Exercise 3

No Fear of Numbers: Introduction to Quantitative Data Analysis in R

Dr. Susanne de Vogel, Data Science Center, University of Bremen

2 December 2025

Research Question: Is there a Life beyond the PhD?

In this workshop, we will work with one coherent example throughout all four parts.We focus on two outcomes of interest:

- `bpsy01`: overall life satisfaction (10 point scale)
- `blcd06`: Children (yes/no)

We will examine how these outcomes are related to different aspects of doctoral researchers' lives and backgrounds:

PhD and Work conditions

- `adbi01`: Status of the doctorate
- `adbi15`: Discipline
- `bdcd17`: Emotional Support during PhD
- `bwdr12`: Perceived Scientific Pressure
- `bemp81`: Monthly Gross income

Attitudes and Well-Being

- `bldc12`: Satisfaction with work-life balance
- `apar15b`: Relationship with parents
- `bpsy05`: Self-Efficacy

Demographics

- `adem01`: Gender
- `adem02`: Age in years
- `apar10`: Highest vocational degree of parents
- `adem03`: Country of birth
- `blcd01`: Relationship status

Bivariate Analysis

In the previous exercise, we focused on **univariate analysis** looking at one variable at a time. In this sheet, we move one step further to **bivariate analysis**: we look at **pairs of variables** to explore how they are related.

In the context of our research story, our main outcomes of interest are:

- general life satisfaction (`bpsy01`), and

- having children (yes/no, `blcd06`).

Before we build regression models, we first explore **how these outcomes relate to other key variables** (e.g. gender, age).

In this exercise, you will learn to:

- analyse **categorical** × **categorical** relationships using cross-tabulations, row/column percentages, chi-squared tests, and grouped bar charts,
- analyse **categorical** × **metric** relationships using group means, standard deviations, boxplots, t-tests and one-way ANOVA,
- analyse **metric** × **metric** relationships using Pearson's correlation, significance tests, and scatterplots with optional trend lines,
- practice interpreting the numerical results and the plots in simple words (e.g. "There is / is not an association between …").
- export your results as Excel-Sheets or images.

2.1. Set-up: Load packages and dataset

1. Install `tidyverse`, `haven`, `janitor`, `sjmisc`, `skimr`,`scales` and `writexl` if not already installed and load these packages.

2. For this exercise, we will use the file *`03_qa_bi_data.sav`* in the *exercises* folder. Define the *exercise* folder as your working directory with the function `setwd("path")`.

3. Choose `mydata3` as a name for your data frame and import it with the following structure: `chosen_name <- read_sav("filename")`.

2.2. Categorical x categorical: Children (Yes/No) by Gender

In this part we look at the relationship between two categorical variables:

- Outcome (dependent variable): `blcd06` – children (yes/no)
- Predictor (independent variable): `adem01` – gender

To explore this, we will create a **contingency table** (gender × children), compute row percentages and column percentages and discuss which is more useful when children yes/no is our outcome, visualize the relationship with a clustered bar chart and run a **chi-squared test of independence**.

1. First convert `blcd06` and `adem01` into *factors* with labels.

Use `mutate()` and `as_factor()` to transform both variables into a factor variable. Save the result again as `mydata3`.

2. Create a *contingency table* for `blcd01` and `adem01` using `tabyl()` with row and column percentages.

Run the plain code with absolute frequencies, including only cases without `NA`, like this:

```
# cross tabulation children and gender
tab_counts <- mydata3 %>% # store results in new data frame tab_counts
  filter(!is.na(adem01), !is.na(blcd06)) %>% # include only cases without NA
  tabyl(adem01, blcd06) # basic cross tabulation (absolute values)

tab_counts
```

2a. To calculate row percentages (within each gender, how many have / do not have children?), include a last function `adorn_percentages("row")` in the pipe and save the results in a new data frame `tab_row`.

2b. To calculate column percentages (Column percentages: within each children-category, gender distribution), include a last function `adorn_percentages("col")` in the pipe and save the results in a new data frame `tab_col`.

2c. Which one is more useful when children yes/no is our outcome of interest?

3. Use the data frame `tab_row` (which already stores your crosstab with row percentages) and export it as an Excel file named `tab_children_by_gender.xlsx` using `write_xlsx()` in your current working directory.

The code should look like this:

```
write_xlsx(tab_row, "tab_children_by_gender.xlsx")
```

4. Create a *clustered bar chart* to visualize the share of respondents with/without children in each gender groups.

The code should look like this:

```
mydata3 %>%                                          # take data frame mydata3
  filter(!is.na(adem01), !is.na(blcd06)) %>%         # keep only cases without NAs
  ggplot(aes(x = adem01, fill = blcd06)) +      # put adem01 on x-axis, colour bars by children
  ↪  yes/no
  geom_bar(position = "fill") +                      # draw stacked bars, each bar scaled to 100%
  scale_y_continuous(labels = percent) +             # show y-axis as %
  labs(                                              # labels
    x = "Gender",
    y = "Share of respondents",
    fill = "Children (blcd06)",
    title = "Share of respondents with/without children by gender"
  )
```

What do we see regarding the gender differences in parenthood?

5. Export the figure to your working directory as a PNG file named "barchart_children_by_gender.png" using `ggsave()`.

The code looks like this:

```
ggsave(filename = "barchart_children_by_gender.png")
```

6. To check whether the gender differences in parenthood are random, calculate the Chi-Square test.

To do so, you can use the simple cross-tabulation data frame `tab_counts` from earlier with the function `chisq.test(tab_chi)`. What does it tell us?

2.3. Categorial x Metric: Overall life satisfaction x children and gender

We now look at how overall life satisfaction (`blcd01`) differs between groups:

- Children (yes/no) – two groups → independent samples t-test
- Gender – more than two groups → one-way ANOVA

For each comparison we will, compute group means and standard deviations, visualize the differences with boxplots and run the appropriate statistical test.

3.2.1 Life satisfaction by children

1. Compare the mean life satisfaction between respondents with and without children.

To do this, use a `%>%` pipe using the `group_by(blcd06)` function and calculate central tendency measures (mean, sd) and distributions conveniently using the `skim()` function‘.

Note: Filter first with `filter(!is.na(bpsy01), !is.na(blcd06)) %>%` to include only valid categories in the calculations.

The code should look like this:

```
mydata3 %>%
  filter(!is.na(bpsy01), !is.na(blcd06)) %>% # keep cases without missings only
  group_by(blcd06) %>%      # group by children yes/no
  skim(bpsy01)              # skim for life satisfaction by group
```

How does the average life satisfaction differ between groups?

2. Create *grouped boxplots* to visualize the differences in overall life satisfaction by respondents with/without children.

The code should look like this:

```
mydata3 %>%                                   # start with the dataset mydata3
  filter(!is.na(bpsy01), !is.na(blcd06)) %>%  # keep only cases without missings
  ggplot(aes(x = blcd06, y = bpsy01)) +       # map children yes/no to x-axis, life satisfaction to
  ↪  y-axis
  geom_boxplot() +                            # one boxplot for each children group
  labs(                                       # labels
    x = "Children (blcd06)",
    y = "Life satisfaction (bpsy01, 0-10)",
    title = "Life satisfaction by children (yes/no)"
  )
```

What do you see?

3. Test whether the mean life satisfaction differs between people with and without children using the *t-test.*

The code goes like this:

```r
t_test_children <- mydata3 %>%                   # store results in new data frame t_test_children
  filter(!is.na(bpsy01), !is.na(blcd06)) %>%     # drop missings
  t.test(bpsy01 ~ blcd06, data = .)              # t-test: blcd01 by children

t_test_children
```

What does the t-test tell us?

3.2.2 Life satisfaction x Gender

1. Compare the mean life satisfaction between respondents of different gender.

To do this, use a `%>%` pipe using the `group_by(adem01)` function and calculate central tendency measures (mean, sd) and distributions conveniently using the `skim()` function'.

Note: Filter first with `filter(!is.na(bpsy01), !is.na(adem01)) %>%` to include only valid categories in the calculations.

How does the average life satisfaction differ between groups?

2. Create *grouped boxplots* to visualize the differences in overall life satisfaction by gender groups.

What do you see?

3. Test whether the mean life satisfaction differs between people with and without children using the one-way *ANOVA* test.

The code should look like this:

```r
anova_gender <- mydata3 %>%                    # store results in new data frame anova_gender
  filter(!is.na(bpsy01), !is.na(adem01)) %>%   # keep cases without missings only
  aov(bpsy01 ~ adem01, data = .)               # ANOVA model: life satisfaction ~ gender

summary(anova_gender)
```

Is there a significant gender difference in overall life satisfaction?

2.4. Metric x Metric: Life Satisfaction and Age

We look at the relationship between overall life satisfaction (`bpsy01`) and gross income in Euro `bemp81`:

We will calculate the **Pearson correlation coefficient**, test if it is statistically significant and visualize the relationship with a scatterplot.

1. First, calculate the Pearson correlation coefficient between life satisfaction `bpsy01` and monthly gross income in Euro `bemp81`.

Filter out missing values and use `summarise()` with `r_pearson = cor(bpsy01, bemp81)`.

The code should look like this:

```r
# calculate pearsons r
mydata3 %>%
  filter(!is.na(bpsy01), !is.na(bemp81)) %>%        # drop missings
  summarise(
    r_pearson = cor(bpsy01, bemp81)                 # Pearson's r
  )
```

```
## # A tibble: 1 x 1
##   r_pearson
##       <dbl>
## 1     0.149
```

How would you interpret the outcome?

2. Draw a scatterplot to visualize the relationship between both variables.

The code should look like this:

```
mydata3 %>%                                       # start with the dataset mydata3
  filter(!is.na(bpsy01), !is.na(bemp81)) %>%  # keep only cases without missings
  ggplot(aes(x = bemp81, y = bpsy01)) +           # age on x-axis, life satisfaction on y-axis
  geom_point(alpha = 0.4) +                       # draw one point per person (slightly transparent)
  geom_smooth(method = "lm", se = FALSE) +        # draw regression line (linear model, no CI band)
  labs(
    x = "Monthly Gross income in Euro (bemp81)",      # x-axis label
    y = "Life satisfaction (bpsy01, 0-10)",           # y-axis label
    title = "Scatterplot: life satisfaction vs. monthly gross income in Euro"
  )
```

What does the scatterplot imply about the relationship between overall life satisfaction and monthly gross income?

3. Finally, apply a significance test.

Use the filtered data again and run `cor.test()` to obtain the correlation, a p-value, and a confidence interval to test whether the correlation is significantly different from zero.

The code should look like this:

```
mydata3 %>%
  filter(!is.na(bpsy01), !is.na(bemp81)) %>%          # drops cases with missings
  cor.test(                                           # run cor.test on the filtered data
    ~ bpsy01 + bemp81,                                # formula: two numeric variables
    data   = .,                                       # use the piped data frame
    method = "pearson"                                # apply pearson method
  )
```

What is your conclusion?

Take Home checklist: Bivariate analysis in R

Step	Question / task	Useful functions / tools
1	What type of variables are you combining (cat × cat, cat × metric, metric × metric)?	Check variable types with `str()`, `summary()`, `skimr::skim()`.
2	For **categorical** × **categorical**: Is there an association between the two variables?	Contingency tables with `tabyl(var1, var2)`, row/column % with `adorn_percentages()`, `adorn_pct_formatting()`, chi-squared test with `chisq.test()`, grouped bar chart with `ggplot(aes(x = var1, fill = var2)) + geom_bar(position = "fill")`.

Step	Question / task	Useful functions / tools
3	For **categorical** × **metric**: Do group means of the metric outcome differ?	Grouped descriptives with `group_by(cat) %>% skim()`, boxplots with `geom_boxplot()`, t-test with `t.test(outcome ~ cat)`, one-way ANOVA with `aov(outcome ~ cat)`.
4	For **metric** × **metric**: Is there a linear relationship between the two variables?	Pearson correlation with `cor()`, significance test with `cor.test()`, scatterplot with `ggplot(aes(x = xvar, y = yvar)) + geom_point()` (optional `+ geom_smooth(method = "lm", se = FALSE)`).
5	Are the numerical results and the plots telling a consistent story?	Compare effect size (difference in means, correlation) with p-values, confidence intervals and the visual patterns in boxplots / bar charts / scatterplots.