Practice session I

Feedback

Introduction to Statistics with R - Week 5
19.05.2023

Everyone did a really good job!

I will address some little issues though

Plotting barographs with categorical variables

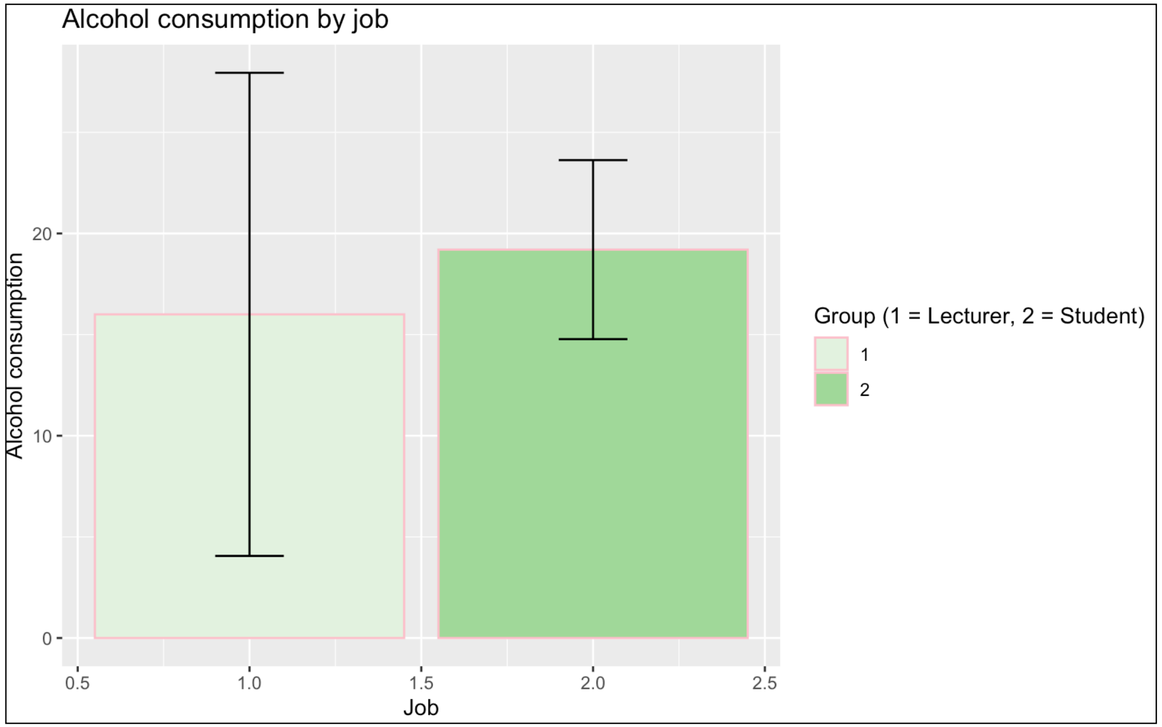

Figure 1: Barplot illustrating alcohol consumption by occupation
- occupation is not defined as a factor

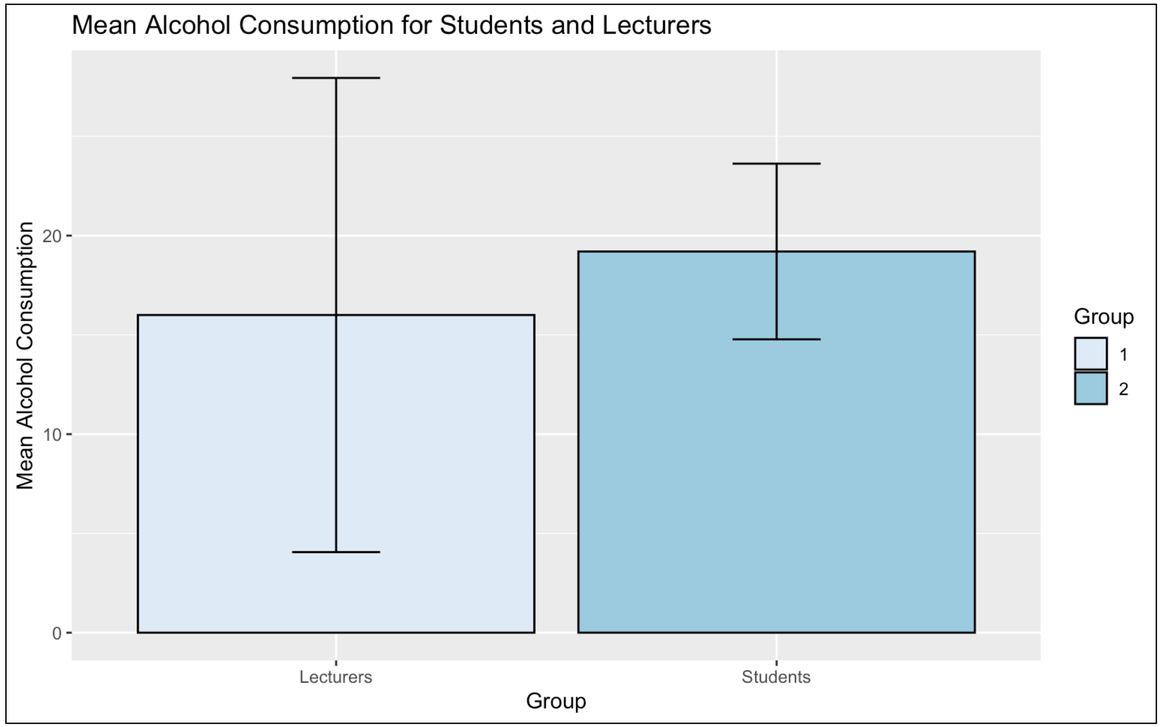

Figure 2: Barplot illustrating alcohol consumption by occupation
- occupation is defined as a factor and recoded (partially)

Whenever you want to plot the relationship between a categorical and a numerical variable:

- **make sure you transform your categorical variable to a factor.**
- **R is not smart enough to know that something is a category, unless it is clearly specified.**
- **When your categories are expressed in numbers, recoding helps (recode() or recode_factor() functions from the dplyr package).**

Data format and ggplot()

Table 1: Data in wide format

A tibble: 32 × 5

subj <dbl>	np_und <dbl>	pa_und <dbl>	np_während <dbl>	pa_während <dbl>
1	984.0000	932.0000	1291.5000	1004.0000
2	1334.2857	1049.7143	1054.5000	816.5714
3	451.0000	523.4286	474.0000	542.5000
4	1333.1429	1401.5000	1010.5000	1108.5000
5	1082.0000	1569.1429	1693.5000	1643.0000
6	1056.0000	341.7143	374.5000	491.3333
7	2590.8571	1043.4286	1468.0000	1480.0000
8	1125.5000	651.5000	1072.5714	961.3333
9	1276.5714	877.1429	1198.0000	880.0000
10	605.3333	524.6667	1020.5714	759.4286

1–10 of 32 rows

Previous 1 2 3 4 Next

Wide data:
- great for “reading” / exploring the dataset, understanding each category
- the wide format is not used all that often in R
- characteristics: data is not stacked, and each variable is distributed in a separate column

Data format and ggplot()

Table 2: Data in long format

A tibble: 128 × 3

subj <dbl>	condition <chr>	reading_times <dbl>
1	np_und	984.0000
1	pa_und	932.0000
1	np_während	1291.5000
1	pa_während	1004.0000
2	np_und	1334.2857
2	pa_und	1049.7143
2	np_während	1054.5000
2	pa_während	816.5714
3	np_und	451.0000
3	pa_und	523.4286

1–10 of 128 rows

The long format is handy for **manipulating: the data is stacked - e.g., in Table 2 a column contains the subcategories of a variable and a separate column contains the corresponding values.**

Most functions in R require the data to be in long format - ggplot() is one of these functions.

For a more in-depth comparison of long vs. Wide data, check out this resource: https://kiwidamien.github.io/long-vs-wide-data.html

Interpreting plots

- I sometimes see that the descriptions of the plots you generate are missing.
- Generating plots is no easy feat, but leaving a plot interpreted leaves the task halfway done.
- If you don't put it in words (as obvious as it might seem), you don't really understand the data patterns deeply.
- Try to describe each plot you see in your own words and try to incorporate some of the vocabulary you have learned so far.
- Who wants to describe Figure 3?

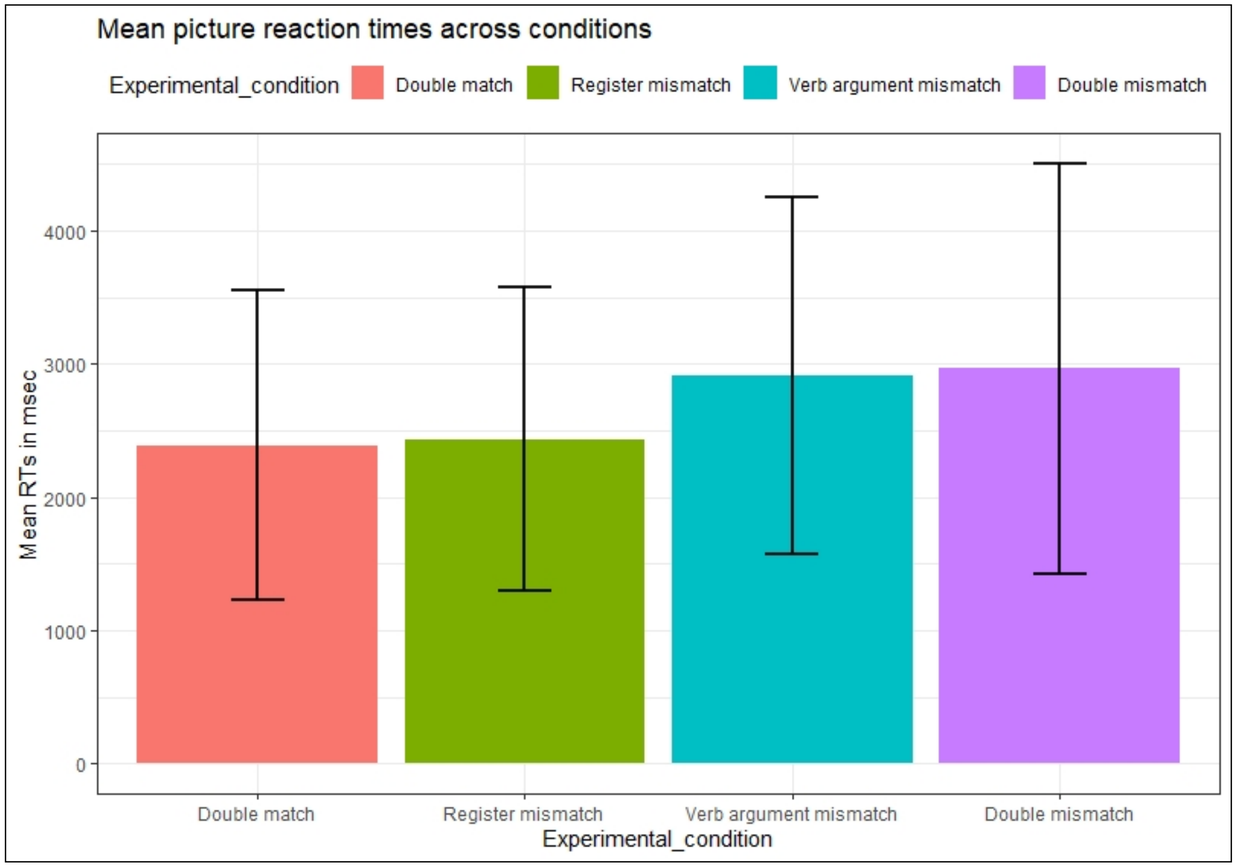

Figure 3: Barplot illustrating mean reaction times in a picture-selection task. Error bars indicate 95% CIs

Interpreting error-bars

- We use error-bars for mainly two reasons:

1) **illustrate the variability of the observations** in the sample we have plotted

2) show **how well the sample mean represents the population mean**

So far, we have used error bars based on 95% confidence intervals:

- in this case, the error bars show the range of values that would capture the population mean 95% of the times

- we can use error bars to ***speculate*** about how the depicted groups might differ, but claims about significant differences can only be made in the context of statistical tests - error bars give us a hint / a first insight.

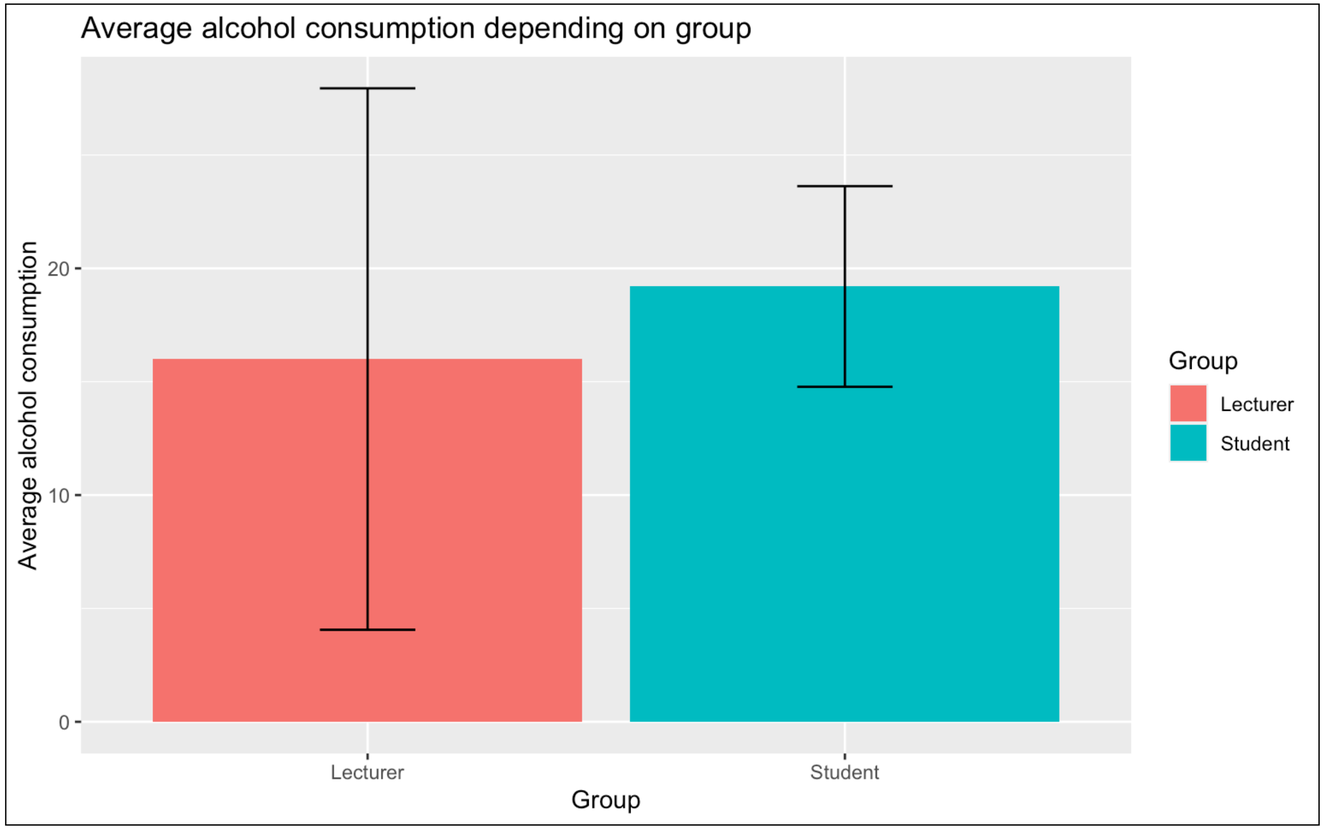

Figure 4: Barplot illustrating mean reaction times in a picture-selection task. Error bars indicate 95% CIs

Interpreting error-bars

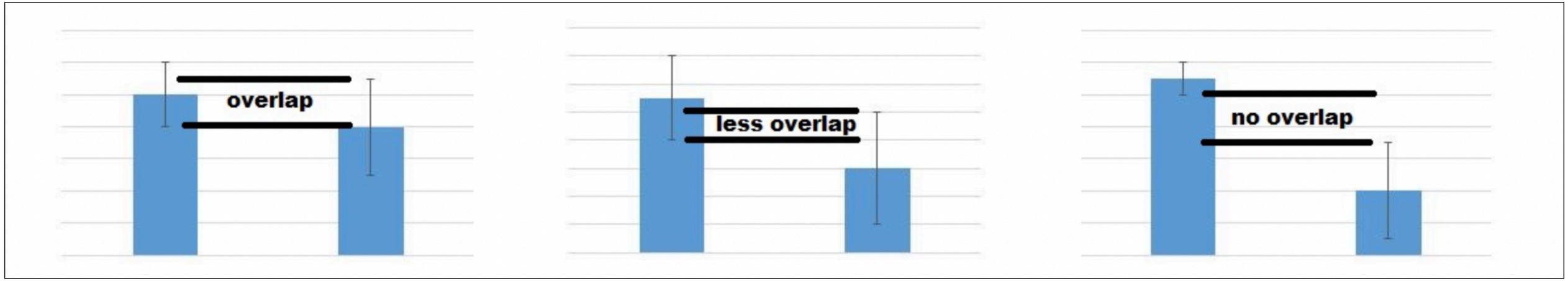

Figure 5: Examples of barplots and error-bar patterns. Source: https://www.biologyforlife.com/interpreting-error-bars.html

When error-bars overlap

- this is a **clue** that the difference between the groups might not be statistically significant

When error-bars overlap less

- this is a **clue** that there might be a difference between the group you are comparing, but you need to check whether the difference is statistically significant or not.

When error-bars do not overlap

- this is a **clue** that the groups you are comparing might be significant different from each other, **but this needs to be confirmed by the results of a statistical test.**

Any questions regarding Script 4?