to do this once.
  2. Load the package into R's memory: `library(gdata)`. You only need to do this once per R-session.
  3. Load the excel file: `mydata <- read.xls("filename.xls")`.

SPSS files For SPSS-files, proceed similarly as to for Excel files. Now, install and load the package `foreign` and load the data using `read.spss`.

- ```
  library(foreign)
  Data <- read.spss(file.choose(), to.data.frame=TRUE)
  ```

5 Indexing

- To retrieve, for instance, the 3rd element in a vector `Residuals` that contains several residuals:

  ```
  > Residuals
  1   2   4   -3    -1   -3
  ```

  use

  ```
  > Residuals[3]
  4
  ```

- If you want to retrieve elements from an object with more dimensions, for instance a data frame that contains rows and columns:

  ```
  > Data
                        Exam 1                          Exam 2

  Student 1        8                                      9

  Student 2        5                                      7

  Student 3        9.5                                   10
  ```

  use

  ```
  > Data[3, 2]
  10
  ```

  or

  ```
  > Data[3,  ]
  9.5   10
  ```

- To find out which element fulfills a certain requirement use `which()`:
  ```

```
> which(Residuals > 1)
2  3
```

or

```
> which(Residuals == 4)
3
```

6 Graphs

The default command for producing graphs is `plot`. Various specific types of plots have their own command, for instance `boxplot()` for boxplots and `hist()` for histograms.

Boxplots per group If `y` is the dependent variable and `group` is categorical, then `boxplot(y ~ group)` gives boxplots per group.

Linear regression plots When you have constructed your regression model (e.g., through `model <- lm(y ~ x)`) then

- `plot(x, y, type="p")` will give a scatterplot of the data;
- `abline(model)` will add the regression line;
- `plot(model)` will give four plots useful for inspecting residuals;
- `qqnorm(y)` will give a QQ-plot, `qqline(y)` will add the line to the QQ-plot;
- `plot(fitted(Model), Model$residuals)` plots the residuals versus the predicted values.

7 Missing Data

When your variables contain missing data, many functions may give an error message. Some functions allow you to perform actions for ignoring missing data, for example: `mean(x, na.rm = TRUE)`. Alternatively, to prevent these warnings it can often be useful to omit missing data from your variable: `x <- na.omit(x)`.

8 Linear Regression

- `model <- lm(y ~ x)` for simple linear regression.
- `model <- lm(y ~ x1 + x2)` etc. for multiple regression.
- see `names(model)` for finding out how to obtain the residuals etc.
- You can obtain useful summaries of the model with `summary(model)` and `Anova(model, type = ''III'')`. The function `Anova()` is included in the `car` package.
- `model$residuals` contains the residuals. You can use these to check assumptions. To produce plots to check regression assumptions, make sure to look at ' Linear regression plots' in section 6.

9 Confidence Intervals

After having specified a regression model, `model <- lm(y ~ x)`, `confint(model)` gives the confidence intervals.

10 Correlations

You can calculate correlations with the built-in function `cor()`. If your variables `x` or `y` contain missing values, you can specify that the missing data should be omitted during the calculation of the correlation: `cor(x, y, use = "pairwise.complete.obs")`.
The `cor()` function does not give p-values though. To get correlations and associated p-values use the `rcorr()` function form the `Hmisc` package. You will need to column-bind your variables for this: `rcorr(cbind(x, y), type="pearson")`.

11 Partial and Semi-Partial Correlations

To compute partial and semipartial correlations you can use the `ppcor` package. After installing and loading the package the `spcor()` and `pcor()` functions compute partial and semi-partial correlations.

- Missing data needs to be excluded using `na.omit()`

```
semiPartial = spcor(na.omit(Data))
```

- To get the semi-partial correlations request the first row of the resulting matrix using `[1, ]`

```
semiPartial$estimate[1, ]
```

12 Calculating means for different levels of one variable

To calculate the means of a variable `y` at different levels of the categorical variable `group` you can use the `tapply()` function on the dataframe, here called Data, that contains both variables.

- `tapply(Data$y, Data$group, mean)`

In a similar way other functions can be applied to every level of a variable.
To calculate the mean of all variables in a data set use `sapply()`.

- `sapply(Data, mean)`

13 t-test

To perform a t-test based on the pooled standard deviations use: `t.test(y ~ x, var.equal = TRUE)`.

14 Contrasts and Post Hoc tests

Pairwise comparisons can be performed using `pairwise.t.test`, where the adjustment method, such as Bonferroni, can be set through `p.adjust.method`.
Applying contrasts in R is not easy and requires some knowledge that won't be taught in this course (but in Statistics 3). A tutorial on contrasts in R can be found at
https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/.

15 ANOVA

When `y` is your dependent variable and `a` is your predictor, then

- `model <- lm(y ~ a)` will create
  1. A regression model if `a` is continuous;
  2. A one-way ANOVA model if `a` is categorical. To make sure `a` is recognized as categorical, you can use `as.factor(a)`.

Using `Anova(model, type = ''III'')` with the `Anova()` function from the `car` package you obtain the summaries of the model using Type III Sum of Squares. When you have two factors, `a` and `b`, then `lm(y ~ a + b)` provides a two-way ANOVA model. In a similar way you can get a multiple regression model.

16 Non-parametric ANOVA

Kruskal-Wallis test

- use `kruskal.test()`

JASP quick guide

The aim of this section guide is to provide a *quick* guide as to where to find some basic options in JASP. You can find more in-depth information on JASP at JASPs' web page https://jasp-stats.org/.
Every week you should save your analyses as a JASP file. To add or correct an analysis you do not have to rerun the analysis. You can go back to a specific analysis menu by clicking on the output that corresponds to the analysis. You can then add or remove options from the analysis.

1 Graphs

In general, graphs can be constructed using
`Descriptives > Descriptive Statistics > Plots`.
In order to be able to use appropriate graphs make sure that the Measurement Type (ordinal, nominal, scale) of your variables are specified correctly. You can change the Measurement Type of your variable in the data view by clicking on the Measurement Type symbol next to the variable name.

Boxplots per group

- `Descriptives > Descriptive Statistics > Plots`.
- Select `Display boxplots` and select the grouping variable in the `Split by` field.

Histograms

- `Descriptives > Descriptive Statistics > Plots`.
- Select `Display distribution plots` and select the desired variable.

Linear regression plots

- `Regression > Correlation Matrix > Plots > Correlation matrix`
  gives visual display of bivariate data. Note that you will have to select the desired variables.

Plots for checking assumptions

- `Regression > Linear Regression > Assumption Checks`

Several plots of the residuals can be made in this section that can be used to check the regression assumptions.

Mean plots

- `ANOVA > ANOVA > Additional Options`
- Select the desired outcome variable as `Dependent Variable` and the desired grouping variable(s) under `Fixed Factors`.
- Under `Descriptive Plots` select the desired grouping variable(s).

2 Filters

Sometimes it is useful to analyze those cases in a data set that fulfil a certain criterion (for instance, to do an analysis on only the control group). JASP allows you to apply filters to your data set to select special cases, for instance according to an equality or inequality constraint. You can find a helpful guide on how to use filters on the JASP website: https://jasp-stats.org/2018/06/27/how-to-filter-your-data-in-jasp/.

3 Linear Regression

Simple Linear Regression and Multiple Regression both use the same dialog boxes. For multiple regression, simply enter more variables in the list of `Covariates`.

- `Regression > Linear Regression` to perform linear regression.
- Specify your linear regression model by selecting variables as `Dependent` and `Covariates`.
- `Statistics > Regression Coefficient`
  ✓ `Confidence intervals`
  ✓ `Descriptives`
  gives basic information for a regression analysis including descriptive statistics and confidence intervals for the model parameters.
- `Statistics > Regression Coefficients`
  `Collinearity diagnostics` adds the VIFs to the **Coefficients** output table.
- `Statistics > Residuals > Casewise diagnostics` provides the Cook's distances. As an extra, summaries of the residuals also become available (all in the **Residuals Statistics** output table).

4 Partial and Semi-Partial Correlations

In JASP semi-partial correlations are called 'Part' correlations.

- `Regression > Linear Regression` to perform linear regression.
- Specify your linear regression model by selecting variables as `Dependent` and `Covariates`.
- `Statistics > Regression Coefficient`
  ✓ `Part and partial correlations`

5 Calculating means for different levels of one variable

- `ANOVA > ANOVA > Additional Options`
- Select the desired outcome variable as `Dependent Variable` and the desired grouping variable(s) under `Fixed Factors`.
- Under `Marginal means` select the desired grouping variable.

6 Contrasts and Post Hoc tests

Contrasts

- `ANOVA > ANOVA`
- Specify your ANOVA model: select the desired outcome variable as `Dependent Variable` and the desired grouping variable under `Fixed Factors`.

- Under `Contrasts` select the desired type of contrast by clicking on the grouping variable in the `Factors` window.
- for confidence intervals tick the bock next to `Confidence intervals` in the `contrasts` menu.

Post Hoc tests

- `ANOVA > ANOVA`
- Specify your ANOVA model: select the desired outcome variable as `Dependent Variable` and the desired grouping variable under `Fixed Factors`.
- go to `Post Hoc Tests`. Select your grouping variable and under `correction` select `Bonferroni`.

7 One-Way ANOVA

`ANOVA > ANOVA`
You can specify your ANOVA model by selecting the desired variable as `Dependent Variable` and the desired grouping variable under `Fixed Factors`.

8 Two-Way ANOVA

Summary statistics for two-way ANOVA designs can be obtained similarly to the one-way ANOVA, all you have to do is add a second grouping variable under `Fixed Factors`.

9 Non-parametric ANOVA

Kruskal-Wallis test Go to `ANOVA > ANOVA > Nonparametric` under `Kruskal-Wallis test` select or drag-and-drop your fixed factor.

10 Bayesian Binomial test

Go to: `Frequencies > Bayesian binomial test` and enter your test value θ under `Test value:`.

- For plots ✓ `Prior and Posterior`.

Week 2: Simple Linear Regression: Inference

The aims of today's activities are to:
(i) Use the Fisher's z transformation;
(ii) Obtain inferential statistics for regression;
(iii) Learn about the different kinds of questions that can be answered with regression analysis.

In a randomized trial, van Apeldoorn et al. (2010) studied the long-term effectiveness of different treatments for panic disorders. A subset of this dataset will be used in our analyses. The goal is to see to what extent psychological distress can be used to predict agoraphobia.
The data file is `van_apeldoorn.sav`. Amongst others, these variables are in the data file:

scl	Sum score of the SCL-90 (symptoms checklist) questionnaire, measuring psychological distress
agv	Sum score of a questionnaire measuring the severity of the agoraphobic problems

2.1 The Fisher z transformation

We start this practical with the manual computations of the Fisher z transformation. This topic is important because (i) it assists you in understanding the methodology behind the computations, and (ii) not all software provide proper support for the Fisher z transformation.

1. Load the data in the software of your choice and look up the sample size n and sample correlation $r = r(\text{scl}, \text{agv})$ using software. Be aware of the presence of missing values.

2. Compute the 95% confidence interval for ρ, the population correlation coefficient.

3. Perform an appropriate test for H_0: '$\rho = 0.3$' versus H_a: '$\rho > 0.3$'.

   (a) Provide the value of the test statistic.

   (b) Provide the p-value.

   (c) What do you conclude based on this test?

4. Compute, by hand, confidence intervals for ρ using levels of confidence other than 95% (already done in Exercise 2). Change the level of confidence as suggested in the table below.

Level of confidence	Confidence interval
90%	
95%	
99%	

5. Describe how the confidence intervals vary with the confidence level.

2.2 Predicting agoraphobia

6. Obtain a visual display showing the relationship between psychological distress score and the agoraphobia score.

7. (a) Describe the shape and direction of the relationship. Estimate its direction and strength (weak?, strong?).

   (b) What property have you just described?

8. (a) Using software, obtain a regression equation predicting agoraphobia from psychological distress. Make sure you obtain confidence intervals for the parameters of the regression equation.

   (b) Record the model parameters and confidence intervals for the model parameters here.

9. (a) Does it make sense to interpret the intercept of the equation? Explain why or why not.

   (b) Explain the slope of the regression equation in a way that makes sense given the units it is measured in.

   (c) Now consider the confidence interval for the slope. Explain the range of plausible slopes by considering the lower and upper bounds of the confidence interval.
   (Hint: How does agoraphobia change with psychoneurotism according to the lower bound? How does agoraphobia change with psychoneurotism according to the upper bound?)

   (d) Comment on the width of the confidence interval for the slope.

10. How good is the prediction of agoraphobia from psychoneurotism? First consider r, r^2, and s.

Week 3: Model Validity

The aims of today's activities are to:
(i) understand the difference between causality and association;
(ii) be able to spot outliers.

3.1 Causality and association

Death penalty

A study published in 1991 (described in (Agresti, 2002), Section 2.3.2) looked at the death penalties issued in the state of Florida after defendants were found guilty of murder. This resulted in the following table:

Victim's race	Defendant's race	Death penalty	No death penalty
Caucasian	Caucasian	53	414
Caucasian	African-American	11	37
African-American	Caucasian	0	16
African-American	African-American	4	139

1. Construct a 2×2 contingency table with defendant's race vs. whether the death penalty was issued or not. For both races, compute the percentage of defendant's that received the death penalty. What do you observe?

2. Now, take into account the race of the victim and compute the percentages death penalty for the four combinations of victim's and defendant's race. What do you observe?

3. How do you explain the discrepancy between the previous two answers.

4. Visit http://www.tylervigen.com/spurious-correlations. This is a website with a collection of variables that are very strongly related, yet there is no clear explanation why these should be correlated. Study three of the examples on this website and think up a mock explanation for this relation. Based on that, learn — and never forget — that correlation does not cause causation.

3.2 Outliers

The website fivethirtyeight reports that the number of hate crimes committed in a US state in 2019 was tied to the income inequality in that state. You can find the article here: https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/. We will redo their analysis, the datafile is `hate_crimes.csv`. We will use the variable `hate_crimes_per_100k_spcl` as an indicator for the number of hate crimes and the variable `gini_index` as a measure of income inequality.

5. Perform the regression analysis using software: Regress hate crimes on income inequality.

6. Plot the regression line including the observed values.

7. Determine the standardized residuals > 3 and record them here. Are there standardized residuals that correspond to unlikely large deviations from the regression line?

8. Which US state does this residual correspond to?

9. (a) Determine all residuals with a Cook's distance > 1 and record them here. You can use the built-in function `cooks.distance()`. Are there any influential points?
   (b) What assumption can be checked by means of the Cook's distance?
   (c) Why is the violation of this particular assumption problematic?
   (d) Do the results give reason to worry?

Week 4: Multiple Regression

The aims of today's activities are to:
(i) illustrate the use of multiple regression;
(ii) explore partial relationships.

4.1 Predicting mental health

Psychologists believe that mental health depends on many factors including recent life experiences and personal circumstances. Agresti and Finlay (1997) provide data from a study by Holzer (1977) of mental health in adults in Florida, USA. The data file is called `Holzer 1977.sav` and includes three variables.

Below is some descriptive information. Variable `mental` is a measure of psychiatric impairment including, for example, anxiety and depression; higher scores indicate higher impairment. Variable `life` is an index of the severity and number of major life events in the past three years; higher scores indicate more and/or greater severity of life events. (You may be familiar with this measure developed by Paykel.) Variable `ses` is an index of socio-economic status that is based on occupation, income, and education; higher scores mean higher status. Also, you can find histograms of the three measures, and also a scatterplot matrix showing the bivariate relationships. You might like to reproduce the scatterplot matrix using software of your choice so you can examine it more closely.

JASP

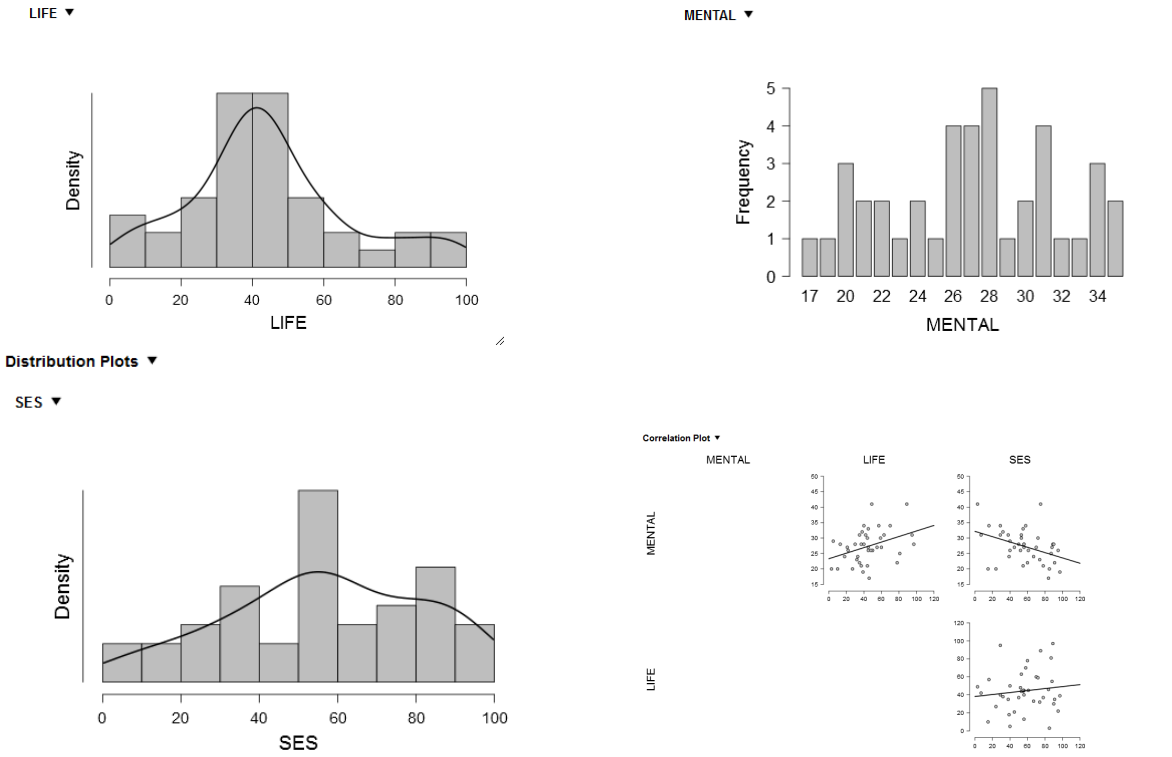
Go to:

`Descriptives > Descriptive Statistics > Plots`

Select `Display distribution plots`.
Go to:

`Regression > Correlation Matrix`

Select MENTAL, LIFE and SES as variables select `Correlation Matrix` under `Plots`.



1. (a) Comment on the univariate and bivariate displays. Describe the shape of the individual distributions and the patterns of relationships in the bivariate displays.
   (b) Are there any unusual features in the data that might be a problem for a regression analysis?

Below are the commands for the multiple regression analysis of these data. Two independent variables are entered – `life` and `ses`. The correlations among all variables and the 95% confidence intervals for the regression coefficients should be included in the output. Run the analysis and look at the output.

JASP

Go to:

`Regression > Correlation Matrix`

Select LIFE, MENTAL and SES.
Then, go to:

`Regression > Linear Regression`

Select MENTAL as `Dependent` and LIFE and SES as `Covariates`. Under `Statistics` select `Confidence Intervals`. Under `Residuals` check `Casewise diagnostics`.

2. (a) How strong are the bivariate relations among all three variables?
   (b) Given the pattern of bivariate relations, do you think both explanatory variables will be useful?
3. Describe the fit of the regression in terms of R^2 and s. Use plain language to explain what these two statistics indicate about the fit of the regression.
4. Examine the model table.
   (a) Why are there two degrees of freedom for the regression?
   (b) The source "residual" is usually given another name – what is it?

5. (a) The F statistic in the model table is the ratio of two variance estimates. What does a large F value mean?
   (b) The p-value in the model table is less than 0.001. Explain in plain language what this p-value means.
6. (a) Examine the table labelled "Coefficients" and write down the regression equation.
   (b) Are the model parameters in the direction that you might expect, given the correlations?
7. Consider the first person in the data file with the following scores:

$$\texttt{mental} = 17, \texttt{life} = 46, \texttt{ses} = 84.$$

   (a) What is the predicted value?
   (b) What is the residual for this person?
8. (a) Report the confidence intervals for the coefficients of the explanatory variables.
   (b) Describe how useful the regression is in terms of the width of these confidence intervals.
   (c) How much variation is there in plausible values for the parameter for the life events variable?
9. Write a short summary of the analysis, including a discussion of how useful you think the regression model is.

4.2 Parental rearing and borderline

The next exercises concern a data set on adolescents with borderline personality disorder symptoms (BPD). The (potential) relations between the severity of these BPD-symptoms and the psychopathology of their mothers is studied in two papers by Schuppert et al. (2015, 2014). The data are provided in `schuppert.sav`.
For the current practical, we consider only the following five variables:

stress	Parental rearing stress
group	Indicating whether a person belongs to the Borderline (coded 2) or the Control group (coded 1)
EWchild	Emotional warmth as recalled by the adolescent
REJchild	Rejection as recalled by the adolescent
OPchild	Overprotection as recalled by the adolescent

The first variable measures the stress as perceived by the mother during the rearing process. The last three variables, collected through the EMBU-C perceived parenting questionnaire, measure three aspects of the adolescent's upbringing, as recalled by the adolescent. These aspects are 'emotional warmth', 'rejection', and 'overprotection'.

10. Throughout this exercise, we will focus on the persons in the data set that are diagnosed with borderline symptoms. Select these cases in your software. To validate that you did this correctly, check that you have $n = 96$ persons remaining in the data with a mean maternal `stress` level of 87.2665. Make sure JASP recognizes `group` as a `nominal` variable by clicking the symbol next to the variable name.
11. Before you proceed with a regression model, it is important to get a feeling for the data. Analyze the descriptive statistics of the four variables and their bivariate correlations.
12. Perform a multiple linear regression which attempts to explain `stress` through the three EMBU-C variables. Include confidence intervals for your parameters.
    (a) Write down the fitted regression model.
    (b) Explain what each of the parameter estimates means in plain language.
13. Assess how well the model fit for the total model is. In your description, take into account the multiple correlation coefficient R^2 and the standard error of the estimate.

14. Assess the performance of the three independent variables.
15. For a person with scores Emotional Warmth = 50, Rejection = 30, and Overprotection = 25, what is the predicted stress level, according to the model?

Week 5: Multiple Regression: Interaction Effects

The aims of today's activities are to:
(i) understand what an interaction among continuous predictors is;
(ii) study the advantage of centering predictors;
(iii) computing and interpreting simple regression equations;
(iv) performing inferences using simple slopes (post hoc probing of interactions).

5.1 What is an interaction between continuous variables?

1. Try to explain with your own words what is being highlighted in the following quote:

   > "When two predictors in regression analysis interact with one another, the regression of y on one of those predictors *depends on* or is *conditional on* the value of the other predictor."
   >
   > (Cohen et al., 2003)

2. Consider the regression model $\widehat{y} = 2 + 0.2x_1 + 0.6x_2$. Interpret the regression coefficients of x_1 and x_2. Try to contextualize your interpretation with the geometric display of the regression surface as shown in the lecture.

3. Now consider the regression model $\widehat{y} = 2 + 0.2x_1 + 0.6x_2 + 0.4x_1x_2$. Again, interpret the regression coefficients of x_1 and x_2. Notice that the interpretation changes due to the presence of the interaction. Try to contextualize your interpretation with the geometric display of the regression surface as shown in the lecture.

5.2 Centering predictors

4. Consider the regression model $\widehat{y} = a + b_1x_1 + b_2x_2 + b_3x_1x_2$. Why is the interpretation of b_1 and b_2 often problematic in practice?

5. Consider the regression model using centered predictors $x_1^c = x_1 - \bar{x_1}$ and $x_2^c = x_2 - \bar{x_2}$: $\widehat{y} = \tilde{a} + c_1x_1^c + c_2x_2^c + c_3x_1^cx_2^c$. Interpret the first-order regression coefficients c_1 and c_2. Does the problem identified in the previous question still stand?

5.3 Example: Predicting physical endurance

Open dataset `C0701DT.sav`. X (age) and Z (number of years of vigorous physical exercise) will be used to predict Y (physical endurance, measured as the number of minutes of sustained jogging on a treadmill). Thus, $y = Y$, $x_1 = X$, $x_2 = Z$. There are $n = 245$ subjects in the sample.

6. Compute the mean, SD, minimum, and maximum of each variable. Fill in the table below using 6 decimals.

Variable		Mean	SD	Min	Max
X	$\bar{X} =$				
Z	$\bar{Z} =$				
Y					

7. Now that you know the means, compute the centered variables $x^c = X - \bar{X}$ and $z^c = Z - \bar{Z}$ (call them Xc and Zc, respectively), where $\bar{X}$ and $\bar{Z}$ are the means of X and Z using 6 decimals places that you saved in the previous exercise. Compute the interaction variable $x^c z^c = x^c * z^c$; call it $XcZc$. You can calculate new variables by pressing the + at the end of the data view window.

Simple regression equations, simple slopes

8. Estimate the regression model including both main effects and the interaction effect based on the centered predictors:

$$\begin{aligned} \widehat{Y} &= a + b_1 x^c + b_2 z^c + b_3 x^c z^c \\ &= \_\_\_\_\_ + \_\_\_\_\_ x^c + \_\_\_\_\_ z^c + \_\_\_\_\_ x^c z^c. \end{aligned} \quad (5.1)$$

9. Write the simple regression equation of Y on x^c at a given value z^{c*} (i.e., consider the number of years of vigorous physical exercise fixed and equal to z^{c*}). To do this, start by replacing all z^c's by z^{c*}'s in Equation 5.1. Next, just rewrite the terms of Equation 5.1 in a way that highlights that $\widehat{y}$ is to be regarded as a function of x^c:

$$\begin{aligned} \widehat{Y} &= (b_1 + b_3 z^{c*}) x^c + (b_2 z^{c*} + a) \\ &= (\_\_\_\_\_ + \_\_\_\_\_ z^{c*}) x^c + (\_\_\_\_\_ z^{c*} + \_\_\_\_\_). \end{aligned} \quad (5.2)$$

10. Similarly, write the simple regression equation of Y on z^c at a given value x^{c*} (i.e., consider the age fixed and equal to x^{c*}). Replace all x^c's by x^{c*}'s in Equation 5.1. Next, rewrite the terms of Equation 5.1 in a way that highlights that $\widehat{y}$ is to be regarded as a function of z^c:

$$\begin{aligned} \widehat{Y} &= (b_2 + b_3 x^{c*}) z^c + (b_1 x^{c*} + a) \\ &= (\_\_\_\_\_ + \_\_\_\_\_ x^{c*}) z^c + (\_\_\_\_\_ x^{c*} + \_\_\_\_\_). \end{aligned} \quad (5.3)$$

11. The simple slope in Exercise 9 is the coefficient of x^c in the simple regression equation. Observe that this simple slope is a function of the moderator z^c. Likewise, the simple slope in Exercise 10 is the coefficient of z^c in the simple regression equation, which is a function of the moderator x^c. Write down the expressions for both simple slopes here:

- Simple slope of x^c (as a function of z^c): $b_{x^c, z^c = z^{c*}} =$ _____.
- Simple slope of z^c (as a function of x^c): $b_{z^c, x^c = x^{c*}} =$ _____.

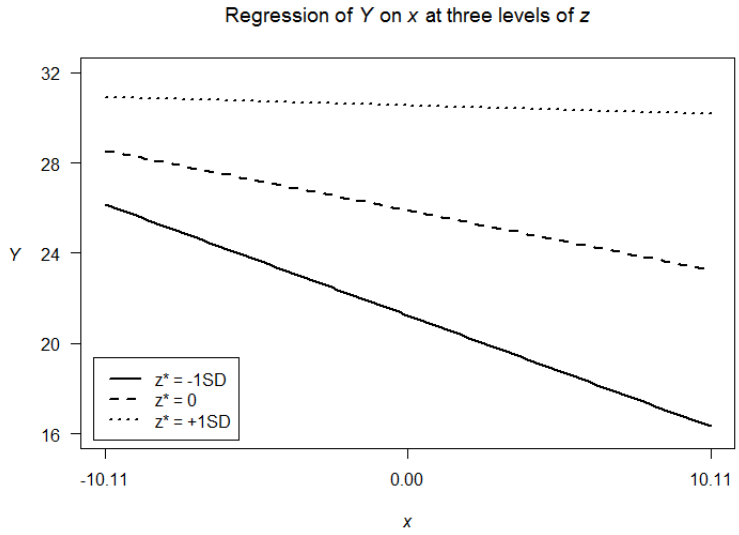
12. Using Equation (5.2), compute three simple regression equations of Y on x^c for three values of z^{c*}: $-1SD_Z = -4.78$, 0, and $+1SD_Z = 4.78$. Interpret each simple regression equation with your own words, taking into account what each variable is measuring.

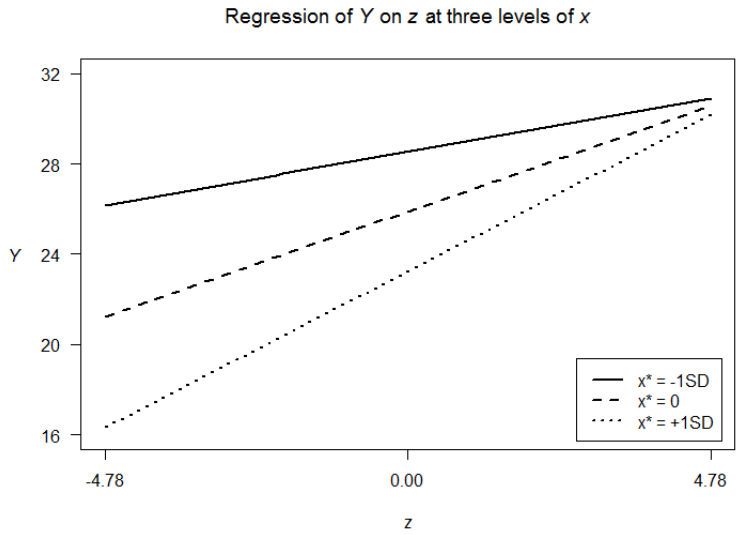
z^{c*}	Simple regression equation
-4.78	
0	
4.78	

13. Similarly, using Equation (5.3), compute three simple regression equations of Y on z^c for three values of x^{c*}: $-1SD_X = -10.11$, 0, and $+1SD_X = 10.11$. Interpret each simple regression equation with your own words, taking into account what each variable is measuring.

x^{c*}	Simple regression equation
-10.11	
0	
10.11	

14. It is possible to plot the simple regression lines estimated in Exercises 12 and 13 in order to visually inspect the interaction effect, as shown below.





Both plots are the equivalents to the 'means plots' and they can be interpreted in a similar way: Parallel lines indicate no interaction between X and Z, whereas non-parallel lines do indicate an interaction between X and Z. This inspection for interactions is not rigorous, it is only a visual indication (i.e., no inferences should be drawn from these plots).

How would you interpret each plot?

Post hoc probing of interactions

The statistical significance of simple slopes can be computed. This is what post hoc probing of interactions is about. Computations are relatively easy once the standard errors of the simple slopes are provided.

15. Let's compute a 95% confidence interval (CI) for a simple slope. We shall focus on the simple regression equation of Y on z^c for $x^{c*} = +1SD_X = 10.11$:

$$\widehat{Y} = \underbrace{1.448}_{\text{b at } x^{c*}} z^c + 23.24.$$

(a) The formula to compute the 95% CI is given by

$$1.448 \pm t^*_{0.975}(241) * SE_{\text{b at } x^{c*}} = 1.448 \pm 1.970\, SE_{\text{b at } x^{c*}},$$

where the t critical value was found using Student's t distribution.

(b) Use the SE provided here: $SE_{\text{b at } x^{c*}} = .205$.

(c) Compute the confidence interval.

(d) Interpret the confidence interval:

'The _____ (increase/decrease) of endurance (Y) with the number of years of vigorous physical exercise (z^c) _____ (is/is not) statistically significant for subjects with a high age ($x^{c*} = +1SD_X$). We are 95% confident that the population's simple slope of z^c is between _____ and _____.'

16. You can also perform a t-test to test the null hypothesis 'the simple slope of Y on z^c at $x^{c*} = 10.11$ is equal to zero' against the alternative hypothesis 'the simple slope of Y on z^c at $x^{c*} = 10.11$ is different from zero'.

(a) Compute the value of the test statistic.

(b) What is the result of this test?

5.4 Extra exercises

The following exercises are to be solved during the practical only if time allows.
If not, you are expected to work on these exercises at home.

Enter the following data in your software package. We will use these data to help answering questions 17 through 18.

X	Z	Y	$x^c = X - \bar{X}$	$z^c = Z - \bar{Z}$	XZ	$x^c z^c$
1	6	12				
3	8	14				
5	3	10				
7	8	11				
9	5	16				

17. (a) Compute x^c and z^c, the centered X and Z variables, respectively. Add the values to the table above.

(b) Estimate the regression coefficients for models $\widehat{Y} = a + b_1 X + b_2 Z$ and $\widehat{Y} = \tilde{a} + c_1 x^c + c_2 z^c$. How do the corresponding coefficients b_i and c_i, for $i = 0, 1, 2$, relate to each other? (Are they equal? Are they different?)

18. (a) Compute XZ and $x^c z^c$, the interaction terms using the noncentered and centered predictors, respectively. Add the values to the table above.

(b) Estimate the regression coefficients for model I $\widehat{Y} = a + b_1 X + b_2 Z + b_3 XZ$ and model II $\widehat{Y} = \tilde{a} + c_1 x^c + c_2 z^c + c_3 x^c z^c$. How do the corresponding coefficients b_i and c_i relate to each other? (Are they equal? Are they different?)

(c) You must have observed in the previous exercise that $b_1 \neq c_1$ and $b_2 \neq c_2$. However, this does **not** mean that the effects of X and Z on Y differ from the effects of x^c and z^c on Y, respectively. The only thing that occurred while centering the predictors was a translation of the regression surface on the (X, Z) plane. This can be seen with an example:

i. For instance, take $X = 6$ and $Z = 4$.
ii. Knowing that $\bar{X} = 5$ and $\bar{Z} = 6$, compute the corresponding x^c and z^c values.
iii. Compute $\widehat{Y}$ using model I with $X = 6$ and $Z = 4$.
iv. Compute $\widehat{Y}$ using model II with x^c and z^c that you found in exercise 18(c)ii.
v. Compare both $\widehat{Y}$ values. What do you conclude?

19. Observe the following:

> "In general, *centering predictors has no effect on the value of the regression coefficient for the highest order term* in the regression equation."
>
> (Cohen et al., 2003)

This general rule explains why first-order regression coefficients are unaltered by centering in regression models with no interactions (Exercise 17b). This rule also explains why the regression coefficient for the higher-order interactions are unaltered by centering in regression models with interactions (Exercise 18b).

Week 6: Multiple Regression: Partial Correlation. Standardized Regression

The aims of today's activities are to:
(i) be able to compute and interpret (semi)partial correlations;
(ii) Compute various variants of the multiple correlations coefficient R^2.

6.1 Predicting mental health

In section 4.1 a multiple regression analysis was performed on the mental health data by Holzer (1977). We will revisit this exercise. The data file is called `Holzer 1977.sav` and includes three variables, with `mental` being the dependent variable.

1. Use software to compute the pairwise correlations between `mental`, `life` and `ses`.
2. Manually compute the semi-partial correlation of `life` and `ses`.
3. What is the interpretation of the semi-partial correlation of `life` and `ses`?
4. Manually compute the partial correlation of `life` and `ses`.
5. What is the interpretation of the partial correlation of `life` and `ses`?
6. The regression equation you used last week is $\texttt{mental}=\alpha+\beta_1\texttt{life}+\beta_2\texttt{ses}$. *Using software* compute the estimates for β_1 and β_2.
7. Using the result of the last exercise, compute the regression coefficients of the standardised regression model $\texttt{mental}=\beta_1^* z_L + \beta_2^* z_S$, where z_L and z_S are the standardised versions of `life` and `ses`.
8. Now, use software to compute the values you computed manually in the previous exercises. Check whether you calculations are correct and indicate where in the output all the values can be found.
9. Compute R^2 using software. Next, compute the Wherry adjustment manually. Interpret these values.

6.2 Predicting hate crime

Now we will use software to calculate the partial and semi-partial correlations when more than two predictors are involved. We will revisit the hate crime data from week 3. The data file is `hate_crime.csv`.

10. Use `avg_hatecrimes_per_100k_fbi` as dependent variable and `gini_index`, `share_voters_voted_trump` and `share_population_with_high_school_degree` as predictors. Compute the partial and semi-partial correlations with your software.

11. What is the interpretation of the partial correlation of gini-index?
12. What is the interpretation of the semi-partial correlation of gini-index?

This week's practical is - as you might have noticed - fairly short. Use the left-over time during the practical to catch up with exercises from previous weeks you haven't finished yet.

Week 7: Multiple Regression: Assumptions

The aims of today's activities are to:
(i) study the assumptions of the regression model;
(ii) practice testing the assumptions of the regression model.

7.1 Predicting mental health

Recall the practical of weeks 4 and 6. We now continue the analysis of the mental health data by Holzer (1977), which was described in Agresti and Finlay (1997). The data file is called `Holzer 1977.sav` and includes three variables.

1. After performing the multiple regression (explaining `mental` by `life` and `ses`), create a QQ-plot of the residuals.

   (a) Why is this graph useful?

   (b) What does it indicate for this analysis?

   (c) What other residual plots might be useful to check normality?

2. Now build two scatterplots:

   - Residuals (y-axis) against `life` (x-axis).
   - Residuals (y-axis) against `ses` (x-axis).

   Add the horizontal line $y = 0$ to each scatterplot (to be used as a reference line). Use these plots for inspections of the linearity and the homoscedasticity assumptions.

   (a) What can you conclude concerning the linearity assumption?

   (b) What can you conclude concerning the homoscedasticity assumption?

7.2 What candy to buy for Halloween?

The website `fivethirtyeight` wanted to know what the most popular candy would be to pass out on Halloween. The data file is `candy-data.csv`. These are the variables in the data file:

winpercent	a measure of the popularity of a candy
pricepercent	indicates how expensive the candy is
sugarpercent	a measure of the sugar content of the candy

3. Perform a linear regression analysis predicting `winpercent` using `pricepercent` and `sugarpercent` as predictors. Write down the regression equation.

4. Make a histogram and QQ-plot of these residuals.

   (a) What assumption do we make about the shape of the distribution of the residuals?
   (b) Do the residuals above meet this assumption?

5. Draw scatterplots: Residuals (y-axis) against each of the predictors (x-axis). Using the scatterplots, what can be said about:
   (a) Linearity?
   (b) Homoscedasticity?
6. Draw scatterplots: Residuals (y-axis) against predicted values (x-axis). Using the scatterplots, what can be said about:
   (a) Linearity?
   (b) Homoscedasticity?
7. Now plot the residuals against the dependent variable `winpercent`, what do you conclude based on this plot?
8. Add an interaction between `pricepercent` $\times$ `sugarpercent`, how does the plot of residuals against the dependent variable `winpercent` change? What do you conclude based on this plot?
9. What do you conclude about candy popularity based on this final model, consider the significance of the parameters and the proportion of explained variance.

Week 8: Regression with Code Variables

The aims of today's activities are to:
(i) examine the relationship between t-test based procedures and simple linear regression;
(ii) practice interpreting the parameter estimates when using regression with code variables.

8.1 Two group comparisons

Larkin et al. (1992) looked at the effects of using heart-rate feedback training on performance in stressful tasks. Their study was quite complex, but we shall focus here on one task – a video game – and on the immediate effects of training. Sixteen male students were randomly assigned to either a treatment group or a control group. The treatment group received video feedback about their heart-rate during training on the video game. We will compare heart-rates of the control and treatment groups after training; all participants' heart-rates were measured for a 6 minute baseline and then for two minutes while playing the video game. Here is a visual summary:

	Training	Heart-rate measurement	
Control	no feedback	6 minute baseline	2 minutes with game
Treatment	heart-rate feedback	6 minute baseline	2 minutes with game

For each participant, the average heart-rate during the baseline ("vbpost" in the data file) and the average heart-rate during the game ("vpost" in the data file) was calculated. The variable we will examine is the difference between the baseline average and the game average. This variable has been computed;the name of the new variable "vdifpost" (thus, vdifpost = vbpost − vpost). The data file is `vdifpost.txt`.

Standard analysis

1. Using a boxplot, examine the distribution of "vdifpost" for the treatment and control groups. The groups are defined by the variable "treat".

2. Compute descriptive statistics that allow a comparison of the groups.

3. Compute a 95% confidence interval for the difference between the treatment and control groups. Record it here:

4. Record the results of the t-test of the null hypothesis that there is no mean difference between the treatment and control groups.

5. What is the size of the effect as given by Cohen's $d = (\bar{x}_1 - \bar{x}_2)/s_p$?

6. Based on the analyses above, what do you conclude about the change in average heart-rate (baseline to video game) for the comparison of treatment and control groups?

The idea of regression with code variables – By hand...

Now consider analyzing the data using simple linear regression with a code variable representing the group (control/ treatment).

7. The group variable, "treat", has already been created in the data file. It can be used as the code variable in regression. Examine "treat" and write down the value labels:

   - treat= 0 for the _____ group.
   - treat= 1 for the _____ group.

8. Write down the statistical model used for predicting "vdifpost" (difference in heart-rate) from the group variable "treat". Note: Write down the equation in terms of parameters (e.g., α and β_1) rather than the estimated parameter values.

9. (a) Write down the estimated regression model for the control group.
   (b) Write down the estimated regression model for the treatment group.

10. (a) In the 'Standard Analysis' above you should have recorded the group means. Using these values to estimate the population parameters, work out the estimated values of α and β_1. (Remember: a is the estimate of α and b_1 is the estimate of β_1.)
    (b) Write down the fitted regression model.
    (c) Consider the confidence intervals for α and β_1. Which one is the confidence interval for the difference between the means of the control and treatment groups?

The idea of regression with code variables – Using software...

Now let's produce the analysis using software.

11. Regress "vdifpost" on "treat". Make sure that you request confidence intervals for the parameters.

12. (a) Examine the output and check if the estimates of α and β_1 correspond to those you have calculated manually.
    (b) Write down the confidence interval for the difference between the group means.
    (c) Compare this confidence interval with the interval found in Exercise 3. Explain what you find.

Why the values of the code variable matter...

13. Consider now what might happen if the code variable "treat" was defined in a different way. Now assume that the code variable for group is defined in the following way:

    - treat = 2 for the control group.
    - treat = 1 for the treatment group.

    (a) Write down the estimated regression model for the control group. Note: Write down the equation in terms of parameters (i.e., α and β_1) rather than the estimated parameter values.
    (b) Write down the estimated regression model for the treatment group. Note: Write down the equation in terms of parameters (i.e., α and β_1) rather than the estimated parameter values.
    (c) Show how to estimate α and β_1.
    (d) How do the estimates differ from those you found earlier? Why do they differ?
    (e) Again consider the confidence intervals for α and β_1. Which one will give a confidence interval for the difference between the means of the control and treatment groups? How does it differ from the confidence interval for the difference that you found earlier?

8.2 More than two groups

Next consider comparisons of several groups. We will use a study by Allen et al. (1991) that examined ways of moderating autonomic responses to stress. They compared the physiological responses of women performing a backward counting task under one of three conditions: (i) experimenter present (control condition), (ii) female friend and experimenter present, or (iii) pet dog and experimenter present.
The data file is: `Allen, Blascovich, Tomaka & Kelsey 1991.sav`. We will consider the mean pulse rate during the tasks; it is called "mean" in the data file.

14. In a comparison of three groups, how many code variables are needed?

15. Consider creating code variables in the following way:

Group	Code variable 1 "friend"	Code variable 2 "pet"
Control	0	0
Friend present	1	0
Pet present	0	1

    (a) How are the groups identified in the data file?

    (b) Recode variable "condit":

    Load the data frame `AllenRecoded.txt` in a new JASP session. This data frame includes the two code variables.

    (c) Look at the the newly created variables in the data set to confirm that the new variables are correctly coded.

16. (a) Write down the statistical regression model for predicting mean pulse rate from "friend" and "pet".

    (b) Now write down the estimated regression model for each group.

    Control: $\bar{y}_{\text{control}} =$ _____.
    Friend: $\bar{y}_{\text{friend}} =$ _____.
    Pet: $\bar{y}_{\text{pet}} =$ _____.

    (c) Use the three equations from (b) to show how to estimate α, β_1, and β_2.

17. Consider the confidence intervals for α, β_1, and β_2. Describe what each of these confidence intervals estimates.

    α:
    β_1:
    β_2:

18. (a) Compute the means for each group and record them here. You can find a helpful paragraph on how to do this in your quick guide at the beginning of the reader. We will use the ANOVA menu for this, you will learn more about ANOVA in week 10.1.

    Control: $\bar{y}_{\text{control}} =$ _____.
    Friend: $\bar{y}_{\text{friend}} =$ _____.
    Pet: $\bar{y}_{\text{pet}} =$ _____.

    (b) Estimate each of the parameters *by hand*:

    α: $a =$ _____.
    β_1: $b_1 =$ _____.
    β_2: $b_2 =$ _____.

19. Using software, find the regression equation and confidence intervals for the parameters.

    (a) Check that your estimates in Exercise 18b are correct.

(b) Record the confidence intervals for each of the parameters:

(c) Write a short summary of the findings including a description of the confidence intervals.

Week 9: Regression Assumptions, Contrasts and Post Hocs

The aims of today's activities are to:
(i) practice assessing the regression assumptions in regression with code variables;
(ii) practice defining contrasts to answer research questions;
(iii) test contrasts.

This practical continues the analysis of the same data from last week. This week we will test contrasts. Before we use contrast we will revisit testing regression assumptions, discussed in week 7, and regression with code variables, discussed in week 8.

9.1 Recap: Regression Assumptions

Allen, Blascovich, Tomaka, and Kelsey (1991) were interested in ways of moderating autonomic responses to stress. One way to reduce stress is by providing some kind of support. The researchers measured physiological responses during demanding backward-counting tasks (mental arithmetic is a very reliable way to induce stress in experiments!). The participants were 45 female dog owners. They were asked to perform the task under one of three conditions: (i) experimenter present (control condition), (ii) female friend and experimenter present, or (iii) pet dog and experimenter present. The researchers believed that less autonomic reactivity would occur with pets present compared with friends present. The pets were expected to provide "non-evaluative" support.
Load the datafile `AllenRecoded.txt`. Redo your regression analysis from last week using `Mean` as dependent variable and the code variables `Friend` and `Pet` as covariates.

1. Now produce a plot to test the assumption of normality of residuals and a plot to test the assumptions of homoscedasticity and linearity. What do you conclude?

As you might have observed, the residuals are normally distributed eventhough our predictors are clearly not normally distributed (they only take on values of 0s or 1s).

9.2 Contrasts for planned comparisons

Now we will use contrasts for planned comparisons. The data file is again: `Allen, Blascovich, Tomaka & Kelsey 1991.sav`.

The researchers were interested in two questions:
(i) What was the difference between the control condition (experimenter only) and the two supported conditions (with friend or with pet dog)?
(ii) What was the difference between the two supported conditions?

1. Describe, in words, a contrast for comparing the control condition with the two supported conditions.

2. Write a null hypothesis for this contrast (first contrast), and then write down the contrast as a combination of population means.

3. What are the coefficients for this contrast? Write them in the table below.

	Control	With friend	With pet dog
Coefficients for first contrast			
Coefficients for second contrast			
Code values used by software			

   (a) Now write down a contrast for comparing the friend condition with the pet dog condition (second contrast).

   (b) What are the coefficients for this contrast? Write them in the table.

4. In the data file, each condition (control, with female friend, with pet dog) is assigned a code value by the statistical software. Write those code values in the table above.

9.3 Predefined contrasts

Some types of contrast are common and re-occur in many settings. For instance, when comparing a control group with several treatment groups, or comparing the scores in subsequent groups.
Many statistical packages have a number of these common types of contrast build in. Statistically, there is no benefit of using these above contrasts where you manually enter the contrast coefficients. The benefit lies in that it saves you time and reduces the probability of making mistakes in entering coefficients. The table below gives a description of a number of contrasts commonly provided and includes examples for the three groups we are investigating.

Name of contrast	Description	Example
Deviation	Compares the mean of each level (except a reference level) to the mean of all of the levels (grand mean). The levels of the factor can be in any order.	Control *vs* Mean of all groups; Friend *vs* Mean of all groups
Simple	Compares the mean of each level to the mean of a specified level. This type of contrast is useful when there is a control group. You can choose the first or last level as the reference.	Control versus Friend; Control versus Pet
Difference	Compares the mean of each level (except the first) to the mean of previous levels. (Sometimes called reverse Helmert contrast.)	Friend versus Control; Pet versus Mean of (Control & Friend)
Helmert	Compares the mean of each level of the factor (except the last) to the mean of subsequent levels.	Control versus Mean of (Friend & Pet); Friend versus Pet
Repeated	Compares the mean of each level (except the last) to the mean of the subsequent level.	Control versus Friend; Friend versus Pet

5. What is the type of contrast called we used above.

9.4 Confidence intervals for contrasts

6. We will investigate the two contrasts that we have just defined. Run an analysis with the two contrasts above included. For this we use ANOVA models, we will learn more about ANOVAs during next week's lecture. You can consult the quick guide in the beginning of your reader for help on contrasts and Post Hoc tests. For now focus on the `Contrasts` table and ignore the `ANOVA` table.

7. Confirm that the coefficients have been correctly defined.
8. Examine the results for each contrast and record them in the table below.

	Mean	95% confidence interval
Contrast 1 (Control vs Support)		
Contrast 2 (Friend vs Pet)		

Note that not all software packages calculate a confidence interval for the contrasts. However, it is easy to calculate the confidence intervals by hand. This is also a required skill for the final evaluation of the course, therefore it is a good idea to practice at this point.

9. Compute the standard error for each contrast by hand. The formula is given by

$$SE_c = s_p \sqrt{\sum_i \frac{a_i^2}{n_i}},$$

where s_p = pooled SD, a_i = contrast coefficient in group i, and n_i = sample size of group i.

   (a) Compute s_p using the software output.

   (b) The contrast coefficients a_i were saved in the table of Exercise 3. Moreover, each group has a sample size equal to 15.

   (c) You can now compute the standard error for each contrast. Save the result below.
   $SE_{\text{contrast 1}} =$ _____.
   $SE_{\text{contrast 2}} =$ _____.

10. Finally, compute the 95% confidence interval for each contrast. Save the result in the table above.

11. (a) What can you conclude about the two research questions, given the confidence interval computed for each contrast?

    (b) Why is the value of the first contrast so close to zero?

9.5 Simultaneous confidence intervals for one grouping factor

What the experts say...

> "We recommend using contrasts as the basis for computing specific effects and their associated confidence intervals."
>
> Masson & Loftus, 2003

It is always preferable to use confidence intervals for predefined contrasts to investigate theoretical questions, rather than simultaneous confidence intervals for all differences among the population means. However, now assume that there were no predefined contrasts for the study we have been analyzing today, so you will perform multiple comparisons.

12. Obtain Bonferroni simultaneous confidence intervals for differences between groups in mean heart rates.
13. Write a short description of the simultaneous confidence intervals that includes an explanation of what they suggest about the role of friends and pets in moderating responses to stress.

Week 10: One-way ANOVA

The aims of today's activities are to:
(i) practice using visual displays for comparing independent groups;
(ii) learn to perform a One-way ANOVA analysis with software.

10.1 Visualizing data for group comparisons

Age of congress members

The website fivethirtyeight was interested in the age of U.S. congress members: https://fivethirtyeight.com/features/both-republicans-and-democrats-have-an-age-problem/. They found that despite Obama inspiring young people, American congress members are still quite old on average. Today we will find out if this is true for congress members of all political parties or if there are differences between parties. The data file `congress.csv` contains the age and party affiliation of all congress members between 1947 and 2014. The variables in the data file are:

`age`	the age of a congress member when elected into congress;
`party`	the party affiliation of the congress member: Democrat, Republican or Other.

1. What is your research hypothesis?

2. (a) Create a side-by-side boxplot and histograms to display any differences between the three parties. Sketch the displays below.

   (b) Describe the distribution within each party and mention any differences highlighted by the plots.

3. (a) Inferential procedures (hypothesis tests and confidence intervals) for designs with several independent groups usually make two requirements about the data distribution. What are they?
   (b) Examine the plots that you have produced. Do the data suggest that the assumptions might not be met?
   (c) Assess the normality of the data using the boxplots created earlier.

10.2 One-way ANOVA

4. Write down the null and alternative hypotheses that are tested using one-way ANOVA.
5. Fit the one-way ANOVA model to our data. Report the descriptives (means, SDs, sample sizes for each group), the 95% confidence interval (CI) for each group, and the test results (the ANOVA table).
6. Start by examining the descriptive statistics. What does the pattern of means suggest about the differences between the groups?
7. Look at the CI for each group. What are these intervals? How are they calculated?
8. What is the value of s_p, the pooled standard deviation (hint: Use the ANOVA table)?
9. Now look at the ANOVA table. Be sure that you know exactly what each entry in the table stands for. In particular, try exploring the numerical relationships between the entries. See whether you can answer these questions:
   (a) What is the sum of the between groups SS and the within groups SS?
   (b) Why are the degrees of freedom equal to 2, 13632, and 13634?
   (c) How can the between and within mean square values be computed using the first two columns of the table?
   (d) How can the F value be computed?
10. What is the p-value? What does the p-value in the table suggest?
11. Fill in the spaces in the following sentence:

    The _____ (larger/smaler) the F value, the smaller the associated p-value (for a fixed pair of df's). Also, the _____ (larger/smaller) the error MS is with respect to the group MS, the smaller the p-value is.

12. Under H_0, both the group and the error MS are estimates of the same parameter. Which parameter is this? What can be said in this case?

10.3 Manual computation of a one-way ANOVA

Although it is a time-consuming and somewhat boring task, your insight in the principles behind one-way ANOVA may greatly benefit from performing all computations by hand at least once. That is what we will do in this exercise.
Data have been collected for three groups with four measurements per group:

Group A	8	10	12	14
Group B	7	8	11	12
Group C	10	12	12	16

The goal of this exercise is to perform a full analysis of variance and to fill in the following table

	SS	df	MS	F
Group	...	...	...	...
Error	...	...	...	
Total	...	...	...	

13. 'Computing' the degrees of freedom is the easiest part. Fill in the correct values in the table.

14. The next step is the most work: Filling in the sums of squares. To obtain the sum of squares, you need to compute three squares *per observation*. These squares are the squares of the terms in the equation

$$\underbrace{(y_{ij} - \bar{y})}_{\text{'total'}} = \underbrace{(\bar{y}_i - \bar{y})}_{\text{'group'}} + \underbrace{(y_{ij} - \bar{y}_i)}_{\text{'error'}},$$

where:

- $\bar{y}$ = overall mean, that is, the mean of the 12 scores available, computed disregarding group membership. So $\bar{y} = \frac{8+10+12+14+7+8+11+12+10+12+12+16}{12}$.
- $\bar{y}_i$ = mean of group i. For example, $\bar{y}_1 = \frac{8+10+12+14}{4}$.
- y_{ij} = score j in group i. For example, $y_{14} = 14$ and $y_{32} = 12$.

(a) Confirm that the group means are 11, 9.5, and 12.5, respectively, and that the overall mean is equal to 11. Fill these values in the table below (columns '$\bar{y}_i$' and '$\bar{y}$').

(b) Subsequently, compute the three terms of the above equation for all twelve observations and record them in the table below (columns '$(y_{ij} - \bar{y})$', '$(\bar{y}_i - \bar{y})$', and '$(y_{ij} - \bar{y}_i)$').

(c) Next, square the values of the three columns that you filled in in last question. Record the results in the table below (columns '$(y_{ij} - \bar{y})^2$', '$(\bar{y}_i - \bar{y})^2$', and '$(y_{ij} - \bar{y}_i)^2$').

(d) As a final step, compute the sum of squaresin the table below. Fill in these values in the table above under 'SS' (=sum of squares).

(e) Once you have the SS and df, obtaining the MS and F is straightforward. Compute them and record them in the table above.

(f) You can double-check your own computations by running a one-way ANOVA analysis on these data using software.

Group	y_{ij}	$\bar{y}_i$	$\bar{y}$	$(y_{ij} - \bar{y})$	$(\bar{y}_i - \bar{y})$	$(y_{ij} - \bar{y}_i)$	$(y_{ij} - \bar{y})^2$	$(\bar{y}_i - \bar{y})^2$	$(y_{ij} - \bar{y}_i)^2$
A	8								
A	10								
A	12								
A	14								
B	7								
B	8								
B	11								
B	12								
C	10								
C	12								
C	12								
C	16								
							$SS_T =$	$SS_G =$	$SS_T =$

15. Test the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$. What is the result?

Week 11: Two-way ANOVA

The aims of today's activities are to:
(i) learn about visual displays for two-way ANOVA designs;
(ii) use software for ANOVA designs with two factors;
(iii) learn about displays for confidence intervals for two-way ANOVA designs;
(iv) practice a complete analysis of a two-way ANOVA design.

11.1 Exploring and visualizing data for two-way ANOVA designs

Age and Memory

Michael Eysenck (1974) investigated the role of age and type of processing on memory. He used a word memory task. He randomly assigned 50 younger participants and 50 older participants to one of five learning groups.
Here are the tasks, in increasing order of complexity, that each of the five learning groups were asked to do while reading a list of 27 words:

Counting	Count the number of letters in each word (lowest level of processing).
Rhyming	Think of a rhyme for each word.
Adjective	Think of an adjective to modify each word.
Imagery	Think of a visual image for each word.
Intentional	Try to memorize each word.

Each participant read the word list three times and then was asked to recall as many words as possible. Eysenck recorded the number of correct words recalled.

The data file is `Eysenck 1974.jasp`.

1. The study has two independent variables or factors. What are they?
2. Open the data file and identify the names for the independent variables or factors.
3. First, we will investigate the distribution of the data. One way to show two factors on visual displays is by using one of the factors to group or cluster parts of the display.
   (a) Obtain two boxplots of number of words recalled, one split for age group and one split for level of processing. Also build a so-called 'means plot': A plot showing the means of the dependent variable for different values of predictor variable(s). In this case, put level of processing on the x-axis, mean number of words recalled on the y-axis and draw the means for both the 'older' and 'younger' groups. As always, make sure your nominal variables are recognized by JASP as nominal (check the symbol next to the variable name in the variable view).
   (b) Examine the display and describe the pattern of results.
   (c) Which of the two boxplot displays do you prefer? Why?

(d) Do there appear to be differences between the levels of processing?

(e) Do there appear to be differences between young and old?

(f) Is the pattern of differences between the levels of processing the same for the younger and older groups?

4. Compute the mean, SD, and sample size in each of the ten experiment groups that result from fully crossing the levels of factors 'age' and 'type of processing'. Record the results in the table below.

	Counting	Rhyming	Adjective	Imagery	Intentional
Younger	Mean = n = SD =	Mean = n = SD =	Mean = n = SD =	Mean = n = SD =	Mean = n = SD =
Older	Mean = n = SD =	Mean = n = SD =	Mean = n = SD =	Mean = n = SD =	Mean = n = SD =

5. Examine the table and describe patterns in the means and standard deviations.

6. Often it is easier to see the patterns of means in a visual display. Which of the plots created above do you prefer?

11.2 Checking the assumptions

7. What are two assumptions about the distribution of words recalled in each condition?

8. What do the boxplots (created earlier) suggest about the shape of the distribution (in each condition)?

9. What do the summary statistics and boxplots suggest about the assumption relating to the variability in each condition?

10. Why might there be so much variation in the standard deviations across the conditions?

11. What do you conclude? Do you think it is reasonable to use ANOVA for these data? Explain why or why not.

11.3 Running the two-way ANOVA

Consider the following rule for examining standard deviations in ANOVA (from Moore, McCabe, & Craig):

> "If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumption of equal standard deviations, and our results will still be approximately correct."

They give some further advice:

> "When the sample sizes in each group are very small, the sample variances will tend to vary much more than when the sample sizes are large. In this case, the rule may be a little too conservative. (...) Careful judgment is needed in all cases. By considering p-values rather than fixed level alpha testing, judgments in ambiguous cases can more easily be made; for example, if the p-value is very small, say 0.001, then it is probably safe to reject H_0 even if there is a fair amount of variation in the sample standard deviations."

We will proceed with the ANOVA keeping Moore et al.'s advice in mind.

12. (a) Consider the following ANOVA table. Start by completing the 'Degrees of freedom' column.

Source	Degrees of freedom	Sum of squares	Mean square	F	p-value
Age					
Process					
Age × Process					
Error					
Total					

(b) Obtain the two-way ANOVA table using software.

(c) Now complete the ANOVA table above. Take care particularly to use the appropriate totals.

13. What do you conclude about:

(a) the effect of age;

(b) the effect of process;

(c) the effect of age and process together?

11.4 Describing and interpreting interactions

Examining the distribution of the data, the model assumptions, and the ANOVA table are part of a complete analysis of a two-way ANOVA design. It is also important to describe the patterns of means and the confidence intervals for the means.

14. Look at the visual display of the sample means that you have sketched above.

(a) Describe the pattern of means for the older group.

(b) Describe the pattern of means for the younger group.

(c) Is your description of the pattern for the old and young group the same?

11.5 Two-way ANOVA: Graphical approach (plotting CIs)

We continue with the two-way analysis, now we focus on the computation and interpretation of confidence intervals in two-way ANOVA.

> "Careful inspection of the means is necessary to interpret significant main effects and interactions. Plots are a useful aid."

As in the case of one-way ANOVA, you can use visual displays of means and confidence intervals as an aid to interpretation. Visual displays can help you describe which means differ and the extent to which they differ.
It is important to use appropriate estimates of variability in calculating the confidence intervals. You can use the individual estimates from each group or a pooled estimate of variability.
Here is some advice from Masson & Loftus (2003):

> "*Designs with Two Levels of Each Factor*
>
> ... For a pure between-subject design, there is only a single MS error term (MS_{within}), representing a pooled estimate of variability. In this case, a single confidence interval can be constructed... This confidence interval can be plotted with each mean and used to interpret the pattern of means. If there is a serious violation of the homogeneity of variance assumption (...), separate confidence intervals can be constructed for each group in the design..."

Let's now examine both approaches to finding the confidence intervals so we can see the pros and cons of using the different variance estimates.

15. Obtain a visual display, for the ten groups, a means plot including each group's SE.

16. Examine the visual display.

    (a) For which, if any, levels of processing are the differences between young and old very small?

    (b) For which, if any, levels of processing are the differences between young and old relatively large?

17. Now let us compute by hand, for each group, the 95% confidence interval based on the pooled estimate of the within group variability.

    (a) Recall the formula to compute the required confidence intervals: $\overline{x} \pm t^* \times SE$, with $SE = \frac{s_p}{\sqrt{n_i}}$. The mean and sample size of each group were already computed in Exercise 4.

    (b) Compute s_p using the software output.

    (c) Compute the standard error of each group's mean. Observe that the same standard error applies to all groups because all groups have equal sample size. Save the result below.
    $SE =$ _____.

    (d) We still need the critical value t^*. How many degrees of freedom for the SSE are there? Get the most approximated value for t^* from the textbook's t-table.

    (e) Finally, compute and record the confidence intervals in the table below (LB=lower bound of the CI; UB=upper bound of the CI). Use two decimal places.

	Counting	Rhyming	Adjective	Imagery	Intentional
Younger	Mean = LB = UB =	Mean = LB = UB =	Mean = LB = UB =	Mean = LB = UB =	Mean = LB = UB =
Older	Mean = LB = UB =	Mean = LB = UB =	Mean = LB = UB =	Mean = LB = UB =	Mean = LB = UB =

18. What conclusions do you draw from the confidence intervals? Do they differ from the conclusions you drew in Exercise 16?

19. What do you conclude about the effects of age and level of processing on memory? Write a short summary statement that explains the findings, and refer to one of the visual displays that you have generated.

Week 12: Introduction to Bayesian Statistics

The aims of today's activities are to:
(i) learn about the Monty Hall problem;
(ii) practice calculating posterior probabilities using Bayes' rule;
(iii) calculate Bayesian credible intervals with JASP.

12.1 The Monty Hall problem

This is a classic probability problem that has been around for almost 50 years, but that only became highly controversial in the early 90s. You can see the Wiki for interesting information: https://en.wikipedia.org/wiki/Monty_Hall_problem.

Here is the problem:

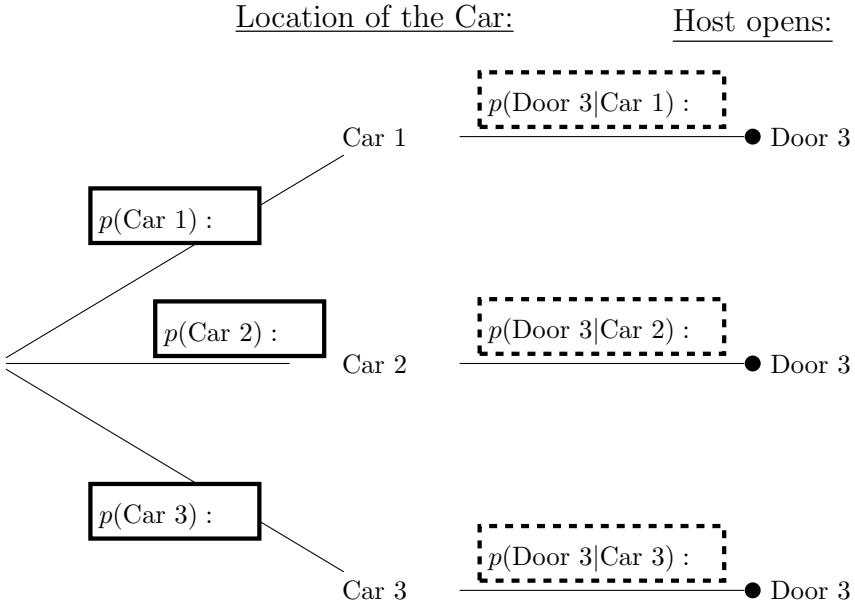
> *Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?*

Let's use the Bayes' rule to try to answer the question. It is important that we completely clarify the assumptions made (taken directly from Wiki):

- The host must always open a door that was not picked by the contestant.
- The host must always open a door to reveal a goat and never the car.
- The host must always offer the chance to switch between the originally chosen door and the remaining closed door.

1. Let's follow the problem in steps. Assume you randomly picked door No. 1. Of course, we have no information about where the car is. Fill in the three solid boxes with probabilities (first tree branching) in the tree below with the prior probabilities for the location of the car, $p(\text{Car } i)$.

   In the tree below 'Car i' means 'the car is behind door No. i'.



2. Next, the host opens door No. 3. Because we know that the host cannot open the door with the car, we know that the car is behind either door No. 1 or 2. The key to understanding the Monty Hall problem lies in realizing that the host opening door No. 3 is more likely for some positions of the car than for others. We will now calculate these conditional probabilities. Fill in the three dashed boxes with probabilities (second tree branching) in the tree above. These probabilities are the conditional probabilities $p(\text{Door } 3|\text{Car } i)$ of opening door No. 3, given that the car is behind door No. i.

   In the tree above 'Door 3' means 'the host opens door No. 3', this is our data.

3. Now we compute the product of the probabilities for each branch of the tree, that is, we compute $p(\text{Car } i) \times p(\text{Door } 3|\text{Car } i)$. Fill in the third column in the table below, including its sum at the bottom.

Car	Door	$p(\text{Car } i) \times p(\text{Door } 3\|\text{Car } i)$	$p(\text{Car } i\|\text{Door } 3)$
1	3		
2	3		
3	3		
	Sum =		

   The sum is equal to $p(\text{Door } 3)$ by the law of total probability:

$$p(\text{Door } 3) = \sum_{i=1}^{3} p(\text{Car } i) \times p(\text{Door } 3|\text{Car } i).$$

4. Finally we use the Bayes' rule to compute $p(\text{Car } i|\text{Door } 3)$ for $i = 1, 2, 3$ for the fourth column in the table above. These are the posterior probabilities for the car location, *after* observing that the host opened door No. 3 with a goat:

$$p(\text{Car } i|\text{Door } 3) = \frac{p(\text{Car } i) \times p(\text{Door } 3|\text{Car } i)}{p(\text{Door } 3)}$$

   Before doing the computations, do make sure you understand the formula above. Afterwards, fill in the last column in the table above.

5. The Monty Hall problem is concerned with assessing whether you increase your chances to win the car by changing to door No. 2, or rather stick to your initial pick (door No. 1). This can now be easily assessed by comparing $p(\text{Car } 1|\text{Door } 3)$ and $p(\text{Car } 2|\text{Door } 3)$ (of course, $p(\text{Car } 3|\text{Door } 3) = 0$ because the host opened door No. 3 so it could not have the car). What do you conclude?

6. It is fascinating how this problem can be solved using relatively elementary mathematics, and yet it has stirred up quite a commotion back in the 90s. A lot of intelligent people fell into the trap of concluding that there is no advantage in changing doors; after all, each door has an initial equal probability of $\frac{1}{3}$ of hiding the car. The mistake in the previous reasoning can be expressed as follows (from the Wiki):

   > *"(...) probabilities are expressions of our ignorance about the world, and new information can change the extent of our ignorance."*

   Discuss below how Bayes' rule is of value in incorporating information into the analysis.

12.2 Coin bias revisited

We revisit the coin bias example used in the lectures. Our scientific goal is to infer about the bias of a coin. Let $\theta = p(\text{heads})$. If the coin is fair then $\theta = .5$, otherwise it is either smaller than .5 (biased towards tails) or larger than .5 (biased towards heads).
The data file `coin.csv` contains the outcome of several coin tosses, all performed with the coin in question.

```
Outcome
Heads
Tails
Heads
.........
```

7. Load the data in **JASP**.

8. Start by looking at the descriptives. What is N (total number of throws) and x (total number of heads) for your data?

9. Carry out a Bayesian binomial test to assess if $\theta = .5$. It should be noted that the Bayesian component of JASP is testing-oriented (hence the menu's name). We don't want to delve into this in Statistics 2, so we will go through the output keeping an eye out for the prior and posterior distributions.

   A prior and posterior for θ are plotted. Focus only on the plot for Outcome - Heads. Neglect the Bayes factors, they are beyond the scope of this week's practical. What is the prior distribution? Try playing with it by changing the beta parameters `a` and `b` (they can be any positive real number). See how the posterior changes as you change the prior (look at the curve of the posterior distribution, and look at the summaries shown on the top-right, namely, the median and the central 95% credible interval). Note that the data are always fixed, so any change in the posterior is solely due to changes in the prior.

10. This might feel uncomfortable to you: How can we make any inference about θ if changing the prior has this effect on the posterior? Discuss with your partner: What should we choose for our prior after all? What do you think should be reported when writing down the results of a Bayesian analysis?

11. It is not the typical situation, but in this simple case we do know the exact form of the posterior distribution (recall the lecture).

    Let's bring back JASP's default prior, which is the uniform in (0, 1) (i.e., beta(1, 1) with $a = 1$ and $b = 1$). Given the value of N and x for these data (see descriptives), identify the exact posterior distribution that is being plotted in JASP.

12. Bayesian inference can be performed *sequentially*. In theory, we can start with a prior, collect data, and estimate the posterior. We then collect more data and use these to update the old posterior (which becomes the new prior) into a new posterior. And the cycle can go on indefinitely, using a golden rule coined by Lindley in 1970:

    > *"Today's posterior is tomorrow's prior."*

You can see this crucial Bayesian property at work! Using the closed-form posterior distribution for a beta prior and binomial likelihood, fill in the following table in sequence. Make sure that the initial prior is beta(1, 1), that is, the uniform distribution. Do compare the last row of the table with the result from the previous exercise. What do you conclude?

Toss	$N =$ Number of throws	$x =$ Number of heads	Posterior
1			beta($a =$, $b =$)
2			beta($a =$, $b =$)
3			beta($a =$, $b =$)
4			beta($a =$, $b =$)
5			beta($a =$, $b =$)
6			beta($a =$, $b =$)
7			beta($a =$, $b =$)
8			beta($a =$, $b =$)
9			beta($a =$, $b =$)
10			beta($a =$, $b =$)
11			beta($a =$, $b =$)
12			beta($a =$, $b =$)
13			beta($a =$, $b =$)
14			beta($a =$, $b =$)
15			beta($a =$, $b =$)
16			beta($a =$, $b =$)
17			beta($a =$, $b =$)
18			beta($a =$, $b =$)
19			beta($a =$, $b =$)
20			beta($a =$, $b =$)
21			beta($a =$, $b =$)
22			beta($a =$, $b =$)

13. Look at the 95% credible interval provided in the JASP plot of prior and posterior. What is the interpretation of this interval? Next, suppose that interval actually were the 95% *confidence* interval; how would you interpret it instead? Identify the differences between both interpretations. **Obs.:** This exercise is crucial, do pay attention to it. It is important that you understand what a credible interval is, and perhaps even more importantly, what a confidence interval is *not*.

14. Which values of θ received a boost of credibility after observing the data? And which values of θ suffered a loss of credibility after observing the data?

    Now look at the second data file `coin2.csv`. This file contains the identical contents of the original file but repeated 10 times. This is a quick-and-dirty means of having a 10 times larger dataset with exactly the same proportion of heads. Load this file in JASP and find the posterior for a beta(1,1) prior.

    If you put both JASP windows side-by-side (for the 'coin' and 'coin2' datasets), you can compare two posteriors for data that have the same sample proportion of heads and that are based on the same prior. This is one way to visualize the effect of sample size on the posterior (recall lecture).

    How do both posteriors relate? Which one is more precise (i.e., less spread out)? What is the difference between both 95% credible intervals? And between both medians?

15. Also recall Cromwell's rule from the lecture. You can artificially see this in action here. Under `Alt. Hypothesis` on the left-side menu, choose either `> Test value` or `< Test value`. See how the prior and posterior distributions change; describe this. Do you think that truncating the range of θ this way is insightful for the problem at hand?

Week 13: Good statistics, bad statistics

The aims of today's activities are to:
(i) become a pro at HARKing;
(ii) re-do another researcher's analysis using open data from OSF;
(iii) draw conclusions about a meta-analysis using a funnel plot.

13.1 *p*-Hack your way to scientific glory

Intentionally doing things the wrong way can be educational which is why it's allowed to do so in this exercise. We will be *p*-hacking and HARKing ('hypothesizing after results are known') using a text and online applet written by Christie Aschwanden on the website of American statistician Nate Silver.

1. Go to http://fivethirtyeight.com/features/science-isnt-broken.
   (The text on this website is part of the exam material.) Read the first part (up to the grey part).
   (a) *p*-Hack your way to scientific glory: in the web applet, find at least three different combinations of variables that yield a significant result.
   (b) As a next step, go HARKing: for at least one set of significant results, try to find a coherent explanation. (We are aware that you are probably not an expert in American politics. Just try to think up something that at least sounds a bit coherent.)
   (c) Compare your results to the two questions before with your neighbors.
   (d) Recall that *p*-hacking and HARKing are instances of bad science. Do not try this at home (or anywhere else). Do good science instead.
2. Visit https://xkcd.com/882/. What Questionable Research Practice is this an example of?

13.2 Open data on the Open Science Framework

The Open Science Framework (OSF) has become a popular website for researchers to share their data and analysis scripts openly so others can re-analyze and verify their analysis. Suessenbach et al. (2019) investigated how dominance motives relate to donating behavior across experimental manipulations. The authors published analysis scripts (R), preregistration plan and data on OSF: https://osf.io/uxtq2/. The paper is published open access and is freely available at: https://onlinelibrary.wiley.com/doi/full/10.1002/per.2184. The main idea is summarized in a blog post: https://www.nicebread.de/dopl/. (For those interested in open science and QRPs the website is a great resource with many blog posts on these subjects.)

3. In their paper they report " the proportion given in the neutral condition correlated strongly with the proportion given in the arousal condition, $r = .55$, $p < .001$"(Suessenbach et al., 2019). We will redo this small part of their analysis to validate the finding.
   (a) The authors published a preregistration report where they specify all steps of their analysis a priori data collection. Why might this be a useful step to prevent QRPs?

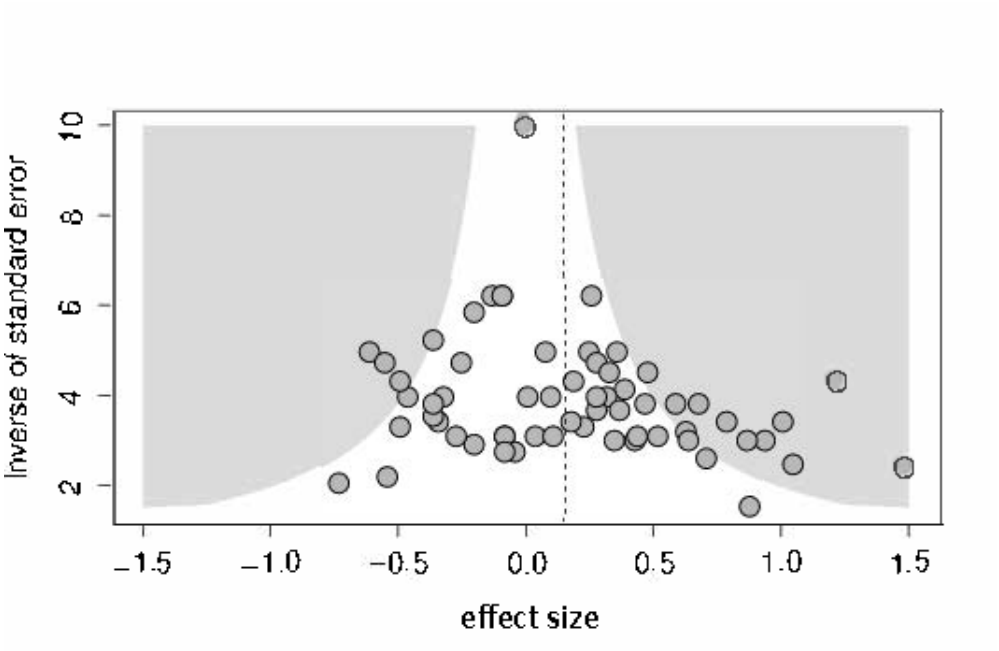
(b) Download the data zip folder form OSF. Unzip the folder and load the data of the second sample of the study: `sample2.csv`. To find out which variable names correspond to the two variables in question consult their codebook (also contained in the data zip folder): `Codebook sample 2.docx`. Calculate the correlation between the proportion given in the neutral condition and the proportion given in the arousal condition with the software of your choice. What do you conclude?

This example is to show you what good science can look like: open data, sharing analysis scripts and a preregistration plan. As you may have noticed though, it can be difficult to make sense of someone else's data and redo their analysis. Good documentation of the data and analysis procedure is crucial for making a study re-analyzable.

13.3 Interpreting Funnel plots

When choosing the best of several options (e.g., cars or apartments) does it help to consciously think about your decision? Some studies found that participants who thought about an unrelated task were more likely to choose the best option compared to participants who performed conscious deliberation. Other studies, however, found no evidence for this so-called unconscious thought advantage (UTA). To settle the question Nieuwenstein et al. (2015) performed a high-powered study ($N = 399$) and combined all previous experiments in a meta-analysis. You can find the open access paper here: http://journal.sjdm.org/14/14321/jdm14321.html.

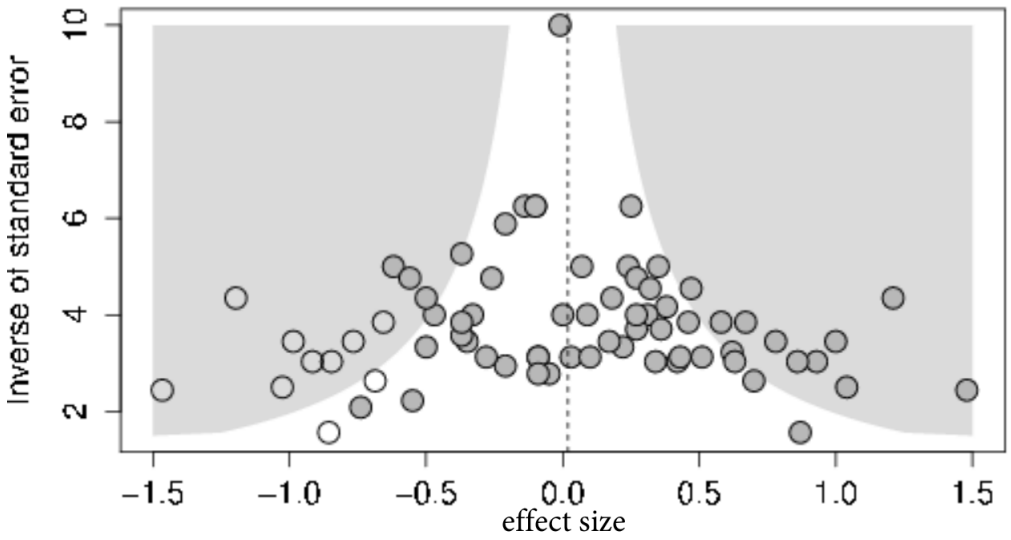
4. Here is the funnel plot combining all published effect sizes.



(a) The study by Nieuwenstein et al. (2015) had the highest sample size of all experiments, which dot corresponds to the experiment by Nieuwenstein et al. (2015)?

(b) What do you conclude based on the plot?

5. To get an impression of the true underlying effect size the researchers filled values into their final funnel plot. These values are based on a complicated statistical method (which we will not discuss in detail here) that makes the plot symmetrical. The filled-in effect sizes are represented by open dots.



(a) Why was it necessary to fill-in effect sizes to get an impression of the true underlying effect-sizes?

(b) What conclusions do you draw about the existence of UTA based on this corrected funnel plot?

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley, 2 edition.

Agresti, A. and Finlay, B. (1997). *Statistical methods for the social sciences*. New Jersey: Prentice Hall.

Allen, K., Blascovich, J., Tomaka, J., and Kelsey, R. M. (1991). Presence of human friends and pet dogs as moderators of autonomic responses to stress in women. *Journal of Personality and Social Psychology*, 61:582–589.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers, third edition.

Eysenk, M. W. (1974). Age differences in incidental learning. *Developmental Psychology*, 10:936–941.

Holzer, III, C. E. (1977). *The Impact of Life Events on Psychiatric Symptomatology*. PhD thesis, University of Florida.

Larkin, K. T., Zayfert, C., Abel, J. L., and Veltum, L. G. (1992). Reducing heart rate reactivity to stress with feedback generalization across task and time. *Behavior Modification*, 16:118–131.

Masson, M. and Loftus, G. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, 57:203–220.

Nieuwenstein, M. R., Wierenga, T., Morey, R. D., Wicherts, J. M., Blom, T. N., Wagenmakers, E., and van Rijn, H. (2015). On making the right choice: A meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment and Decision Making*, 10(1):1 – 17.

Schuppert, H. M., Albers, C. J., Minderaa, R. B., Emmelkamp, P. M. G., and Nauta, M. H. (2014). Severity of borderline personality symptoms in adolescence: Relationship with maternal parenting stress, maternal psychopathology, and rearing styles. *Journal of Personality Disorders*, to appear.

Schuppert, H. M., Albers, C. J., Minderaa, R. B., Emmelkamp, P. M. G., and Nauta, M. H. (2015). Severity of borderline personality symptoms in adolescence: Relationship with maternal parenting stress, maternal psychopathology, and rearing styles. *Journal of Personality Disorders*, 29(3):289 – 302.

Suessenbach, F., Loughnan, S., Schönbrodt, F. D., and Moore, A. B. (2019). The dominance, prestige, and leadership account of social power motives. *European Journal of Personality*, 33(1):7–33.

van Apeldoorn, F. J., Timmerman, M. E., Mersch, P. P. A., van Hout, W. J. P. J., Visser, S., van Dyck, R., and den Boer, J. A. (2010). A randomized trial of cognitive-behavioral therapy or selective serotonin reuptake inhibitor or both combined for panic disorder with or without agoraphobia: Treatment results through 1-year follow-up. *Journal of Clinical Psychiatry*, 71:574–586.