Statistics for Psychologists - PSYC122

Margriet Groen and Rob Davies

2022-01-21

Contents

Chapter 1

Intro

This is a collection of tuition material written for Psychology undergraduates at Lancaster University. At the moment the content represents the "lab materials" for the PSYC122 module in first year. As was the case for PSYC121, they feature tuition of programming with R, building on the skills you developed last term.

1.1 Analysis labs and 'pre-lab work'

Some parts should be completed before you attend the lab session (watching lectures, reading chapters, pre-lab activities). All the links to the different materials and activities are also in the 'to-do list' for the relevant week on Moodle.

Chapter 2

Week 11: Correlation

Written by Margriet Groen (partly adapted from materials developed by the PsyTeachR team a the University of Glasgow)

Today we will take a look at correlation as a measure of association between two numerical variables. We will create scatterplots to visualise correlations, we will run a correlation analysis and we will practise interpreting and reporting the results.

2.1 Lectures

The lecture material for this week is presented in two parts:

1. **Theory** Watch this part before you complete the reading and the pre-lab activities.
2. **How to** Watch this part either after the 'Theory' part of after you've completed the pre-lab activities 1 to 3. Definitely watch it before you come to your lab session.

2.2 Reading

The reading that accompanies the lectures this week and next week is from **the free textbook by Miller and Haden**.

Chapter 10 gives you a brief overview of what correlation and regression are. **Chapter 11** introduces correlation in more detail. Both chapters are really short but provide a good basis to understanding correlational analysis. Please note, in Chapter 10 you might encounter some terminology that is unfamiliar

to you. It talks about ANOVA, which means Analysis of Variance and about GLM, which means General Linear Model. Having a quick look at Chapter 1 of Miller and Haden also helps with that.

2.3 Pre-lab activities

After having watched the lectures on correlation and read the textbook chapters you'll be in a good position to try these activities. Completing them before you attend your lab session will help you to consolidate your learning and help move through the lab activities more smoothly.

2.3.1 Pre-lab activity 1: Visualing correlations

Have a look at **this visualisation of correlations** by Kristoffer Magnusson.

After having read Miller and Haden Chapter 11, use this visualisation page to visually replicate the scatterplots in Figures 11.3 and 11.4 - use a sample of 100. After that, visually replicate the scatterplots in Figure 11.5.

Each time you change the correlation, pay attention to the shared variance (the overlap between the two variables) and see how this changes with the changing level of relationship between the two variables. The greater the shared variance, the stronger the relationship. Also, try setting the correlation to r = .5 and then moving a single dot to see how one data point, a potential outlier, can change the stated correlation value between two variables.

2.3.2 Pre-lab activity 2: Guess the correlation

Now that you are well versed in interpreting scatterplots (scattergrams) have a go at **this online app on guessing the correlation**.

This is a very basic app that allows you to see how good you are at recognising different correlation strengths from the scatterplots. We would recommend you click the "Track Performance" tab so you can keep an overview of your overall bias to underestimate or overestimate a correlation.

Is this all just a bit of fun? Well, yes, because stats is actually fun, and no, because it serves a purpose of helping you determine if the correlations you see in your own data are real, and to help you see if correlations in published research match with what you are being told. As you will have seen from the above examples, one data point can lead to a misleading relationship and even what might be considered a medium to strong relationship may actually have only limited relevance in the real world. One only needs to mention Anscombe's

Quartet to be reminded of the importance of visualising your data, which leads us to the final pre-lab activity for this week.

2.3.3 Pre-lab activity 3: Anscombe's quartet

Anscombe (1973) showed that four sets of bivariate data (X, Y) that have the exact same means, medians, and relationships can look very different when plotted. You can read more about this **here**.

All in this is a clear example of why you should visualise your data and not to rely on just the numbers.

2.3.4 Pre-lab activity 4: Getting ready for the lab class

2.3.4.1 Remind yourself of the basics of how to work with RStudio.

You might want to re-watch some of the videos John and Tom provided in PSYC121:

- Video on how to upload a zip file and import data (3.5 mins)
- Video on basic operations in RStudio (8 mins)
- Video on using scripts and using the console (3 mins)

2.3.4.2 Create a folder and a Project for Week 11.

Click here for the instructions from Week 6 of PSYC121 if you are unsure.

2.3.4.3 Get your files ready

Download the 122_week11_forStudents.zip file and upload it into the new folder in RStudio Server you created at the previous step. If you need them, here are the instructions from Week 2 of PSYC121.

2.4 Lab activities

In this lab, you'll gain understanding of and practice with:

- constructing and interpreting scatterplots
- running correlation analysis and interpret the results
- reporting the results in APA format
- constructing a correlation matrix in APA format

- when and why to apply correlation analysis to answers questions in psychological science

2.4.1 Lab activity 1: Interpreting correlation

2.4.1.1 Question 1

Below are scatterplots that show the relationship between 'how much you know about correlation and how attractive you appear to members of the opposite (&/or same) sex'. Choose the type of correlation (strength and direction) displayed in each graph using one of the following:

- Perfect positive correlation
- Perfect negative correlation
- Strong positive correlation
- Strong negative correlation
- Moderate positive correlation
- Moderate negative correlation
- Null correlation

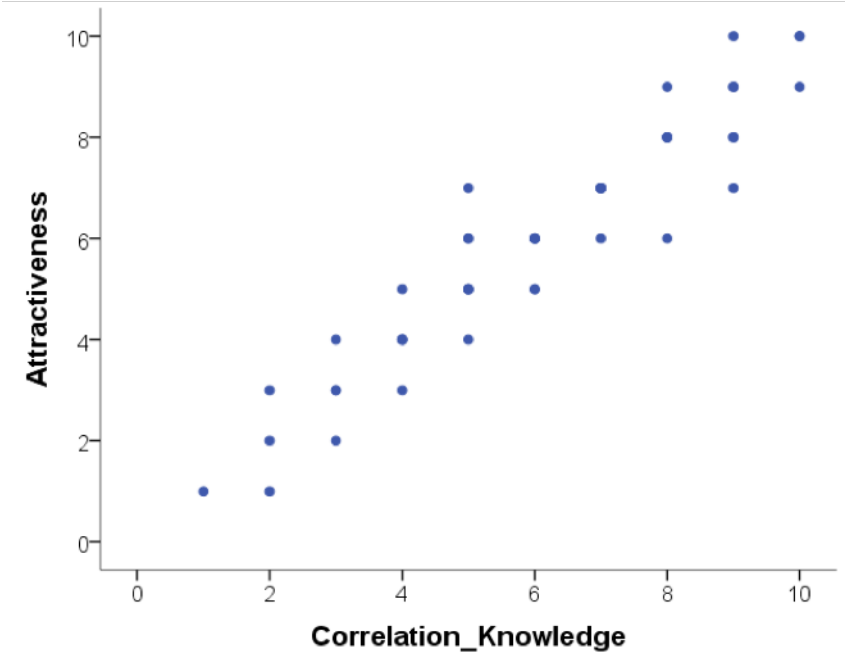


Figure A

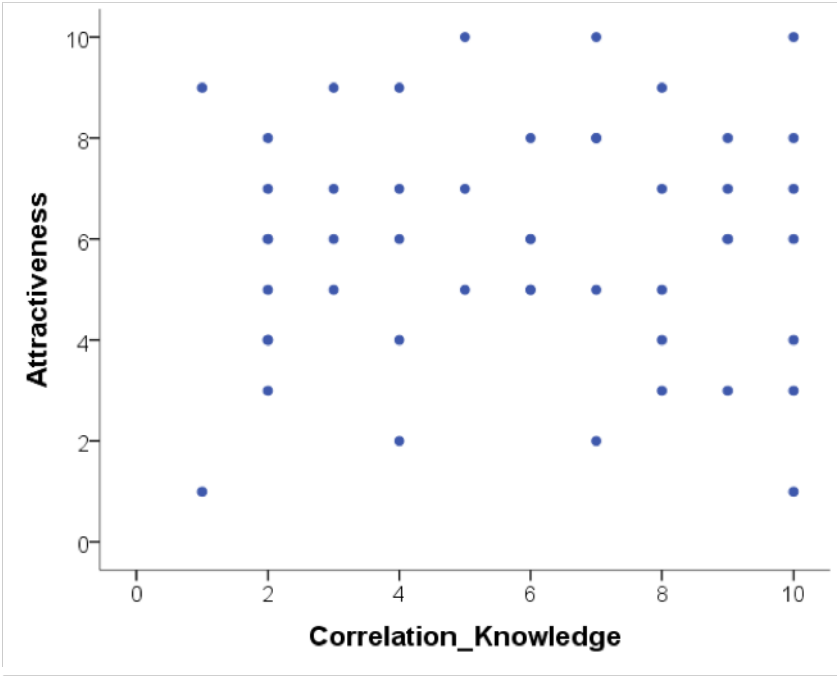


Figure B

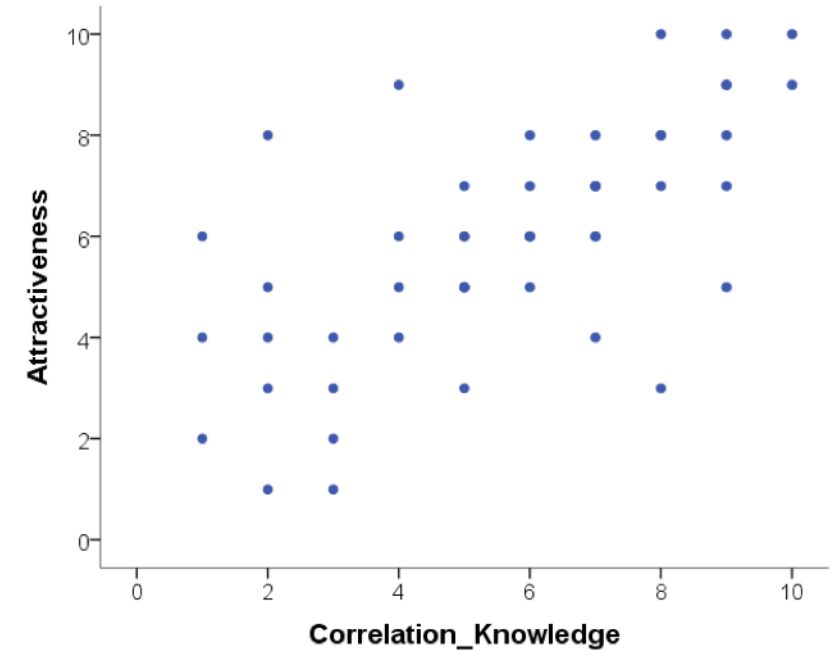


Figure C

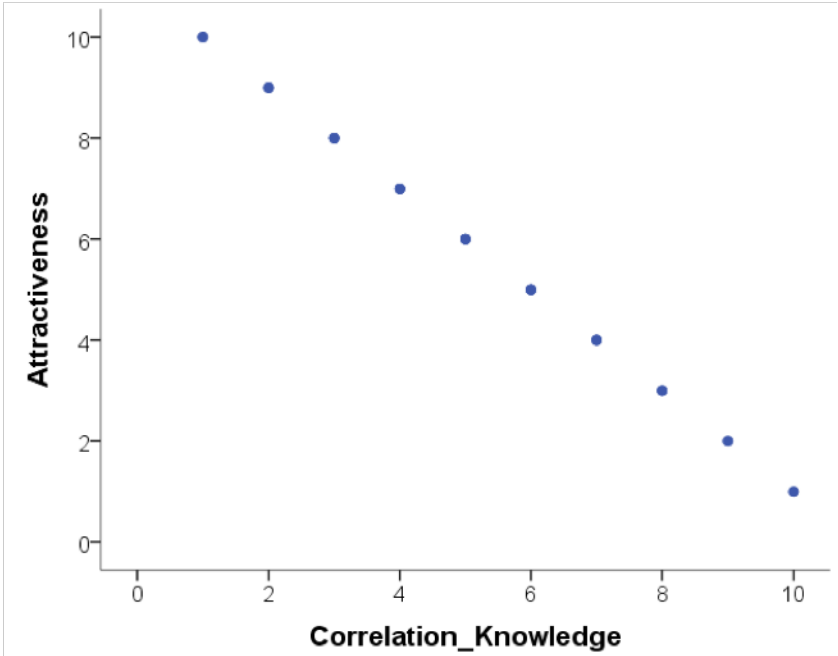


Figure D

2.4.1.2 Question 2

Suppose it was observed that there is a correlation of r = -.81 between a driver's age and the cost of car insurance. This correlation would mean that, in general, older people pay more for car insurance.

TRUE or FALSE? Explain why.

Note: explain your chosen answer based on the statistic given, not on why you think the correlation may or may note make 'logical' sense).

2.4.1.3 Question 3

Suppose that there is a correlation of r = .87 between the length of time a person is in prison and the amount of aggression the person displays on a psychological inventory administered at release. This means that spending a longer amount of time in prison *causes* people to become more aggressive.

TRUE or FALSE? Explain why.

2.4.1.4 Question 4

A significant correlation was found between having great hair and performance in correlation labs. The correlation coefficient was .7. How much variance in correlation lab performance can the 'greatness' of your hair explain?

- 51%
- 70%
- 49%
- 30%
- Who cares I've got great hair.

What was the reason for your answer?

What is this 'new coefficient' called?

2.4.2 Lab activity 2: Visualising, calculating and reporting correlations

Going back to the data discussed in Chapter 11 of Miller & Haden, you'll remember it contains data from 25 8-year-old children on:

- a standardised test of reading ability (Abil)
- intelligence (IQ)
- the number of minutes per week spent reading in the home (Home)
- and the number of minutes per week spent watching TV (TV)

In the video on 'How to conduct correlation analysis using R' we looked at the correlation between reading ability and intelligence. Now, let's look at the correlation between number of minutes per week spent reading in the home and watching TV.

The folder you were asked to download under 'Pre-lab activity 4: Getting ready for the lab class' contains the datafile ("MillerHadenData.csv") as well as the R-script from the 'How to …' video (122_wk11_howtoExample.R) that you can use here and adapt.

1. Load the 'broom' and the 'tidyverse' libraries by running the first two lines of code.
2. Read in the data. You should now see an object with 25 observations and 5 variables in the 'Environment'. Click on it to view it.
3. Construct a scatterplot of the relationship between 'Home' and 'TV'.
4. What can you tell from the scatterplot about the direction of the relationship?
5. Conduct the correlation analysis.

6. What is the correlation coefficient (Pearson's r)?
7. What is the p value?
8. Is the correlation significant at the $p < .05$ level?
9. What are the degrees of freedom you need to report?
10. How much variance in 'time spent reading' can be accounted for by 'time spent watching TV'? (Hint: you can use the Console in RStudio as a calculator.)
11. Write a few sentences in which you report this result, following APA guidelines.

2.4.3 Lab activity 3: More correlations

Researchers were interested in the relationship between hazardous alcohol use and impulsivity (making unplanned, rapid decisions without thinking or 'acting on a whim'). To investigate the relationship, 20 participants completed both the alcohol use disorder identification test (AUDIT; Saunders, Aasland, Babor, de la Fuente, & Grant, 1993) and the Barratt's Impulsiveness Scale (BIS-11) (Patton, Stanford, & Barratt, 1995). The datafile ("alcoholUse_Impulsivity.csv") is in the folder you were asked to download under 'Pre-lab activity 4: Getting ready for the lab class'. Again, the R-script from the 'How to …' video (122_wk11_howtoExample.R) is useful here.

1. Load the 'broom' and the 'tidyverse' libraries by running the first two lines of code.
2. Read in the data. You should now see an object containing the data in the 'Environment'. How many variables does it have?
3. Construct a scatterplot of the relationship between 'Hazardous Alcohol Use' and 'Impulsivity'.
4. What can you tell from the scatterplot about the direction of the relationship?
5. Conduct the correlation analysis.
6. What is the correlation coefficient (Pearson's r)?
7. What is the p value?
8. Is the correlation significant at the $p < .05$ level?
9. What are the degrees of freedom you need to report?
10. How much variance in 'impulsivity' can be accounted for by 'hazardous alcohol use'? (Hint: you can use the Console in RStudio as a calculator.)

11. Construct a correlation matrix to display the correlation coefficient in a table.
12. Give three logically possible directions of causality, indicating for each direction whether it is a plausible explanation in light of the variables involved (and why). No, this is not a trick question -—I know that correlation does not infer causation, but think critically! New studies/ideas are constructed by thinking what the previous study doesn't tell us about what could be happening with the variables of interest.

Job completed — Well done!

2.5 Answers

When you have completed all of the lab content, you may want to check your answers with our completed version of the script for this week. **Remember**, looking at this script (studying/revising it) does not replace the process of working through the lab activities, trying them out for yourself, getting stuck, asking questions, finding solutions, adding your own comments, etc. **Actively engaging** with the material is the way to learn these analysis skills, not by looking at someone else's completed code…

2.5.1 Lab activity 1: Interpreting correlation

1. Scatterplots

- a. strong positive correlation
- b. null correlation
- c. moderate positive correlation
- d. perfect negative correlation

2. FALSE Explanation: The correlation coefficient is negative and therefore infers a negative correlation. As such, older people pay less for car insurance: as age increases, car insurance costs decrease.
3. FALSE Explanation: This is a bit of trick question as it has the sneaky 'cause' word in. The correlation coefficient is a positive number, suggesting a positive relationship between length of time in prison and aggression. However, causation cannot be inferred from correlation and therefore we cannot know whether time spent in prison CAUSES aggression, and rather we suggest a relationship between the two that as length of time in prison increases, aggression increases.
4. c 49% The 'coefficient of determination' or 'R-squared' tells us the proportion or variance in one variable that can be predicted if we know the

other variable. We can determine this by squaring the r. Therefore, $.7^2 = .49$, $R^2 = .49$.

2.5.2 Lab activity 2: Constructing scatterplots and calculating correlations

You can download the R-script that contains the code to complete lab activities 2 and 3 here: **122_wk11_labActivities2_3.R**

1. *See R script*
2. *See R script*
3. *See R script*
4. What can you tell from the scatterplot about the direction of the relationship? **There is a negative association between 'Home' and 'TV'. This means that the longer a child spends watching TV, the shorter they will read at home.**
5. Conduct the correlation analysis. *See R script*
6. What is the correlation coefficient (Pearson's r)? $\boldsymbol{r = -.65}$
7. What is the p value? $\boldsymbol{p < .001}$
8. Is the correlation significant at the p $< .05$ level? **Yes, because the p-value is smaller than .005**
9. What are the degrees of freedom you need to report? **23**
10. How much variance in 'time spent reading' can be accounted for by 'time spent watching TV'? **42%**
11. Write a few sentences in which you report this result, following APA guidelines. **Something along the lines of: A Pearson's correlation coefficient was used to assess the relationship between time spent reading at home and time spent watching TV at home. There was a significant negative correlation, $\boldsymbol{r(23) = -.65}$, $\boldsymbol{p < .001}$. As time spent watching TV at home increased, time spent reading at home decreased.**

2.5.3 Lab activity 3: Hazardous alcohol use and impulsivity

1. *See R script*
2. How many variables does it have? **3**
3. *See R script*

4. What can you tell from the scatterplot about the direction of the relationship? **There is a positive association between 'hazardous alcohol use' and 'impulsivity'. This means that as a participant's score on 'hazardous alcohol use' goes up, their score on 'impulsivity' also goes up.**
5. *See R script*
6. What is the correlation coefficient (Pearson's r)? $\boldsymbol{r = .54}$
7. What is the p value? $\boldsymbol{p = .014}$
8. Is the correlation significant at the $p < .05$ level? **Yes**
9. What are the degrees of freedom you need to report? **18**
10. How much variance in 'impulsivity' can be accounted for by 'hazardous alcohol use'? (Hint: you can use the Console in RStudio as a calculator.) **29%**
11. Construct a correlation matrix to display the correlation coefficient in a table.

Table 1. ***A correlation matrix showing the relationship between hazardous alcohol use and impulsivity.***

	Hazardous alcohol use	Impulsivity
Hazardous alcohol use	-	
Impulsivity	.54*	-

***p < .05**

12. Give three logically possible directions of causality, indicating for each direction whether it is a plausible explanation in light of the variables involved (and why). No, this is not a trick question -—I know that correlation does not infer causation, but think critically! New studies/ideas are constructed by thinking what the previous study doesn't tell us about what could be happening with the variables of interest.

Just really looking for reasoning here.

Examples:

- **Being more impulsive may make people consume more alcohol.**
- **Consuming more alcohol may make people more impulsive.**
- **An outgoing personality might influence both your level of impulsivity and you are more likely to be socialising in the pub and consuming alcohol. So the same 'third factor' may influence both our variables of interest.**

Chapter 3

Week 12: Correlation 2

Written by Margriet Groen (partly adapted from materials developed by the PsyTeachR team a the University of Glasgow)

Today we will continue a look at correlation as a measure of association between two numerical variables. We will review assumptions associated with correlation, discuss some issues important to be aware of when interpreting correlation results and finally, we'll talk about intercorrelation.

3.1 Lectures

The lecture material for this week is presented in two parts:

1. **Correlation – Assumption, issues and intercorrelation – Theory**
2. **Correlation – Assumption, issues and intercorrelation – How to**

3.2 Reading

The reading that accompanies the lectures this week is (the same as last's week) from **the free textbook by Miller and Haden**.

Chapter 10 gives you a brief overview of what correlation and regression are. **Chapter 11** introduces correlation in more detail. Both chapters are really short but provide a good basis to understanding correlational analysis. Please note, in Chapter 10 you might encounter some terminology that is unfamiliar to you. It talks about ANOVA, which means Analysis of Variance and about GLM, which means General Linear Model. Having a quick look at Chapter 1 of Miller and Haden also helps with that.

3.3 Pre-lab activities

After having watched the lectures on correlation and read the textbook chapters you'll be in a good position to try these activities. Completing them before you attend your lab session will help you to consolidate your learning and help move through the lab activities more smoothly.

3.3.1 Pre-lab activity 1: Online interactive tutorial to practise your data-wrangling skills

Data comes in lots of different formats. One of the most common formats is that of a two-dimensional table (the two dimensions being rows and columns). Usually, each row stands for a separate observation (e.g. a participant), and each column stands for a different variable (e.g. a response, category, or group). A key benefit of tabular data is that it allows you to store different types of data-numerical measurements, alphanumeric labels, categorical descriptors-all in one place.

It may surprise you to learn that scientists actually spend far more of time cleaning and preparing their data than they spend actually analysing it. This means completing tasks such as cleaning up bad values, changing the structure of tables, merging information stored in separate tables, reducing the data down to a subset of observations, and producing data summaries. Some have estimated that up to 80% of time spent on data analysis involves such data preparation tasks (Dasu & Johnson, 2003)!

Many people seem to operate under the assumption that the only option for data cleaning is the painstaking and time-consuming cutting and pasting of data within a spreadsheet program like Excel. We have witnessed students and colleagues waste days, weeks, and even months manually transforming their data in Excel, cutting, copying, and pasting data. Fixing up your data by hand is not only a terrible use of your time, but it is error-prone and not reproducible. Additionally, in this age where we can easily collect massive datasets online, you will not be able to organise, clean, and prepare these by hand.

In short, you will not thrive as a psychologist if you do not learn some key data wrangling skills. Although every dataset presents unique challenges, there are some systematic principles you should follow that will make your analyses easier, less error-prone, more efficient, and more reproducible.

In the online tutorial, you will see how data science skills will allow you to efficiently get answers to nearly any question you might want to ask about your data. By learning how to properly make your computer do the hard and boring work for you, you can focus on the bigger issues.

You'll be practising the `select()`, `filter()`, `mutate()`, `arrange()`, `group_by()` and `summarise()` functions from the dplyr package.

You've used these functions before, but if you'd like to quickly remind yourself what they do, watch the video (~10 min) on **Data wrangling: dplyr and pipes**. As the title suggests, I also explain in the video what a 'pipe' (this thing: `%>%`) is and you'll be practising with that as well.

If you're ready to begin, go to the tutorial linked to below. There is no need to install or download anything. Each tutorial has everything you need to write and run R code, right in the tutorial.

- **Working with Tibbles** Practise how to extract values form a table, subset tables, calculate summary statistics, and derive new variables.

3.3.2 Pre-lab activity 2: Getting ready for the lab class

3.3.2.1 Get your files ready

Download the 122_week12_forStudents.zip file and upload it into the new folder in RStudio Server you created (see last week's Pre-lab activity 4 for instructions on how to do that.

3.4 Lab activities

In this lab, you'll gain understanding of and practice with:

- constructing and interpreting histograms and qq-plots
- constructing and interpreting a matrix of scatterplots
- running intercorrelation analysis and interpret the results
- correct for multiple comparisons when running intercorrelation analysis
- constructing a correlation matrix in APA format
- when and why to apply correlation analysis to answers questions in psychological science

3.4.1 Lab activity 1: Assumptions of Correlation Analysis

3.4.1.1 Question 1

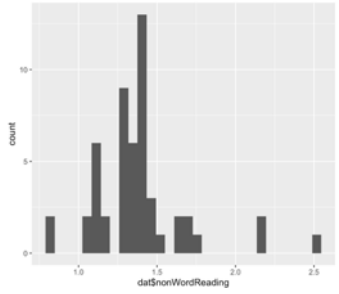
Correlation would be an appropriate form on analysis for researchers interested in the relationship between a. Dog (breed) and height (cm) of owner b. Speed of swimming (mph) and area of tank (cm) c. Number of cows sitting and rain fall (mm) d. Total llama saliva (ml) expelled and gender of visitors e. b and d f. b and c

3.4.1.2 Question 2

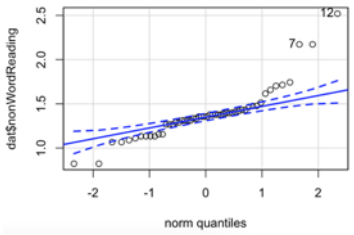
When would you use Spearman's rho analysis instead of Pearson's r? a. When there are clear outliers in the data b. When the data is not normally distributed c. When the relationship between X and Y is curvilinear d. a and b

3.4.1.3 Question 3

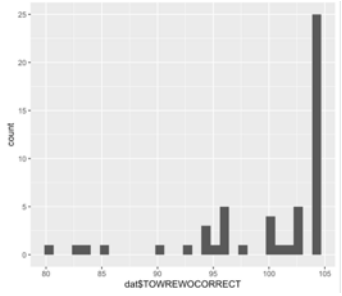
Using the histograms and qq-plots below, which of these variables satisfies the normality assumption? Explain your answers.



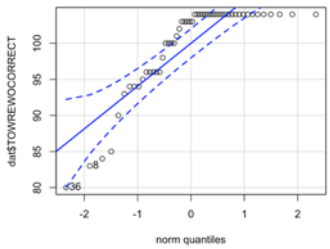
Histogram non-words



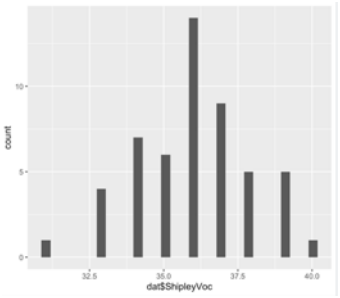
QQ-plot non-words



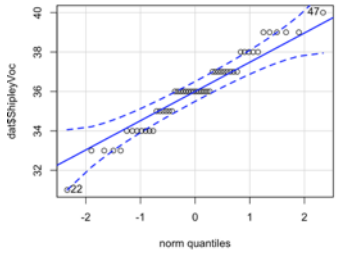
Histogram words



QQ-plot words



Histogram vocabulary



QQ-plot non-words

3.4.1.4 Question 4

Why should correlation analysis not be conducted on variables with a curvilinear relationship?

3.4.2 Lab activity 2: Attitudes towards vaping

WILL BE ADDED EARLY NEXT WEEK

3.5 Answers

When you have completed all of the lab content, you may want to check your answers with our completed version of the script for this week. **Remember**,

looking at this script (studying/revising it) does not replace the process of working through the lab activities, trying them out for yourself, getting stuck, asking questions, finding solutions, adding your own comments, etc. **Actively engaging** with the material is the way to learn these analysis skills, not by looking at someone else's completed code…

The answers to the questions and the script containing the code will be available after the final lab session has taken place.