

An Introduction to Correlation

In these exercises we're going to learn more about different types of correlational analyses, which seek to estimate the extent to which pairs of continuous variables are linearly associated with each other¹.

In these exercises we'll also learn about some functions we can use to standardise our variables and create new variables (using the Compute function). We will also get some more practice at exploring our data - think of this as an opportunity to practice doing "assumption checks" for yourself!

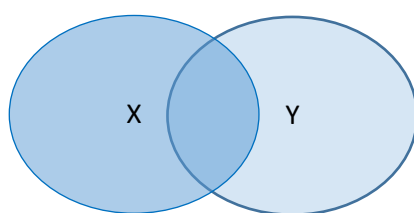
Relevant chapters for this week in the Andy Field Textbook are:

Chapter 6: The beast of bias

Chapter 8: Correlation

Exercise 1: Correlation

An approach commonly used when exploring linear relationships between variables is to calculate a correlation coefficient. Note, this type of **correlational analysis does not allow us to infer causality² between the variables involved**. Instead, correlational analysis will tell us how much response on a pair of variables 'co-vary'. In other words, it estimates the proportion of overall variability across two variables that is shared between them. This is analogous to measuring the overlapping section between the two circles in the illustration below, which represents a proportion of the total area of **both** circles:



If X represents how much people vary in their response to one variable (e.g. how much people within a sample vary in terms of how depressed they report feeling) and Y represents the same property for another variable (e.g. variation in how anxious they report feeling), then the overlap represents the extent to which knowing about variation in one of these things can help us understand variation in the other. The more overlap the stronger the correlation and the more useful it is to know about someone's level of Depression in helping us understand their level of Anxiety, and vice versa. Want a more technical and detailed explanation? See: **Andy Field's Textbook Chapter 8, Sections 8.2.1.**

At the top of the next page there is a description of a (fictitious) observational study we'll be using as an example dataset.

The data for this exercise can be found in the file **session_3_data_Ex1.sav**. Please open it in SPSS.

¹ The parameter these tests try to estimate is the rate at which the value of Variable Y will change in response to changes in Variable X. In other words, if Variable X increases by 1 unit to what degree would we also expect Variable Y to increase or decrease in value?

² Causality is the proposition that one variable acts mechanistically to directly influence (i.e. cause) a response in another variable.

An assessment centre for the College of Policing is interested in the relationships that may exist between the various assessments they give to applicants when they apply. Applicants complete a face-to-face interview, a physical fitness test and two exams (one on their numerical skills and one on their written language skills). The scores on these assessments are then considered by a recruitment panel who then decide whether to offer an applicant a place. The panel is worried though that they may be over-assessing. Are some assessments so similar in what they measure that it's unnecessary to run both of them? They would like to develop more standardised and transparent admissions processes based on using composite, combined scores, derived from all assessments. The panel would then only consider applicants who score above a certain threshold on this overall grading scale.

Task 1: Checking assumptions

In Task 2 we're going to run a type of correlational analysis known as a Pearson's r , to give us some correlation coefficients for relationships between the different assessments in this Police College Exam. Prior to running a Pearson's correlation analysis though, we must first check that the data we're using satisfies the assumptions for running this type of analysis³. These are listed out below, in red:


- **Outcomes are measured on a continuous (interval) scale.** For which of the three assessments does this not apply?

The written language skills assessment is a dichotomous categorical variable (i.e. you can either Pass or Fail). So, it is not suitable to use a Pearson's correlation to test for relationships between this and the other assessments.

- **That each of the variables involved has a normal distribution⁴** (if we want to infer significance).

-

To test this assumption we need to explore the data. Checking for a sufficiently 'normal' looking distribution within each of the continuous variables. To do this:

1. Select **Analyze > Descriptives > Explore...**
2. Select the three continuous variables you are interested in correlating and move all over using the  button into the **Dependent List**

³ You will recall that we've talked about why it is important to check assumptions for statistical tests in several of the previous sessions.

⁴ A normal distribution is the "bell-shaped curve distribution" we discussed a lot in the previous session, when talking about sampling distributions. It's also known as a Gaussian distribution. Note, here it is the symmetrical shape of the distribution of observations of your *variable* we care about, not whether these observations are centred on 0 (i.e. it is the shape of the curve not its position on the x-axis that defines a normal distribution). Whether the distribution centres on 0 is something the correlation will test. That is, can we reject the null hypothesis that data is normally distributed around zero?

3. Click on **Plots** to open another pop-up window and in this window check **Histogram** and **Normality plots with tests**, close this pop-up by then clicking **Continue**
4. To run these explorations of your data click **OK**

As you'll recall from this morning, there's a lot of output. You only need to focus on the parts listed below for this exercise though...

Tests of normality

- Formal significance tests for normality are provided by the Kolmogorov-Smirnoff (K-S) test and the Shapiro-Wilk test. Sample size can influence the likelihood of achieving a statistically significant result in both of these tests, with small deviations from normality being picked up as significant when the sample size is very large and quite noticeable deviations going undetected in smaller samples (in other words, these tests can fall foul of both statistical under- and over-powering of your sample). **For these reasons I would always favour looking at the histograms and Q-Q plots when judging normality.** As a rule of thumb I'd want to see a clear and consistent suggestion from both these tests that there is non-normality (i.e. $p < .05$) **and** further evidence from the plots before defining a given variable as non-normally distributed. For further criticism of the reliability of these tests see: **Andy Field's Textbook Chapter 6, Sections 6.10**

Histograms and QQ-plots

- These are two methods of visualising the data in your variables, which are useful for judging the normality of its distribution.

When looking at **histograms** you want to check that the observations conform (approximately) to the 'bell-shaped' curve seen in a normal distribution. If observations are not normally distributed values will likely either be bunched more to one end of the graph (skewed), have two or more peaks (bi-modal distribution) or have a single, centrally located peak but with this being overly 'peaked' or 'flattened' compared to how we'd expect a normal distribution to look (this is called Kurtosis).

For **Q-Q plots** (specifically the 'normal' rather than 'detrended' graph) normality is indicated by observations all typically falling on or near to the diagonal line. When values consistently fall above or below the line this suggests kurtosis. When points make a 'snaking' pattern back and forth across the diagonal line as it rises this is indicative of skew.

What is your assessment of the distribution of each of the continuous variables and would you consider all of them suitable for analysis using a Pearson's r correlation test?

Of the three relevant variables here (*math*, *physical* and *interview*) I'd only be majorly concerned by the (Face-to-Face) *interview* variable's distribution. This looks skewed in the histogram and there's obvious 'snaking' on the Q-Q plot, as well as by both normality tests signalling concerns (i.e. significant results in what is, broadly speaking, quite a small sample).

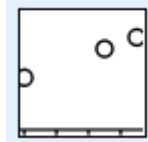
For the remaining variables, the Physical Fitness test's KS normality test is significant BUT the Shapiro-Wilk isn't and visual inspection of that variables suggests the data looks relatively normal in its distribution. You might possibly argue there's some 'snaking' on the Q-Q plot for the math test too but again, on other graphs this data looks reasonable.

Before we move on, we also have one more assumption we need to check...

- *Linear relationships between variables seem plausible*

To check this assumption we'll need to explore the pairs of variables we're interested in, using a series of scatterplots, which we also learned about this morning. To produce an appropriate scatterplot for the *Physical* vs. *Maths* tests comparison:

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:



3. From the **Variables:** list drag-and-drop in the *physical* variable along the X-axis and the *math* variable along the Y-axis⁵.
4. Then Click **OK**

Now create equivalent scatterplots for the other linear relationships you wish to test, and assess whether the plots suggest a linear-relationships is plausible in all of these cases (i.e. *physical & interview*, *interview & math*). Are there grounds, given what you see in the scatterplots, to believe it is worth testing for linear relationships in all three cases?

The scatterplots suggest a moderately strong *positive* linear relationship is definitely plausible and worth testing for between the Maths and Physical test. There is general trend for scores to increase on the Y axis as they also increase along the X axis.

For the relationships between the Interview and other two tests these look to be on the less plausible side of things, particularly for the Maths and Interview measures.

Note: whilst you may have concerns over the interview scores' normality, this doesn't actually preclude analysing linear relationships between it and math or physical fitness. Its non-normality only precludes interpretation of any test for statistical significance SPSS runs for you whilst calculating the Pearson's correlation. In fact, there are other options for doing correlations (besides Pearson's) that do still allow you to test significance, that you can opt for when the assumption of normality is violated.

Task 2 (Running Correlational Analyses)

Having considered whether our assumptions are met we're now going to run six bivariate (bivariate meaning: "2-variable") correlations. These will allow us to investigate how plausible it is that linear relationships exist between the variables in our dataset. To perform this analysis:

1. In SPSS select **Analyze > Correlate > Bivariate**
2. Select the three variables we're interested in correlating with one another (i.e. *interview*, *physical* and *math*) and move them into the **Variables** box.

⁵ You could, if you wanted, reverse this order (e.g. *math* on Y and *physical* 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.

3. Check that the option for **Pearson** and **Two-tailed**⁶ are both checked
4. Also tick the options for **Kendall's tau-b** and **Spearman**, so that SPSS gives us these alternative correlation tests. These alternatives are more suitable when assumptions of normality are violated. Why might we be more interested in these results for some of the bivariate correlations we're running?

Spearman and Kendall's tau correlation methods work on ranking data across variables and seeing how closely rankings match, rather than calculating the co-variation coefficient.

As such, **they do not require a normal distribution**. So, whilst they are more conservative in estimating the strength of a correlation, they are also more suitable methods for testing the statistical significance of the relationship between the (skewed) interview data variable and the other two variables.

5. Open the **Options** sub-window and also ask for **Means and Standard Deviations**
6. Then click **Continue** and **OK** to run these tests

Your SPSS output will contain some **Descriptive Statistics** for each variable and **Correlations** tables (often referred to as a 'correlation matrix') that report on the analysis having been repeated three times – once per 'type' of correlation calculated (i.e. Pearson, Kendall's and Spearman).

Summarise your output for the **Pearson** correlations in the APA formatted table below. The 2nd row has been completed for you as an example:

Table 1.

Descriptive statistics and Pearson correlations between application centre assessments (N=31).

Measure	Mean	SD	1	2	3
1. Interview Score	6.42	2.49	-	.278	.546**
2. Physical Assessment	6.21	0.56		-	.566**
3. Numeracy test	51.39	18.16			-

* $p < .05$, ** $p < .01$

Extra Tips:

In this table, note that we don't place a 0 before the decimal point for the correlation coefficient and significance (p) values. We do for the Mean and Standard Deviation though. This is because in APA style we denote proportional values (i.e. values that can't be greater than 1) by not placing a 0 in front of the decimal place. Putting a zero implies these values could under some circumstances be larger than zero.

Why do you think it is that in this APA formatted table we only fill in values on one side of the table (i.e. above the diagonal, marked with the '-' symbol)?

⁶ Always default to using 2-tailed, as this is a more conservative test of statistical significance

This cuts out unnecessary repetition we get in the SPSS output. The correlation between Interview Score and the Numerical exam (top right in the Correlation table in SPSS) is the inverse of the correlation between the Numerical Exam and Interview Score bottom left). So reporting them both is redundant.

Based on the results in the table:

- **Ignoring violated assumptions momentarily**, what do the size of the correlation coefficients (i.e. our parameter estimates) and their significance values tell us about the extent to which performance on each variable is related to the others?

The size of the correlation coefficient for the Numeracy test with both the other assessments is quite large and statistically significant for both. The 'effect size', which is one way to describe r can be interpreted as 'large' based on the cut-off points suggested by Cohen (i.e. $>.5$) but only just so.

To me, I'd be interpreting this as evidence that whilst some assessments do require the individual to tap into some shared/similar underlying abilities (represented by quite a large proportion of shared variation) each assessment also has quite a large chunk of its variability left unassociated with the other variables. Thus, it would be useful and informative to consider retaining all of them in the recruitment process.

- Are there any of these relationships you would prefer to analyse and report using the other types of correlation we requested (i.e. Spearman or Kendall's tau)?

The two correlations involving Interview rating (given the answer you've hopefully given at the bottom of page 3) are probably better investigated "non-parametrically" due to this variable being skewed. **Notes:** non-parametric tests don't require the assumption of normality to be met. Also, given the small sample size I'd also probably opt for Kendall's tau, as opposed to Spearman's, in this instance.

- Lastly, there is one of the four assessments we chose not to include in our correlation table. Why? What type of correlation might be appropriate to investigate relationships between this variable and the other three? Hint: see **Andy Field Chapter 8, Sections 8.4.5**

You could use a biserial correlation to look at relationships between the categorical grade applicants receive for their Written test performance (i.e. pass vs. fail) and the other measures). You could do this because this variable is biserial in that it is (presumably) created from applying a cut-off to a continuous variables. By that I mean people could pass by a lot or pass by a little and in that sense it is not a truly dichotomous categorical variable (i.e. it is not what would be defined as 'point biserial'). If I wanted to calculate a biserial correlation this cannot be done easily in SPSS but there are other statistical packages that can be used.

Task 3 (optional): Standardising and Weighting Scores

Before we leave this dataset behind we're briefly going to cover:

- 1) How to manipulate/transform some of the variables in this dataset to allow us to make direct comparisons, within participants, for their performance on the different variables. Yes, this is possible in spite of them being measured on different scales originally!

- 2) How to combine scores on individual assessments in a more sophisticated way to arrive at an overall (weighted average) score.

If the Policing College wished to create a transparent selection procedure, which in some way combined how individuals performed on these four assessments into an overall score, this would presently be quite difficult. Assuming we did not make any further improvements/alterations to the dataset what problems would you foresee in doing this? *Hint: why would calculating an 'average' score not be possible?*

The problem with the 4 variables, as they currently stand, is that we cannot make direct comparisons between them because they are measured on very different scales (e.g. some are continuous, some are categorical and even between continuous variables there are differences in the size of the increments used in each scale).

For example, is a participant who scores 6.4 on their Physical test and 6.4 on their interview test equally as fit as they are adept under interview? Perhaps they are better in one of these domains than the other but it is easier to 'score points' on one of these two scales! This confounds direct comparison. It is for these reasons that we often standardise individual participant responses to variables, relative to the responses made by the rest of the sample. Doing this prior to analyses allows us to make direct comparisons across variables more easily.

Standardising Continuous Variables

To compare participant's responses across a set of continuous variables that are measured on different scales we need to first transform all variables so they are measured on the same scale. In statistics, when variables are normally distributed, the most common approach to doing this is to convert raw-scores into **z-scores**. The z-score is a 'special case' of the normal distribution, with a mean of 0 and a standard deviation of 1.

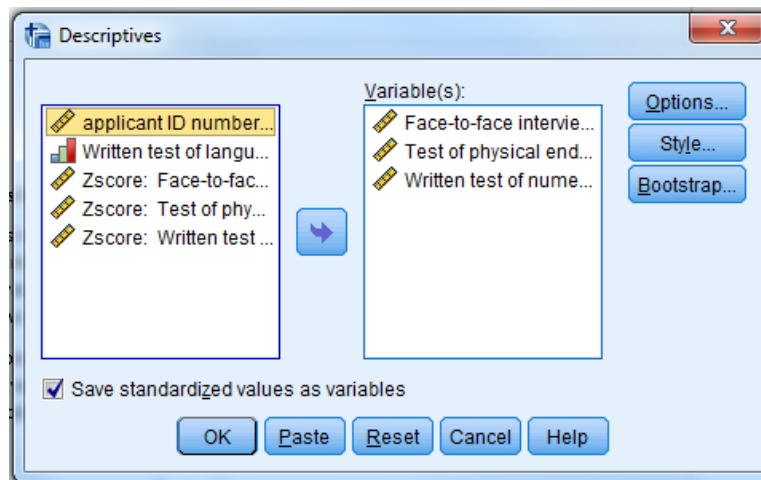
You can convert the scale of any continuous variable from its natural units (e.g. cm, kg, percentage points) into a z-score in SPSS. It does this by creating a new variable in which every participant's value for the original (i.e. 'raw-score') has had the sample mean (M) subtracted from it, before dividing that value by the sample standard deviation (SD) for that variable too:

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

- To create z-scores for the three continuous variables we have (i.e. *Interview*, *Physical*, and *Math*) select **Analyse > Descriptive Statistics > Descriptives**
- In the window that pops up (see top of next page) move these three variables into the **Variable(s)** box and ensure the **Save standardized values as variables** option is ticked before pressing **OK**.
- Doing this will create z-score transformed versions of the variables you've selected in your dataset (i.e. check your **Data** and **Variable View** in SPSS and you'll see we now have 3 extra variables added to our dataset (*Zinterview*, *Zphysical* and *Zmath*).



- Use the Explore function to look at the data for your new z-score transformed variables. What has changed and what is still the same about these variables after this rescaling into standardised z-scores?

All three variables now have a mean of 0 and a standard deviation of 1 (i.e. they're measured on comparable scales). The relative rank position of each individual on those scales though has remained unaltered (i.e. the distribution of responses within each variable and thus the shape of the distribution).

- In the case of the first candidate (i.e. id = 1), compare their performance on the z-scored mathematics and physical assessments. Which do they perform relatively better on? Similarly, answer this question for the final candidate (i.e. id= 31).

Case 1 is better on Math than Physical and 31 is better on physical than math. Be aware that the mean for each of these variables is 0 so scores that are positive and larger in value reflect better performance but scores that are negative and larger in value are worse (i.e. 'more negative')

Important Note: for the purposes of a later exercise we have z-score transformed the *Interview* variable but why is this arguably less valid than z-score transforming either the Numeracy or Physical assessments? Also, are there any steps we could have taken before z-score transforming this the Interview variable to address any concerns we might have (see **Andy Field Chapter 6, Sections 6.12.4** for ideas)?

The interview score is not normally distributed and z-scoring assumes an approximately normal distribution. To enforce a more normal distribution on the interview data (before z-scoring it) we could have tried transforming the data (as described in the section of the textbook highlighted) to reduce skew in this data. For example, a square-root, log or reciprocal transformation could've been experimented with to normalise this variable's distribution.

Using categorical variables to exclude cases and editing filters to create new Categorical Variables

Let's presume that the Police College automatically reject any applicants that fail the written assessment.

- To exclude cases from our dataset where participants have failed this test use the **Select Cases** function we discussed this morning. **Reminder:** use the function to 'Filter out unselected cases' using the '*If condition satisfied*' command to specify that for the categorical variable *written* you only want to select cases where a 'Pass' has been achieved.
- In Data View rename the **filter_\$** variable this creates as *AutoRej* and relabel it as 'Automatic Rejection based on failing written language assessment'

Lastly, the college wants the admissions panel to only consider applicants whose performance, after combining scores on the three remaining assessments, is classifiable as "above average". Combining scores across these assessments is now feasible because we've converted them all to the same standardised scale (i.e. we have a z-score version of each variable).

However, in calculating this combined score the college view some of the assessments as more important than others and want this adjusted for in the averaging process. Specifically, they view the interview as being twice as important as the other two assessment. To do this we need to calculate a **weighted average**, which adjusts for the relative contribution different assessment's z-scores make to the overall average. To do this use we can use SPSS's **compute variable** function to create a new variable from responses to our existing ones by following these instructions:

- Select **Transform > Compute Variable**
- Give the **Target variable** you're creating a name (e.g. *AssessmentScore*)
- In the **Numeric Expression** box enter the following formula then click **OK** to create the new variable:

$(Z_{\text{interview}} * 0.5) + (Z_{\text{physical}} * 0.25) + (Z_{\text{math}} * 0.25)$

Note: This formula tells SPSS the weights (highlighted in red) that we want to apply to each of the variables in calculating the Assessment Score. *Zinterview* score makes a 0.5 (i.e. 50%) contribution, whilst *Zphysical* and *Zmath* each make smaller equal contribution of 0.25 (25%). Thus *Zinterview* is 'double weighted' compared to the other two component variables (i.e. counts for twice as much as any one other component). If using this method to apply weights to variables before combining them in the future then you must make sure that the overall sum of your weightings adds up to 1 (i.e. $0.5 + 0.25 + 0.25 = 1$ [i.e. 100%]). Obviously, the variables you're combining must also be measured on the same standardised scale too!

As a conclusion to this exercise can you tell me how many candidates the admissions panel would review if they accepted all candidates who (1) passed the written assessment and (2) scored above the median on the overall assessment score?

Hint: You'll need to first checking the median value for *AssessmentScore* using the **Explore** function but then you could run check 1 and 2 by hand OR challenge yourself to use **Compute Variable** to create a new categorical variable based on these criteria!

10 individuals would be selected.

This extra exercise is meant to illustrate the power of the Compute function, as a tool you can use to derive new variables from your existing data. It is also a much more reliable method of doing such calculations than working out such things 'by-hand' (i.e. more prone to human error, if doing the same calculation/checks over many rows of participants in your datasets).