Lab 12: SPSS II

Table of contents

Descriptives categorical data **3**
 Lab class . . . 3
Selecting and sorting variables . . . 3
Calculating descriptive statistics . . . 4
Recoding variables . . . 5

Descriptives continuous data **8**
 Lab class . . . 8
Calculating descriptive statistics . . . 8
Boxplots . . . 10
Removing participants from SPSS data files . . . 13

Quiz 3 **15**

Explore, apply, reflect **16**
Exercise 1 . . . 16
Exercise 2 . . . 17
References . . . 18

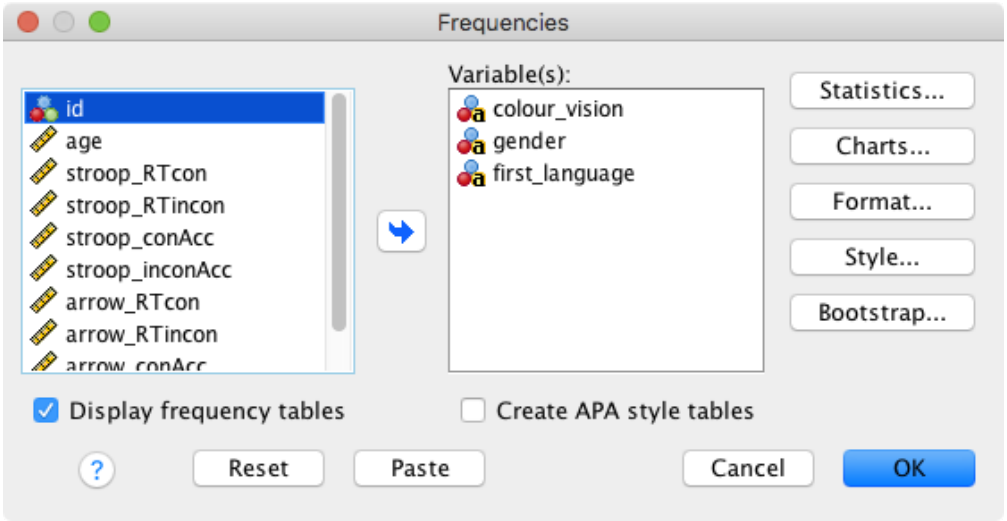
Descriptives categorical data

Lab class

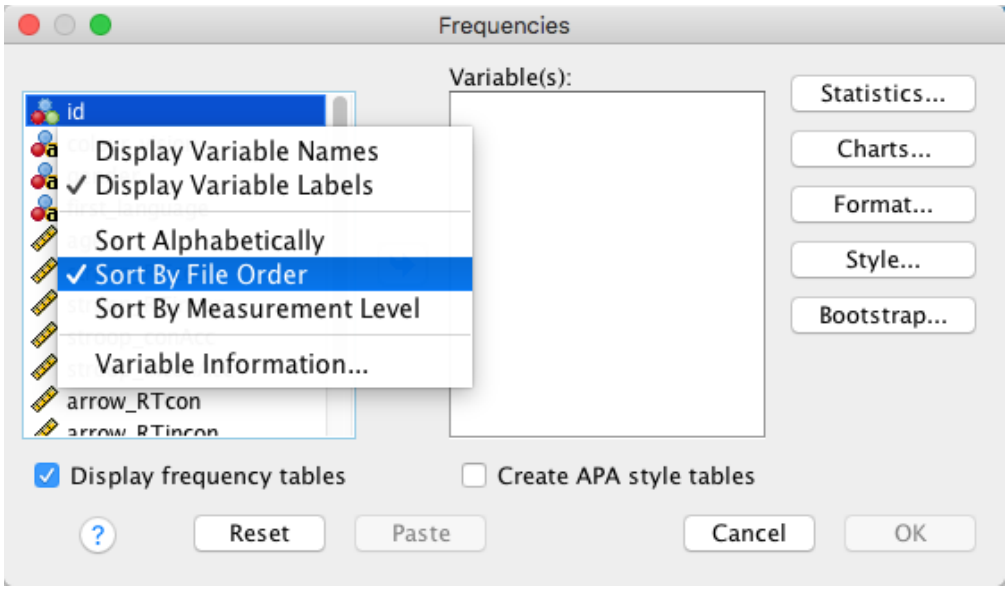
This week we will cover descriptive statistics. In particular, we are going to focus on screening and cleaning data as a step that should precede the calculation of descriptive statistics. We will use the same data as last week. You can either use your SPSS data file to which you added the interference effects or you can re-download the .csv file from last week.

Selecting and sorting variables

Initially, let us have a look at the categorical variables (i.e., `colour_vision`, `gender` and `first_language`). To calculate some descriptive statistics, go to **Analyze** → **Descriptive Statistics** → **Frequencies** and add all categorical variables to "Variable(s)" (it is not necessary to add `id`, as this will not provide us with useful information). Note that what SPSS refers to as "APA style tables" is not really APA style (see here for actual APA style tables).



When the **Frequencies** window opens, you can right-click on the list of variables and change how they are sorted (e.g., you can sort them alphabetically or by measurement level.)



SPSS often requires moving variables from left to right, and vice versa. You do not need to do this individually for each variable. You can use these shortcuts instead:

- Select all variables: `Cmd + A` (as always, Windows users should use `Ctrl` instead of `Cmd`)
- Select a continuous range of variables: select first variable, hold down the `Shift` key, and select the final variable the full range of variables between the first and the final variable will be selected
- Select a few non-adjacent variables: select one, hold down the `Cmd` key, and select the other variables

If you want to move a single variable and there is just one place where it can go, you can simply double-click on it. You can also drag and drop individual variables.

Calculating descriptive statistics

Frequencies offers us the following options:

- **Statistics**: Make sure none of these are selected (apart from the mode, computing these statistics is not meaningful or informative for categorical data and we can easily get the mode from the frequency table we'll create).
- **Charts**: Make sure None is selected (you can choose "Bar Charts", but it won't provide you with useful information that you won't already get from the frequency table).
- **Format**: Not currently of interest.
- **Style**: Not currently of interest.
- **Bootstrap**: Not currently of interest.

Finally, make sure that **Display frequency tables** *is* checked, and click on "OK".

After clicking "OK", an output window will open, displaying a number of tables. The first table allows you to check for missing values:

Statistics

		colour_vision	gender	first_language
N	Valid	156	157	157
	Missing	12	11	11

It turns out that we have up to 12 missing values for our categorical variables.

The next tables about the valid values for each of the variables. Let's have a look at the output for `colour_vision`:

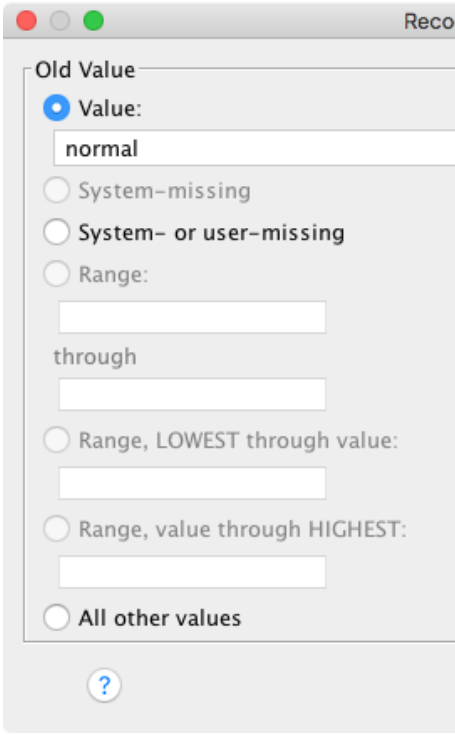
colour_vision

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	n	5	3.0	3.2	3.2
	normal	1	.6	.6	3.8
	y	135	80.4	86.5	90.4
	Y	8	4.8	5.1	95.5
	yes	6	3.6	3.8	99.4
	Yes	1	.6	.6	100.0
	Total	156	92.9	100.0	
Missing	missing	12	7.1		
Total		168	100.0		

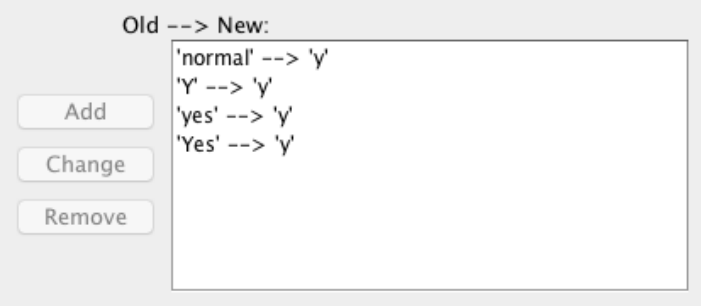
Recoding variables

In PsychoPy, the question was "Normal colour vision?" and the response options were given as `y` and `n`. However, we notice that not all participants have entered the instructed values. (Please note that, like Python, SPSS is case-sensitive, and `y` and `Y` are considered to be different values.) There are similar issues with the other variables. However, for all of the cases it seems that we can easily decode what the participants meant. Therefore, it will be relatively straightforward to correct the values. We will use the `colour_vision` variable to demonstrate how to do this.

- Go to **Transform Recode into Same Variables**.
- Move `colour_vision` to the field "Variables".
- Click on "Old and New Values".



- Enter the "Old Value", then the "New Value", and then click on "Add".
- You should end up with the `Old --> New` field looking like this:



Click "Continue" and "OK". Run **Frequencies** again to check that the recoding has worked:

colour_vision

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	n	5	3.0	3.2	3.2
	y	151	89.9	96.8	100.0
	Total	156	92.9	100.0	
Missing	missing	12	7.1		
Total		168	100.0		

If you want to be on the safe side, go for **Recode into Different Variables**. In this case, SPSS will keep the original variable and you can easily correct errors if something went wrong during recoding. However, the process is a bit more involved. First, you need to provide

SPSS with a new variable name (e.g., `colour_vision_recoded`), then you need to add the old-new changes as just described, but in addition you must tell SPSS to "Copy old value(s)" and click on "Add", so that SPSS displays `ELSE --> Copy` in the `Old --> New` field. You might also need to indicate that "Output variables are strings".

What you could also do is to initially recode into a different variable, check if everything is correct and then delete the original variable. In this way, your total number of variables will not increase and it will be easier to keep track of things.

Descriptives continuous data

Lab class

Calculating descriptive statistics

For continuous variables, there are a number of options, all found under **Analyze Descriptive Statistics**:

- **Frequencies**
- **Descriptives**
- **Explore**

Unfortunately, there is quite some overlap between these options, which can make it hard to remember the differences between them. I spent some time going through these and have summarised commonalities and differences for you. **Frequencies**, **Descriptives** and **Explore** all offer the following:

- Mean
- Standard deviation, variance
- Minimum/maximum, range
- Standard error of the mean
- Kurtosis
- Skewness

Frequencies and **Explore** offer in addition:

- Median
- Histograms
- Percentiles (more options under **Frequencies**)

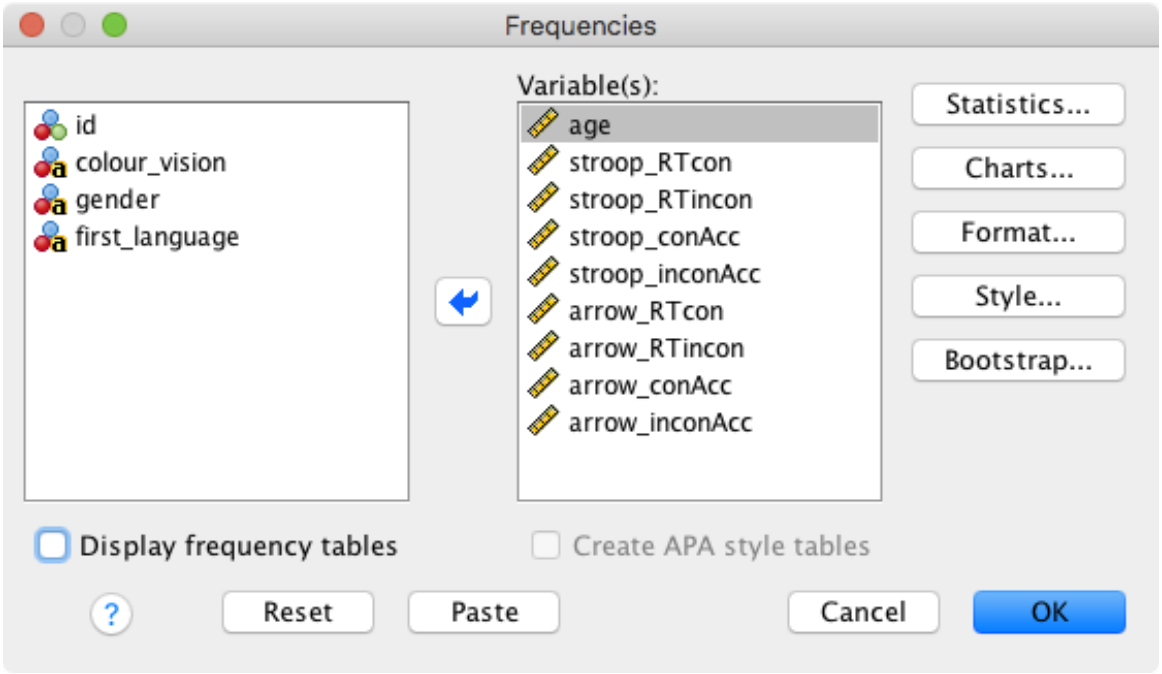
Unique to **Frequencies**:

- Histogram with normal curve overlaid

Unique to **Explore**:

- 95% confidence interval for the mean
- 5% trimmed mean (mean after removing the top 5% and bottom 5% of the values)
- Interquartile range
- Identification of “outliers” (simply the 5 highest and lowest values for each variable)
- Stem-and-leaf plots
- Normality plots (Q-Q plots), including significance tests
- Boxplots

I would encourage you to have a look at **Explore** at some point to see what you can learn from the detailed information that **Explore** provides. However, for our present purposes **Frequencies** is sufficient. We are going to compute frequencies for all of our scale variables:



Here, we have the following options:

- **Statistics**; select these:
  - Mean
  - Median
  - Std. deviation
  - Minimum
  - Maximum
  - S.E. Mean (standard error of the mean)
- **Charts**: Choose "Histograms".
- **Format**: Not currently of interest.
- **Style**: Not currently of interest.
- **Bootstrap**: Not currently of interest.

Finally, you might want to uncheck **Display frequency tables** (as they won't be informative), and click on "OK".

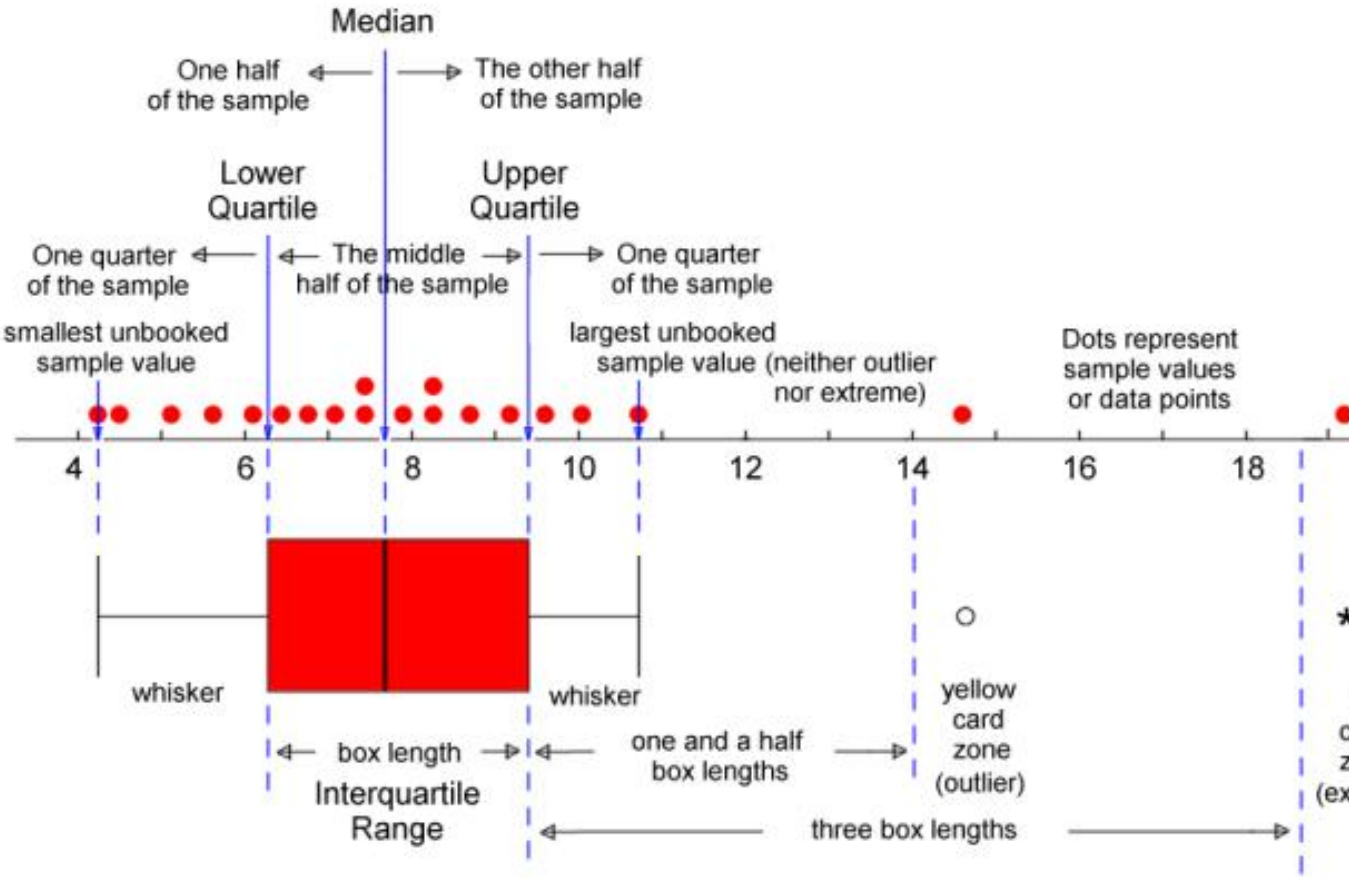
Let's have a look at the output for the arrow flanker task (the Stroop task will later be part of an exercise). Some things to note from the **Frequencies** table and the histograms:

- Accuracy is generally very high (a ceiling effect), resulting in distributions that deviate very clearly from a normal distribution. This very pronounced deviation from normality makes it problematic to run parametric statistics such as *t*-tests on the accuracies (e.g., to test questions such as "Is there a higher error rate in the incongruent condition?")

  - If a statistical test requires normality and your data are not normally distributed, a frequent recommendation is to transform the data. This topic goes beyond what we will cover in our lab class, but Andy Field talks about this in some detail in his chapter "Correcting problems in the data".
  - Please also note that not everyone recommends data transformations though (see "To transform or not to transform…" in the same chapter of Andy Field's book).
  - An alternative approach would be to run a non-parametric statistical test.
- One participant has a very low accuracy (around chance).
  - This might happen for a number of reasons: They misunderstood the instructions, they misremembered the stimulus-response mapping, or they did not pay attention to the task.
  - If a participant's performance is close to chance, it is often better to remove them from the analysis.
- Our raw RTs tend to be positively skewed (i.e., they have a long tail on the right side of the distribution); this is not particularly problematic for two reasons:
  - We have a large sample size; as a result, the sampling distribution of the mean will be normally distributed anyway (look up the *central limit theorem* in the statistics book of your choice).
  - Most inferential statistical tests we will run will investigate the interference effects; these values typically more closely approximate a normal distribution (compute the histograms for the interference scores to check this).

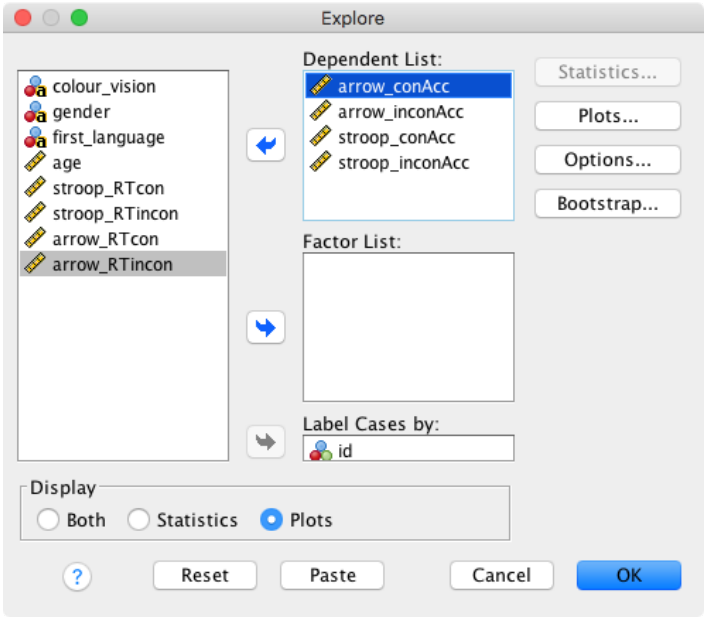
Boxplots

Based on the above results, we conclude that we should have a closer look at the low-performing participants. Let's use *boxplots* (also called *box-and-whisker plots*) for this. Boxplots are often a good way to get a quick overview of potential outliers. The most comprehensive explanation of SPSS boxplots I came across is this:



Please note that other software packages might have different rules for drawing the whiskers and defining outliers/extremes. Also note that SPSS uses the terms “outliers” and “extremes” differently from how I used these terms in the Excel labs.

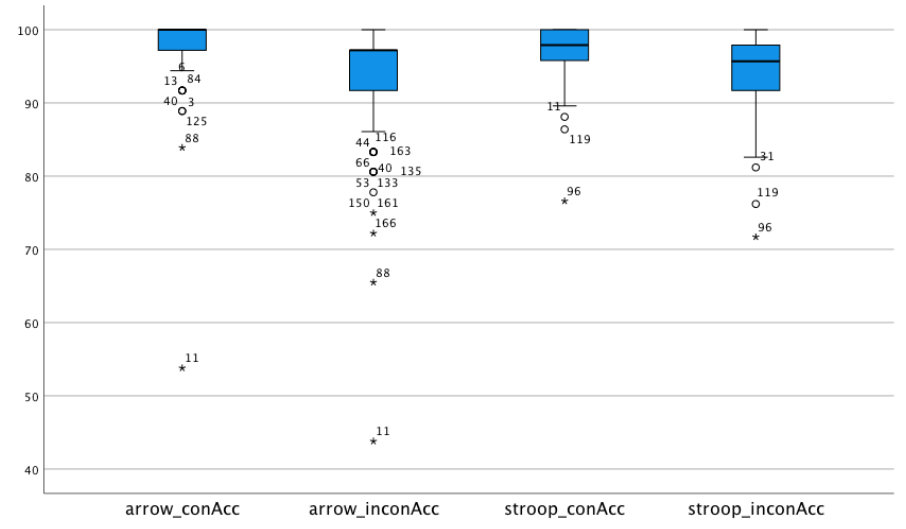
To get the boxplots: **Analyze Descriptive Statistics Explore**. Let’s explore the accuracies in the flanker and the Stroop task. We also ask SPSS to label cases by participant ID (this will only apply to outliers and extremes).



Options

- **Statistics**: Greyed out because we selected Display Plots.
- **Plots**: Only select "Dependents together".
- **Options**: Choose "Exclude cases pairwise".
- **Bootstrap**: Not currently of interest.

Click on **OK** and inspect the output.



For the arrow flanker task, the following participants are identified as extreme: 11, 88, 166, and 161. Apart from participant 11, the accuracies for those are around 70%. I would tend to consider this a low, but still acceptable performance. Participant 11, however, is at chance performance. I would therefore exclude them from any analyses involving the flanker task.

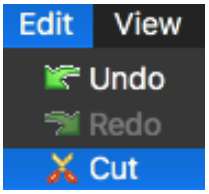
For the Stroop task, participant 96 is identified as extreme. Remember that there were 4 response alternatives for the Stroop task, so chance performance would be 25%. Again, I would tend to keep this participant as their performance is not close to chance.

Now, the question is if we should remove participant 11 from *all* analyses (so, not just those involving the flanker task). Note that they are also marked as an outlier for their accuracy in congruent Stroop trials. Let's have a closer look at their performance in the Stroop task: It turns out that they made more errors in the congruent as compared to the incongruent condition. While they are not the only participant for whom this is the case, they are the participant with the most pronounced difference (the congruent error rate is 9.2% higher). Taken together, I would tend to remove this participant from all analyses due to their unusual behaviour.

Removing participants from SPSS data files

So, how are we going to remove this participants' data from all analyses?

- Save the data file under a new name (e.g. adding `_cleaned` to the file name).
- Click on the row with participant 11's data, such that the row is selected (all columns are highlighted in colour).
- Click on "Edit" and "Cut" to remove the participant.



- Re-save the data.

In this example, we have excluded a participant because we considered them an absolute outlier (i.e., we mainly relied on their *absolute* accuracies to remove them). We have not used statistical criteria (e.g., SD or median absolute deviation) to remove participants. Next week, we will show you how to use SDs for outlier removal.

Removing participants with missing data

We have a number of participants for whom we don't know the age, the gender, the first language, or their colour vision abilities. We might decide that the best approach might be to also remove these participants from all analyses. Here is how to do this:

- Click on **Data** → **Select Cases**.
- Click on "If condition is satisfied".
- Click on "If...".
- Copy and paste the following code into the text field at the top:

```
colour_vision ~= "missing" & gender ~= "missing"  & first_language  ~= "missing" & ~ S
```

- This tells SPSS to only select participants who do *not* have a missing value for any of the above variables. The tilde symbol (~) means NOT.
- Note how the approach differs for string variables (`colour_vision`, `gender` and `first_language`) and numeric variables (age). For the numeric variable, we need to use `SYSMIS` (which stands for system missing value).
- Click on "Continue".
- Select "Filter out unselected cases". This allows you to later remove the filter if you want to exclude a subset of participants only temporarily. To reset the filter, go to **Data** □ **Select Cases** □ select "All cases".

Note that **removing participants with incomplete data might not be necessary**. We have done this here to show you how it works and because there was only a relatively small number of participants affected. Whether or not it will be necessary to remove participants with missing data depends on how relevant knowing these data are for your analysis. For example, for our current example data set and analysis, knowing the age or gender does not appear to be critical.

On the other hand, you might suspect that having a first language other than English or having impaired colour vision are important moderators of Stroop task performance. As a result, you might decide to exclude non-native speakers of English and participants with impaired colour vision from the analysis (and ideally you would back this decision up using published studies). For this reason, you might then also decide to remove participants whose first language or colour vision abilities are unknown.

For an actual research project, you should **think about these exclusion criteria in advance** and ideally **preregister them together with the other details of your study**.

Quiz 3

Quiz details

- Available from Friday, 6 February, 4pm in the "Quizzes and assignments" section on the PSGY1001 Moodle page.
- **Deadline: Friday, 13 February, 4pm**.
- There are 17 questions overall. The first six are about Excel formulas in general. The remaining 11 require you to perform calculations on a data set using Excel.
- The quiz is not time-limited.
- If you have a support plan and need a deadline extension, please fill in the coursework extension form for students with support plans.
- You have one attempt at the quiz and you will not be able to retake the quiz.
- This is one of the three quizzes that together contribute 10% to your overall module mark (see **?@sec-assessment**).

Q: How difficult will this quiz be?

A: Over the past few years, the average quiz mark has ranged between 60 and 66%. For this quiz, it is really important that you follow the instructions closely and use the information provided in Labs 9 and 10!

Use of AI tools

- You will be permitted to use AI tools to revise for the quiz. E.g., an AI tool could explain Excel formulas and functions to you, generate illustrative examples, or generate questions.
- You will not be permitted to use AI tools when completing the quiz.

Explore, apply, reflect

Note that the solutions for the exercises below assume that you have:

- Removed the participant with close to chance performance.
- Removed participants with incomplete data.

If you have completed these steps, there should now be 156 cases in your dataset.

Exercise 1

Recode the variables `gender` and `first_language`.

`gender` should be recoded into `f`, `m` and `x`. `x` was the option for non-binary gender identities.

`first_language` should be recoded into `English` and `other`.

Run **Frequencies** again to check that the recoding was successful.

Show/hide results

colour_vision

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	n	5	3.2	3.2	3.2
	y	151	96.8	96.8	100.0
	Total	156	100.0	100.0	

gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	f	130	83.3	83.3	83.3
	m	25	16.0	16.0	99.4
	x	1	.6	.6	100.0
	Total	156	100.0	100.0	

first_language

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	English	135	86.5	86.5	86.5
	other	21	13.5	13.5	100.0
	Total	156	100.0	100.0	

Exercise 2

Follow the instructions from the chapter Descriptives for continuous data to calculate the descriptive statistics for the RTs and accuracies from the arrow flanker task and the Stroop task.

Show/hide results

		age	stroop_RTcon	stroop_RTincon	stroop_conAcc	stroop_inconAcc	arrow_RTcon	arrow_RTincon	arrow_conAcc	arrow_incc
N	Valid	156	156	156	156	156	156	156	156	
	Missing	0	0	0	0	0	0	0	0	
Mean		18.64	680.42	818.97	97.254	94.054	458.72	507.17	98.462	94.
Std. Error of Mean		.089	11.013	13.741	.2227	.3847	5.570	5.406	.2028	.4
Median		18.00	655.00	776.50	97.900	95.700	442.50	495.00	100.000	97.
Std. Deviation		1.107	137.555	171.628	2.7820	4.8055	69.575	67.518	2.5333	5.8
Minimum		17	481	552	86.4	76.2	353	387	83.9	6
Maximum		24	1468	1417	100.0	100.0	885	836	100.0	10

References