

IMPORTANT NOTICE

This is the first of a series of “Practicals” I created for teaching research methods to masters student when I was a lecturer at the University of Leeds. I’ve adapted them so that they’ll work pretty well as standalone exercises for you to use in this summer school, to teach yourself how to use SPSS.

Throughout the practicals there are references to chapters in the old course textbook I used to use, which you can refer to for more details and advice. An electronic copy of this textbook is available via the University of Edinburgh Library, link below:

Discovering Statistics Using IBM SPSS Statistics / Andy Field (Fifth Edition). (2017). SAGE LTD.

https://discovered.ed.ac.uk/permalink/44UOE_INST/iatqhp/alma9924461179002466

Note: There is a more recent 6th Edition of this textbook (released in 2024). If you access that version you may find the chapter/section reference don’t match up exactly. Although things should be fairly similar.

Practical 1: Management and Manipulation of Data

This practical is intended as an introduction to the SPSS software programme. It we will only cover how to get data into this programme and some basic tools that you can use to describe and explore your data prior to running more substantive statistical analysis.

The relevant chapters for this week in the Andy Field Textbook are:

Chapter 4: The IBM SPSS Statistics environment

Chapter 6: The beast of bias

Section 1: Describe your data


Exercise 1: How to summarise Continuous data

Having got your dataset loaded into SPSS and formatted the way you want it to (i.e. “wrangled” your data)¹, we’ll start looking at how you describe columns within your dataset that contain continuous (scale-type) data.

The first variables we’ll focus on is *Gest_Age_Wks*. This column contains data on children in this sample’s gestational age (in weeks) when they were born. As you may be aware most pregnancies last around 40 weeks, children born around this time are described as being born “at term”. Children can be born much earlier than this though, and those that are born particularly early (before 37 weeks gestational age) are often referred to as being born ‘prematurely’.

¹ **Relevant reading in companion textbook on data importing and saving:** Chapter 4, Sections 4.7 and 4.11, also see Sections 4.5 and 4.6 for info on how to **enter and edit data** into a dataset file.

To explore the gestation age data in this dataset and describe it in terms of measures of both its 'central tendency' and 'dispersion' we can use the following steps: Go to **Analyze > Descriptives > Explore**

1. In the **Dependent list** use the  button to add **Gest_Age_Wks**
2. For now, in the **display** option just select **Statistics**. We'll come on to exploring data using plots in the next section.

You should see the following **Descriptives** table appear in your output:

Descriptives			Statistic	Std. Error
Gestational age in weeks, at birth	Mean		37.35	.273
	95% Confidence Interval for Mean	Lower Bound	36.82	
		Upper Bound	37.89	
	5% Trimmed Mean		37.64	
	Median		40.00	
	Variance		22.423	
	Std. Deviation		4.735	
	Minimum		26	
	Maximum		42	
	Range		16	
	Interquartile Range		8	
	Skewness		-.878	.141
	Kurtosis		-.765	.281

To practice for yourself, repeat these steps to extract a summary of descriptive statistics for the **prosocial_sum** variable. This is a measure from a teacher-completed standardised questionnaire (the Strengths and Difficulties Questionnaire [SDQ]). This subscale from the SDQ rates children's levels of prosocial behaviour on a scale from 0-10.² This, and other, SDQ subscales are often treated as a continuous measure in quantitative analyses.


To give further context, I've included a brief summary of the SDQ on the next page, with some explanation of how it is scored. In these outputs, see if you can work out which of the outputted statistics are measures of the central tendency of the data in this sample and which are measures of dispersion.

Exercise 2: How to summarise Categorical data

The next variable we're going to take a look at is **p_confidence**. This represents the parent's response to a question about their confidence in their parenting, six months after the child was born². They were asked: "Overall as a parent, do you feel that you are..." and were able to make one of three responses: 1 = "Average or Lower"; 2 = "Above Average"; 3 = "A very good parent". Note: we also have some cases where this question was not asked (i.e. the data is missing) and these are coded in the dataset as "-91".

² Note: this is not 'real' data though. This is entirely fabricated data (i.e. not collected from real people), which I created for the purposes of giving us some data to work with in this practical.

This is an example of an ordinal categorical variable. We can describe this variable statistically in terms of count data (i.e. how many of our sample responded in each of the three possible ways listed in the last paragraph?). We can then derive proportions/frequencies from these counts, which are often a bit easier/quicker to interpret. To do this within SPSS:

1. Go to **Analyze > Descriptives > Frequencies**
2. Highlight the **p_confidence** score then using the 'arrow'  button and move it into the **Variable(s)** window, then click the **OK** button.
3. In the output window that will pop up you will get a Frequency table displayed. This should look like the example below:

Parental Levels of confidence, based on self-report

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Average or lower	85	28.3	28.9	28.9
	Above average	142	47.3	48.3	77.2
	A very good parent	67	22.3	22.8	100.0
	Total	294	98.0	100.0	
Missing	-91	6	2.0		
Total		300	100.0		

The **Strengths & Difficulties Questionnaire [SDQ]** (Goodman, 1997) is a widely used, brief assessment of child mental health. There are teacher, parent and self-report versions of this 25 item questionnaire, with each item comprising of a behavioural statement (e.g. My child is '... considerate of other people's feelings'; '... constantly fidgeting or squirming') that is scored by the parent with a response of either 0 ('Not True'), 1 ('Somewhat True') or 2 ('Certainly True').

The questionnaire can be broken down into five subscales (Emotional Problems, Conduct Problems, Hyperactivity, Peer Problems and Prosocial ability), each assessed by five different items. On each of these subscales is scored by summing response values for the five corresponding items, meaning an individual's score on a subscale can range from 0-10. In the dataset you have you only have data for two of these five subscales. In the case of the Prosocial Scale a higher sum score is indicative of an individual being more prosocial (e.g. helpful, sharing), on the peer problems scale a higher sum score is indicative of more difficulties with peers (e.g. bullying, fewer friends). The dataset you are working with comes from a study where all teachers within a large primary school completed an SDQ for consenting children in their class at the end of the school year.

References

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry*, 38(5), 581–586.

What is the most frequent, common, response, amongst those who answered, and how frequent was it?

Above average. 48.3% responded this way (note: look at Valid percent, which discounts for missing responses).

In how many cases is data missing?


6

To practice for yourself, repeat these steps to extract a summary of descriptive statistics for the *Ethnicity* data column, this codes the child's ethnicity into one of five standard categories used in census reporting: 1 = "White", "2 = Asian or British Asian", 3 = "Black, Black British, Caribbean or African", 4 = "Mixed or Multiple Ethnic Groups", 5 = "Other").

Extra Tips:


SPSS only understands categorical variables that are coded numerically (i.e. each category within a variable of this type needs assigned a unique numeric value). You give each code a label in **Variable View** and these are what are viewable in **Data View**. If you want to see the underlying numerical code in Data View you can toggle the labels on and off with this switch:

Value Labels on



	Gender
8	Male
8	Female
5	Male
8	Female
8	Male

Value Labels off



	Gender
8	1.00
8	2.00
5	1.00
8	2.00
8	1.00

The left hand column of this frequency table will lists the categorical response values that exist for this variable (i.e. "White", "Asian or British Asian", etc...). The Frequency column then indicates how many participants fall within each of these categories, with this represented as a percentage further to the right.

Is there something you notice about the response values that seems odd, given what I already told you about how it was categorised?

You should notice that there are response values of -91 here too but in the case of this variable they are NOT being recognised by SPSS as code used to designate a response that is "missing". Instead, SPSS is treating these cases as sixth, valid, response category!?

Using a numerical code (e.g. -91) to log missing data is standard practice in most datasets. Can you think of reasons why retaining information on participants who did not complete all or part of your study is important, rather than just deleting them from your final dataset?

In many studies participant data is not missing at random and there is systematic bias in what is missing. Also, participants may at times provide partial but not fully complete responses and therefore their data may be able to be included in some but not all aspects of a planned analysis. Deleting data from participants who don't complete your study in full is not good practice because it is effectively re-writing history. For example, it will prevent you from analysing demographic differences that may exist between full and partial completers of your study. Imagine you are interested in evaluating the benefits of a new treatment for depression. In such a study it would be important to know if there were systematic differences between the patients who agreed to participate compared to those who didn't and also between those who completed a 6-month intervention period and those who didn't. After all, participants are more likely to try new treatments if their existing treatments are proving ineffective and they're more likely to stick with a trial if they feel that treatment is effective and having benefits!

Getting back to this dataset. Why is -91 a particularly sensible code to use in the context of inputting values for missing data in this dataset?

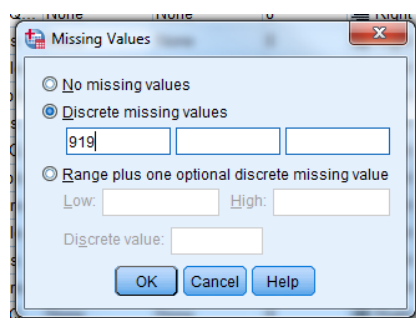
-91 is good as a code because it is not an 'easy' typo. These are not two keys that are close to each other on the keyboard.

It is also a particularly good value to use for the data in this dataset because negative values are implausible/impossible response values for all the variables it includes too.

I have emphasised the importance of noticing the missing data value here because it is important we pay attention to outlying values whilst exploring any dataset, considering all plausible explanations for why they may arise. **This is something you'll look at again in Section 3.**

To round up this part of the exercise

1. Go to **Variable View**
2. Select the **Missing** column for *Ethnicity* and in the **Missing Values** window that pops up click on **Discrete Missing Values** and specify '-91' as a missing value (see window below for example).



One final thing...

You'll notice that SPSS allows you to code missingness in a number of different ways (i.e. you don't need to use only one single value). Why might this be useful? *Hint: What are all the possible reasons you might have missing data?*

You could use different codes to distinguish between data that is missing due to different reasons. For example, using one code could denote participants that failed to understand instructions, another code to denote data lost due to equipment errors corrupting certain files. You could even define 'missing' values as those below a certain threshold, which might indicate making too many errors on a cognitive test, invalidating further interpretation because you doubt the participant has understood the instructions.

After having the inputted Missing Values info into variable view, what do you notice is different when you repeat the steps to create the frequency tables (i.e. **Analyze > Descriptives > Frequencies**)

-91s are now partitioned off in a separate set of rows and denoted as 'missing'. You'll also note you get two '**Total**' rows in each table, one that includes the missing values and one that excludes them. The **Valid Percent** column also excludes the missing values in calculating the proportion of the sample in each of the, genuine, ethnicity categories here.



Exercise 3: How to summarise Combinations of data

Often, we don't just want to summarise each individual variable in our datasets in isolation. Sometimes it is useful (and theoretically interesting) to look at how response on one variable may vary across categories of a second (categorical variable). Here are two examples of that below:

Splitting Gestational Age Descriptive Statistics by Sex at Birth

We are going to apply a split to the Descriptive Statistics we extracted in Exercise 1, calculating these measures of Central Tendency and Dispersion within groups defined by their **Sex_at_birth**. In this dataset, newborns were categorised as either 1 = "Male" or 2 = "Female". To apply this split:

Go back to **Analyze > Descriptives > Explore**

1. In the **Dependent list** use the  button to add **Gest_Age_Wks**
2. In the **Factor list** use the  button to add **Sex_at_birth**
3. Again, in the **display** option, just select **Statistics**, for now.

You should see the a new **Descriptives** table appear in your output, which is broken up into repeating sections (first section reports descriptive stats for the Males in the sample, second for Females).



Look at the mean gestational age values for males and females in this output. Is there a difference here?

Trick question! The answer is we can't now if there's a meaningful difference unless we test whether this difference is statistically significant. Which is something we'll come back and do this afternoon.

Splitting Parental Confidence Descriptive Statistics by Ethnicity

As you might've guessed by now, we're going to go back and split **p_confidence** responses by **ethnicity**. Here's how we'd do that:

1. Go to **Analyze > Descriptives > Crosstabs**

2. Use the  button to slide *p_confidence* in (R)ows:
3. Use the  button to slide *ethnicity* in (C)olumns:
4. To make sure we get some percentages outputted, as well as counts, click on the **Cells...** side menu and within it tick the option for **Column³**, under **Percentages**, and then go back to the main menu by clicking **Continue**.
5. Finally, click the **OK** button to execute.

In the output window that will pop up you will get a Crosstabulation table that reports, in columns, the categorical responses to the Parental Confidence Question within each ethnic group.

Note: Unfortunately, SPSS doesn't let you know how many missing values are present anymore in the Crosstabs, in the way it does when you use the Explore and and Frequencies functions.

To practice for yourself, repeat these final two exercises but this time split *prosocial_sum* by *Sex_at_birth*, using **Explore**, and *ethnicity* by *Sex_at_birth*, using **Crosstabs**.

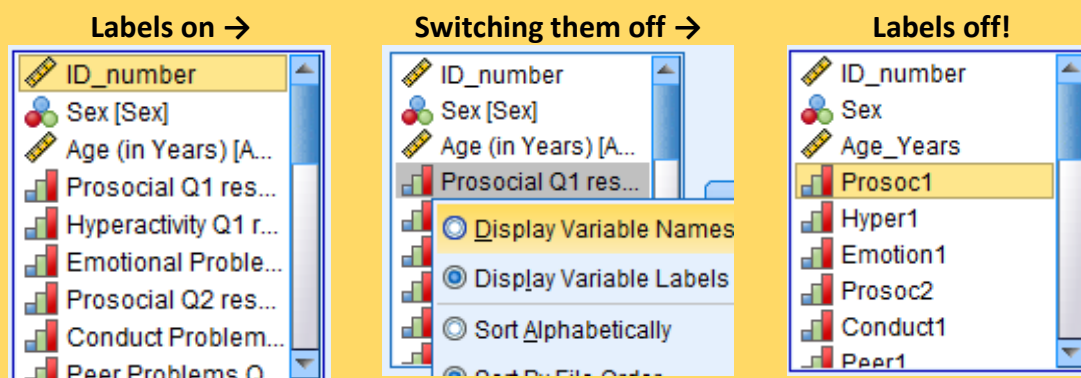
Are there any meaningful group differences in the mean *prosocial_sum* between males and females and are there any groups that are under-represented in the crosstabulation, which you think it might be unwise to try and make general statements about (even after inferential statistical testing)?

Trick question, again! You shouldn't be trying to infer about group differences in prosocial behaviour based on descriptive statistics alone. **Here endeth the lesson that one shouldn't read too much into descriptive statistic on their own!!!**

Beyond the "White" and "Asian and Asian British" groups we have quite small numbers in all the other ethnicity categories, especially when we split by sex (e.g. only two males of 'other' ethnicity). We'll revisit this issue in part 3 of this session but until then, think on why this might be a problem, in terms of generalising from your sample?

Extra Tips:

Often when using SPSS it is simpler to view the variable names rather than their longer labels (which can be a bit text-heavy in the small windows SPSS forces you to work in). To switch to variable names when carrying out an analysis right click on a variable and toggle from **Display variable labels** to **Display variable names** in the options.



³ You could ask for Row and Total percentages to be outputted in this table too but that gets confusing fast. For now, stick with column, which will report the percentage responses to the Parenting Confidence question within each Ethnicity category.

Section 2: Visualise your data

There are many ways you can choose to visualise relationships in your data and I'm only going to be able to scratch the surface all the ways you could think about representing your data visually.

There are also several different reasons why we might want to visualise our data. At the end of a project, when reporting its results it might help with their communication. That's what exercise 1 and 2 will focus on here. In Exercise 3 I'll also teach you about a different type of graph, which is useful to know about to help you explore your data (i.e. to check it over before doing any inferential statistics on it).

Exercise 1: Scatterplots

A scatterplot is appropriate to use if you have a pair of continuous variables that we want to visualise the relationship between. Here, we're going to look at the relationship between participants Prosocial and Peer-Problems sum scores on the SDQ, to produce an appropriate scatterplot of that bivariate relationship:

1. Go to **Graphs > Chart Builder**
2. In the '**Choose From**' menu select the **Scatter/Dot** option, and then click on the **simple scatter** option, which looks like this (Note: icon may vary slightly depending on the exact version of SPSS you're using):



3. From the **Variables:** list drag-and-drop in the *prosocial_sum* variable along the Y-axis and the *peer_probs_sum* variable along the X-axis⁴.
4. Then Click **OK**

To practice for yourself, create two more scatterplots that plot *Gest_Age_Weeks* on the Y-axis and *prosocial_sum* and *peer_probs_sum* on the X-axes, respectively. Imagine that we are at the end of a project and reporting the fact that we've found a statistically significant relationship between Gestational Age and Prosocial Behaviour. Looking at the Scatterplot you've created for this, what more might you want to try and infer about this relationship⁵

Mild positive trend here, where prosocial behaviour is more likely to be higher in those of greater gestational age? It's not a strong relationship though, even though it is significant (more on this in session 3!).

⁴ You could, if you wanted, reverse this order (e.g. *prosocial_sum* on Y and *peer_probs_sum* on X). We would only normally set variables to specific axes in the event we were implying a causal relationship. If that were the case, we'd normally put the predictor on the X axis and the Outcome on the Y axis.

⁵ We CAN infer using descriptive statistics AFTER the results of our inferential tests tell us it is legitimate to interpret patterns we think we're seeing in visualisation of our data, like this scatterplot.

Exercise 2: Bar Graphs

When we have a continuous variable and one (or more) categorical variables we can present these relationships using bar-charts (and clustered bar charts).

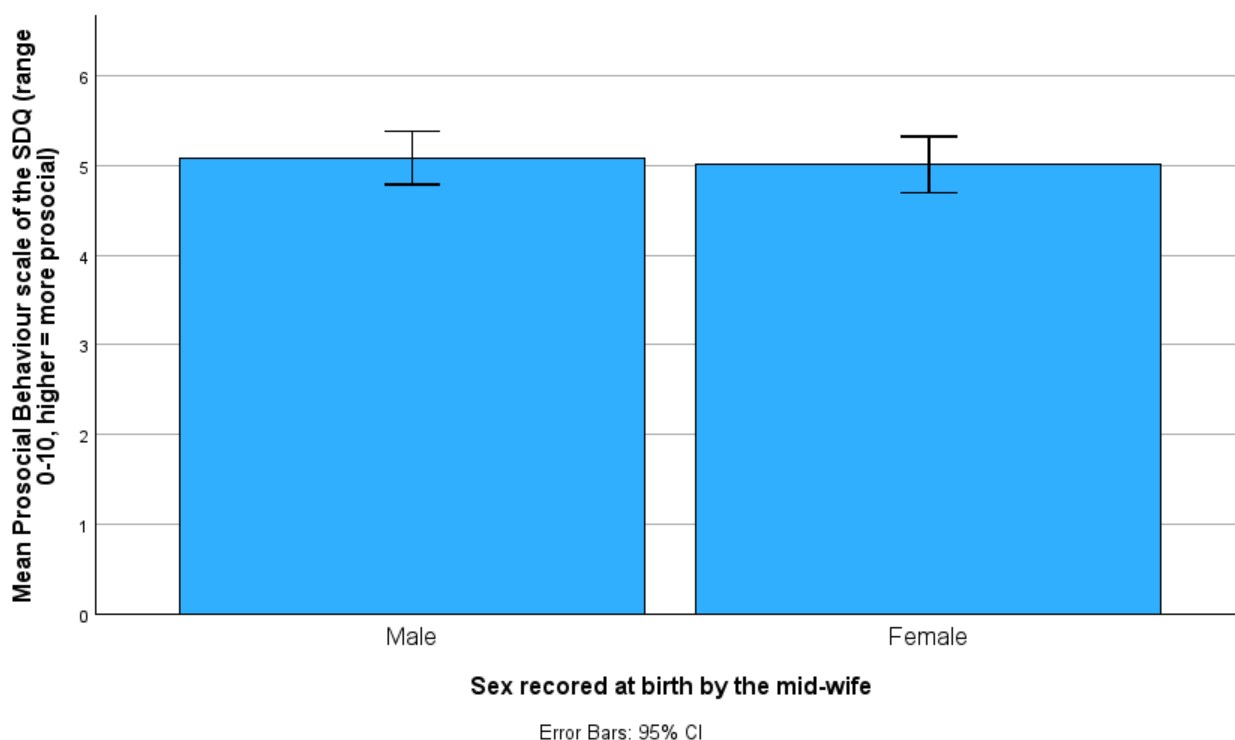
Here's some instructions on how to do that, to look at group differences in Prosocial Behaviour split by Sex and then by Sex and Ethnicity. In the later example, we'll only compare the two (larger) ethnic groups in our sample (i.e. "White" and "Asian or British Asian"). This will involve us learning how to use the select function in SPSS too, to exclude rows temporarily from our dataset (often useful for excluding missing data or focussing in when doing certain sub-group analyses).

Simple Bar Chart (Prosocial Behaviour by Sex at Birth)

To create this graph go to **Graphs > Chart Builder...** and follow these steps:

- In the **Choose from:** option make sure **Bar** is selected in the list to the left
- Double-click on the icon for a **Simple Bar** graph (this is the first one on the top left with three blue columns side-by-side)
- Drag and drop your continuous variable (*prosocial_sum*) onto the Y-axis
- Drag and drop one of your categorical variable (*Sex_at_birth*) on to the X-axis
- In the **Element Properties** tab (on the right) open the drop-down menu underneath the word **Statistic:** and select what you want the height of the bars to represent. Usually, you want to select **Mean** here but you have several measures of central tendency you could choose from.
- By selecting the **Mean**, in the previous step, you should then be allowed to click the option below this to **Display error bars** (also in the **Element Properties** tab). I normally want my error bars to represent 95% confidence intervals but if you want them to instead represent Standard Error you can make this change and decide whether you want to set these at 1 or 2 standard errors in width.
- Lastly, click **OK** to create the graph in your output file

You should get a graph that looks like this:



Note: this is a pretty ugly, poorly formatted graph (unfortunately SPSS is pretty bad for this). In the next example I'll give you some examples of how you can format these graphs further to get them closer to being in APA format.

Clustered Bar Chart (Prosocial Behaviour by Sex at Birth and Ethnicity)

First, for this graph we only want to use a sub-sample of our dataset, who are either designated as 1 = "White" or 2 = "Asian or British Asian" by the Ethnicity variable. To apply this exclusion criteria we use the **select** function, like so:

1. Go to **Data > Select Cases**
2. Tick the option for '**if condition is satisfied**' then click the '**if...**' button
3. In this new window you need to tell SPSS under what circumstances you want a participant's response to be included (not any rows that don't meet this rule will then be excluded *temporarily* from the dataset).
4. In this instance an appropriate command to give SPSS would be:

Ethnicity <= 2

1. To execute this instruction click **Continue** then **OK**

This command instructs SPSS to only retain participants with a Ethnicity value less than (<) or equal to (=) 2. What do you think this will do?

It will only retain participants whose ethnicity is coded as categories 1 (i.e. "White") or 2 (i.e. "British or British Asian"), excluding the other ethnicity categories, which we only have very small numbers of within our sample. We might want to exclude them for fears these numbers within our sample are so small they are unlikely to be a representative sample of these populations (i.e. may be incautious to try and generalise about these groups when numbers in our sample are that small?)

When you go back to the data you'll now see that some have been 'struck out' on the **Data View**, if you look carefully, you'll see these are all cases that have an ethnicity value of 3 or more:

	Child_ID	prosocial_sum	peer_probs_sum	Gest_Age_wks	Sex_at_birth	p_confidence	Ethnicity
1	1	4	5	31	1	3	1
2	2	5	4	30	2	2	3
3	3	8	1	42	2	3	1
4	4	5	4	32	1	2	3
5	5	5	5	42	1	2	1

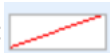
Can you explain why this select cases function is a better way of managing outliers than simply deleting them from your dataset?

Importantly, the cases that I've excluded from the dataset are still there, they're not permanently deleted. So, if I've made a mistake I can go back and fix this. Also, imagine I am in a position to collect more data in a year or two's time, to add to this existing dataset. At that point I might have enough numbers in these smaller ethnic groups to have a representative sample that includes them. In summary, NEVER delete unless you're absolutely sure data needs to be PERMANENTLY removed.

Having selected the rows we want included in our graph, we can create this slightly more complex graph by go to **Graphs > Chart Builder...** and following these steps:

- In the **Choose from:** option make sure **Bar** is selected in the list to the left
- Double-click on the icon for a **Clusteredp10 Bar** graph (this is the first one IN from the top left, with three PAIRS of blue and green columns side-by-side)
- Drag and drop your continuous variable (*prosocial_sum*) onto the Y-axis
- Drag and drop one of your categorical variable (*Sex_at_birth*) on to the X-axis
- Drag and drop your remaining Independent variable (*Ethnicity*) into the **Cluster on X: set color** box.
- Click the option to **Display error bars** in the **Element Properties tab** to the right. I normally want my error bars to represent 95% confidence intervals but if you want them to instead represent Standard Error you can make this change and decide whether you want to set these at 1 or 2 standard errors in width.
- In the **Element Properties tab** (on the right) open the drop-down menu underneath the word **Statistic:** and select what you want the height of the bars to represent. Usually, you want to select **Mean** here but you have several measures of central tendency you could choose from.
- By selecting the **Mean**, in the previous step, you should then be allowed to click the option below this to **Display error bars** (also in the **Element Properties tab**). I normally want my error bars to represent 95% confidence intervals but if you want them to instead represent Standard Error you can make this change and decide whether you want to set these at 1 or 2 standard errors in width.
- Still in **Element Properties**, in the **Edit Properties of:** scroll down menu switch to **GroupColour (Bar1)**. Then, within this side-panel go to the **Order:** box and check that only "White" and "Asian or Asian British" are listed. The other three categories should be in the **Excluded:** box and if they aren't you can drag and drop them over to that box.
- Lastly, click **OK** to create the graph in your output file

As in the previous exercise, you should get another ugly graph outputted, which has a lot of chart clutter on it. To fix that and improve the formatting of this graph...

- Double click on the graph in our output to open it up in the **chart editor** and make the following changes to it:
- Click on the title then press Delete on your keyboard to remove it. This information will be communicated in your Figure caption
- Click on and remove the 'Error Bars: 95% CI' text. Again, **this is information you also should include in your Figure caption**
- Double-click on the horizontal grey lines on the graph and remove them by selecting:  and clicking apply in the properties box that will pop up in response
- Double click on the 'Training Method used' text and edit it down to 'Ethnicity coded...' – this is overly wordy as it stands. Similarly edit down the caption on the X and Y axes (i.e. removing detail that is better given in your figure caption).

This next step is not necessary for APA format but to improve the chart further you could also:

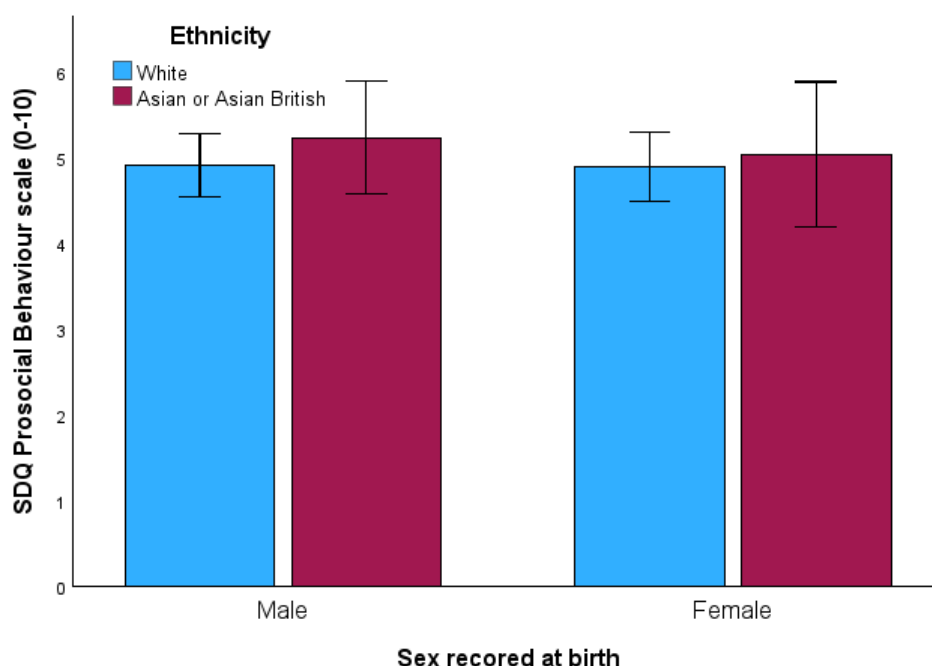
- Click on the legend to highlight it and move it over into the blank space in the top left of your chart area (i.e. compact your chart where possible)

Once you are done in the chart editor go to **File > close** to return to your output file. To export your graph from SPSS, so you can include it in your report, you can either:

- Right-click on the graph, select **Copy As > Image** and paste the copied image into your report
- If that doesn't work, Right-click on the graph, select **Export**, choose a **file name** and destination you want to export the graph to by clicking **Browse**. By default this should be set to save out into a word document, which you can then cut-and-paste the graph from. Make sure the **Document Type** is set for **Word/RTF (*.doc)**.

The final graph should look like the example below and it is **very important** you also include a figure caption (as I have here) directly above it in your write-up:

Figure 1. Mean Differences in Prosocial Behaviour score by Sex at Birth and Ethnicity. Error bars represent 95% confidence intervals



Exercise 3: Boxplots

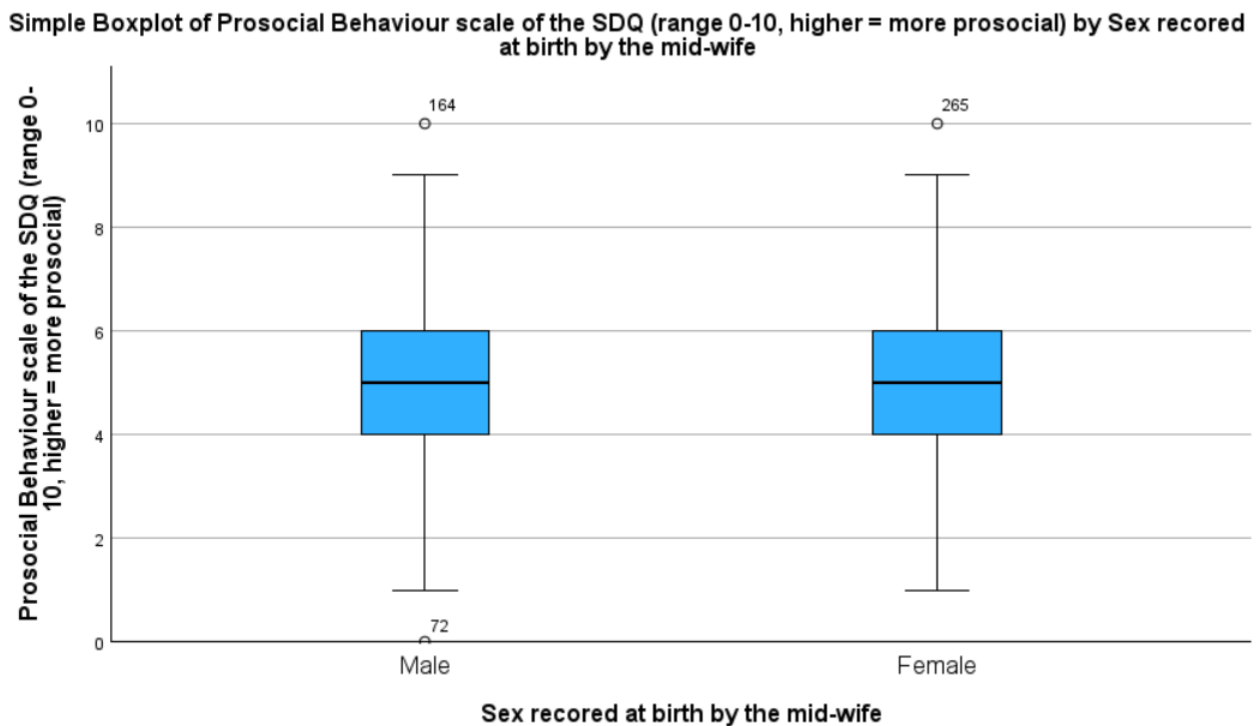
When exploring data at the beginning of a study, boxplots can be a useful visualisation of the central-tendency and dispersion of continuous variables, in the sample as a whole or within levels of a categorical variable. They are particularly helpful to us when we want to spot extreme values in these distributions, which we might want to know about before we go further on in our analysis.

Asking for a boxplot is very similar to asking for a bar chart (see Exercise 2 in this Section). So, if we instead wanted a boxplot of participant's Prosocial Behaviour by Sex category, here's how we'd request that output from SPSS:

To create this graph go to **Graphs > Chart Builder...** and follow these steps:

- In the **Choose from:** option make sure **Boxplot** is selected in the list to the left
- Double-click on the icon for a **Simple Boxplot** graph (this is the first one on the left of the three options)
- Drag and drop your continuous variable (*prosocial_sum*) onto the Y-axis
- Drag and drop one of your categorical variable (*Sex_at_birth*) on to the X-axis

You should get a graph that looks like this:



Note: Numbered values that fall outside the 'whiskers' of the boxplot are evaluated as SPSS as being outliers (i.e. small numbers of cases that are very distant [and thus unrepresentative] of the mean). The number is the row within your SPSS dataset that this case sits upon.

For example, go to row 164 in your dataset, in Data View, and you'll see this case has a prosocial sum score of 10, as indicated in the boxplot:

	Child_I D	prosoci al_sum	peer_p robs_s um	Gest_A ge_wk s	Sex_at _birth	p_confi dence	Ethnicit y
164	164	10	0	42	1	3	1

Section 3: Exploring your data

IMPORTANT: For this last set of exercises I would like you to use a different dataset from the one we've been using up until now. Please click on the link to the following (slightly changed) SPSS dataset on the website: **session_1_data_Ex3.sav**.

Before we move on from discussing Descriptive Statistics, to discussing Inferential Statistics, we need to consider our sample's representativeness and **whether there might be any features of our data** (summarised using the descriptive techniques we've learned about this morning) **that violate common assumptions required for statistical inference testing**. To explain further, most inferential statistical tests will run just fine irrespective of whether the data you feed into it violates its assumptions or not. The big problem is, data that violates a test assumption is MUCH more likely to lead to biased estimates and results being generated from that model, which may result in mis-leading and unreliable results from it. **In summary, if you've run statistical analysis on data that you know violates the assumptions of that analysis then you should make the reader aware of this when reporting your results and discuss the implications of this in your dissertation** (i.e. you need to be *even more* cautious with your interpretations than usual).

Going to use what we've learned already to explore our data, looking for three specific things.

Note: what assumptions you need to check will in part depend on the specific statistical tests you go on to use, below are just three very common issues, that often present problems and need thought about when preparing any dataset:

Exercise 1: Exploring Extreme Values

Let's revisit the outliers we found for Prosocial Behaviour when creating boxplots for the first time in Section 3 Exercise 3 (see page 13).

To recreate this boxplot in this new dataset:

go to **Graphs > Chart Builder...** and follow these steps:

- In the **Choose from:** option make sure **Boxplot** is selected in the list to the left
- Double-click on the icon for a **Simple Boxplot** graph (this is the first one on the left of the three options)
- Drag and drop your continuous variable (**prosocial_sum**) onto the Y-axis
- Drag and drop one of your categorical variable (**Sex_at_birth**) on to the X-axis

The graph should look the same as it did in the previous dataset because, in the case of these two variables, the same data is identical.

Ask yourself how you might want to handle the three outliers (72, 164 and 265), would you want to keep or exclude them from further analysis, and how would you justify this decision?

To practice for yourself, swap out **prosocial_sum** and replace it with **peer_probs_sum** in the boxplot. Are there any outliers in this visualisation of your data and why might you want to handle them differently?

I would exclude case 43 from any analysis involving the Peer Problems scale because its response value for Peer Problems is implausible (i.e. 100 > than the maximum value this scale should go up to). If I had more background information I might want to see if I can verify if this is a typo but given I can't tell whether this is a mistyped 1, 10 or 0 I need to act cautiously and just exclude it.

Cases 72, 164 and 265 are different, they are less comment but plausible response values in terms of their Prosocial behaviour score. I would definitely want to retain these (even though they are outliers). If I was being super-cautious I could re-run my analysis several times with and without them included in an approach often referred to as an 'outlier sensitivity analysis'.

Further advice when doing a sensitivity analysis: If you find broadly similar results with or without exclusion, I would tend to report the set of results with the outliers included but note in my write-up that the dataset contains outliers that do not substantially affect the model. You can always report the model with them excluded in your supplementary materials (for the purposes of full disclosure) too.

If the results conflict depending on whether outliers are present, I would report both sets of results and discuss this conflict in your discussion – in reality, you have to be very careful and cautious in reporting such inconsistency. Often the only option is to run a follow-up study that tries to investigate the discrepancy (e.g. studying more closely a subgroup of the population you're interested in, which might be less numerous and respond in a different way from the majority, in the context of the topic your studying).


Exercise 2: Exploring Skewness in Continuous Variables

Skew is when a dataset's distribution is not symmetrically distributed around its centre-point. Skew can either be positive (a longer 'tail' of values above the centre point, which pulls this estimate up) or a strong negative (a longer 'tail' of values below the centre point, which pulls this estimate down) in its directions.

It doesn't matter whether it's positive or negative both can be issues for inferential tests that assume a 'normal' parametric distribution within your variables (Note: not all test have this assumption though).

There are three methods we can use to identify skew. Here we'll focus on visually inspecting for skew but we could also interpret the skew value (reported when you use the **Explore** function) or you could formally test your variables for a non-normal distribution using specific tests.⁶

To illustrate this, let's **Explore** the *Gest_Age_Wks* variable, this time using plots rather than statistics. Do this by executing the following steps:

1. Go to **Analyze > Descriptives > Explore**
2. In the **Dependent list** use the  button to add *Gest_Age_Wks*
3. Now, in the **display** option select **Both**. So the output will now include the regular statistics we had previously plus some extra **Plots** (AKA graphs).

⁶ Lots of people use tests like the Shapiro-Wilks or Kolmogorov-Smirnov test but there are reservations over their validity, particularly when working with large samples.

4. Open up the the **Plots** side-menu and within it untick **Stem-and-Leaf** and instead tick **Histogram**.
 - *Note:* There is an option here to also request normality plots and tests but as I've mentioned already, I'm not personally a fan of these tests, so I'll leave this option unticked – you can tick it if you want though!
5. Then click **continue** to close the submenu and **OK** to get the output.

You'll notice the data is very skewed here – do you think this is a problem/issue?

The answer here is nuanced. No, there's nothing wrong here in the sense that this data is in line with what we'd expect in the target population (i.e. most births occur at around 40 weeks, give or take a few weeks). Births much earlier than that are much less common and, to be honest, would probably be even less frequent were this real (not fake) data you were analysing.

The issue remains that this skew may violate the assumptions of the statistical tests you go on to use. There are various ways we could try and handle it (e.g. transforming the data to normalise its distribution, using a matched-samples approach) but which would make most sense would really depend on the project and the specific data we were working with, that's where your supervisor would have to provide guidance.

To practice for yourself, explore the distribution of the [peer_probs_score](#) too. Are there any issues here and how might you fix them?

Whether there are any issues with skew in this variable will depend on whether or not you've already excluded case 43 as an outlier (i.e. using the **Select Cases**) function. If you exclude the outlier you'll find this is a reasonably normally distributed variable, if you haven't, you'll see an extremely skewed plot.

It may be interesting to you to play around and see what impact excluding the outlier has on the descriptive statistics here (e.g. how does the mean and confidence interval change when you exclude case 43)?

Exercise 3: Exploring low cell counts for Categorical Variables

I've already hinted at the fact that, when looking at purely categorical data, it can be a particular problem when cell-counts for certain combinations are very low (close to zero).

In fact, when using chi-squared tests (as we will later today) it can be a problem if there are less than five observations in a cell, when running a **crosstab** (function introduced in Section 1, Exercise 3, bottom of p6). This violates an assumption certain types of chi-squared tests have (more on this later).

Looking back at your **crosstabs** for [ethnicity](#) by [Sex_at_birth](#), are there any cells within this crosstab that violate this assumption?

Both observations of males and females in the “other” ethnicity category are below 5. In other words, there too little data here to judge well what the true distribution of Sex in the wider population would be, specifically for the “Other” ethnicity group. Making any claims that this distribution may differ significantly from the distributions seen in the other groups specious.

More generally, as I’ve discussed already, I think you’d agree that trying to make inferences whether distributions of one categorical variable differ depending on which level of another categorical variable a participant falls within are pretty questionable when you’ve got little to no data on these distributions, because you’ve very few participants to represent these distributions. That gets us on to thinking about ‘sampling’ though, which is definitely a conversation for later (i.e. session 2).