**Spearman’s Rank Correlation in R**

***Exploring the link between experimental design, statistical analysis and the writing of scientific reports with a focus on linear models and non-parametric alternatives such as Spearman’s rank correlation.***

**Introduction**

This tutorial is designed to guide you through the process of designing an experiment so that you can conduct the appropriate statistical test for the research question you have in mind. You will also learn about the different statistical tests, focusing on linear models and their non-parametric alternatives. You will use your R coding skills to test whether a dataset matches the assumptions of a linear model and learn what you can do if it doesn’t. Having found (spoiler alert) that the dataset doesn’t meet these assumptions you will use R once more to conduct a Spearman’s Rank Correlation test and find out how to present these results in a scientific report.

This tutorial is aimed at coders who have done the following Coding Club tutorials (or equivalent, I’m not just here to advertise Coding Club 😉 ):

* <https://ourcodingclub.github.io/tutorials/intro-to-r/>
* <https://ourcodingclub.github.io/tutorials/troubleshooting/>
* <https://ourcodingclub.github.io/tutorials/data-manip-efficient/>
* <https://ourcodingclub.github.io/tutorials/datavis/>

**Tutorial Aims**

* Understand the importance of considering your statistical analysis when creating your experimental design
* Learn which statistical tests apply to your dataset
* Learn about different linear models and their non-parametric alternatives
* Learn the assumptions of linear regressions and what to do if your data does not match them
* Use R to test your data matches linear model assumptions
* Run a Spearman’s Rank Correlation and linear regression in R
* Interpret the results of a Spearman’s Rank Correlation and linear regression
* Create a plot to visualise the outputs of both statistical tests
* Learn the appropriate way to refer to the outputs of statistical tests in the results section of a scientific report

**Tutorial Steps**

1. Basics of Experimental Design
2. Choosing a Statistical Test
3. Linear Models
4. Linear Regressions and Their Assumptions
5. Testing the Assumptions of Linear Regressions
6. Transforming Your Data to Match Linear Regression Assumptions
7. Non-Parametric Alternatives
8. Interpreting Your Results
9. Visualising the Relationship Between Continuous Variables
10. Presenting Your Results
11. **Basics of Experimental Design**

Considering your statistical test during the planning stage of your experiment will make the writing of your scientific report much easier. Some things can’t be fixed by data manipulation but require thinking about the ecological system which is central to your study and incorporating that into your experimental design.

Here is a simple but indescribably useful workflow that can help you with experimental design.

Start by considering the **general** **background** for your study. What system do you want to examine? What will the bigger picture significance of your study be? You then want to narrow this down into a **specific** **context** through which you can address this topic. This is where you first start to brainstorm about the practical aspects of your data collection. You can only answer so many questions with one experiment and are likely to be constrained by access to equipment, study organisms, time etc. You then want to frame a **research** **question** within this context. Avoid starting your research question with “why”. Try using one of the following:

* What? (has an effect)
* How? (does a process function)
* How? (large/important is a given effect)

This will make it easier to create predictions in the response and explanatory variable format.

Before you come up with your predictions you need to generate a **hypothesis**. Your hypothesis gives an explanation for the pattern you expect to see in your data. If you are conducting experiments at the undergraduate level it is unlikely you will be testing to see if an effect exists so it is best to avoid having a hypothesis that says “(Explanatory variable) will have an effect on (response variable)” but rather to be clear on what the effect will be e.g. metabolic rate will increase with temperature. You will show a more robust understanding of the system you are examining by phrasing it this way.

It is also good to generate a **null** **hypothesis**, though it is not standard practice to state this explicitly in a scientific report, it will help you understand what the results of your statistical test represent. A null hypothesis is not the opposite of your hypothesis per se. If your hypothesis were that total base saturation would decrease with sand content because there will be fewer binding sites for cations and the null hypothesis would be that sand content will have no effect on the total base saturation of the soil and not that total base saturation will decrease with increasing sand content. The hypothesis tends to be less specific than your prediction(s).

Your **predictions** should refer to your response and explanatory variables, but these might be more specific than those in your hypothesis. Your predictions might reflect restrictions on your experimental design where you had to choose a proxy measurement to represent one of your variables and therefore the variables aren’t the same for the hypothesis and predictions. Predictions are what you use to verify your hypothesis.

Going through these steps will mean that half of your scientific report is basically done before you even collect any data and will give you a useful resource to refer to in coming up with your statistical test. This workflow also happens to be a good way to structure your introduction in a scientific report. Starting with important background information and gradually adding more detail so the reader can see your thought process and motivation for coming up with your final experimental design.

Other things you want to consider when designing your experiment with your statistical test in mind are the effect size (magnitude of the effect you are measuring), amount of random variation and number of replicates. These factors will affect the statistical power of your experiment. Statisticalpower is the probability that a particular experiment will detect an effect assuming that there is an effect to be detected.

The first two factors are mostly biologically determined and are therefore harder to control. But effect size can be increased based on the treatments you apply. Random variation can also be accounted for during your experiment by controlling for confounding variables. But if you’re up for a challenge you can always try accounting for confounding variables at the data manipulation stage of your experiment by creating a linear mixed effects model instead. This goes beyond the scope of this tutorial but luckily for the thrill seekers amongst you there is already a [Coding Club tutorial](https://ourcodingclub.github.io/tutorials/mixed-models/) on this. The number of replicates is purely down to experimental design and is the factor you have the most control over. This [link](https://www.statmethods.net/stats/power.html) shows you how to calculate the optimum number of replicates/sample size using R, to maximise the statistical power of your experiment so your linear model or non-parametric alternative is more likely to show a tangible result.

Going through these steps will make it more likely that you will gain a significant p-value if there is an effect to be detected. We will get onto p-values later but if you haven’t heard of them before that’s pretty much the dream in frequentist statistics.

1. **Choosing a Statistical Test**

This table describes how you choose your statistical test based on the nature (continuous or categorical) of your response and explanatory variables. Categorical variables can be split into discrete categories whereas data points on continuous variables are not categorized based on being either one thing or another but can sit anywhere along a continuum. For instance, if you are collecting data on eye colour, this would be a categorical variable because you would have people with blue *or* green *or* brown eyes. (Of course, you could say there are people with green-ish blue eyes but for the sake of this tutorial let’s not make things needlessly complicated.) But if you measured human height these data points would span a range of values and therefore be a continuous variable.

*Diagram of table with potential statistical tests based on nature of variables, linear models highlighted*

1. **Linear Models**

***Linear Model Definition***

When the response variable is continuous we tend to run linear models. Linear model is a general term that encompasses many types of commonly used statistical test. Basically, a linear model is an equation designed to represent the relationship that your data indicates exists between your response and explanatory variable. The equation takes a different format depending which linear model you choose. A statistical test is then conducted on this equation to see how closely the model matches the raw data you collected. If the model fits your data well this usually means there is a pattern to see as we can assume that that pattern is reflected in the linear model. This is also why good linear models can be used to predict the value of the response variable at a certain explanatory variable value for which we have no measurement. If your mode is a good fit and you dataset is statistically significant (the patterns seen in your data are unlikely to be due to chance), you can conclude that the pattern shown by the linear model is likely to be representative of the true relationship between your response and explanatory variables. You then look at the outputs of the linear model to determine whether the nature of that relationship is the same as you predicted i.e. whether it supports your hypothesis.

***Definition of Each Linear Model***

An analysis of variance ([ANOVA](https://www.scribbr.com/statistics/anova-in-r/)) is used to examine the relationship between a continuous response and a categorical explanatory variable. Unlike a T-test, ANOVAs can be used where the explanatory variables has more than two categories. Linear regressions are used to understand the relationship between two continuous variables. Specifically, if one significantly affects the other. An analysis of co-variance (ANCOVA) is a combination of an ANOVA and a regression. An ANCOVA tests whether the means of a continuous response variable are equal across levels/treatments of a categorical explanatory variable whilst controlling for covariates. It is a type of linear mixed effects model in that it controls for potential confounding variables. A multiple regression is the same as a linear regression in that it describes the linear relationship between continuous variables but it uses multiple explanatory variables to explain the response of a single response variable.

1. **Linear Regressions and Their Assumptions**

There are three main reasons to run a linear regression: to test whether the explanatory variable has a significant effect on the response variable, to quantify that relationship or to predict values of the response variable for which you have a measure of the explanatory variable but not the response variable.

When you run a linear regression you are essentially generating a straight line equation to describe the observed relationship between your variables and testing its ability to represent this relationship precisely and accurately. The equation comes in the same format as the equation of a straight line (y = mx + c) but different terms are used to indicate that we are referring to a linear model, observe: y = β1x + β0 + ε. Where y is your explanatory variable, x is your response variable, β0 is still the y-intercept and β1 is still the gradient of the slope but there is an additional character “ε”. Epsilon is representing the error, the magnitude of the residuals. This epsilon character quantifies the variation in x that cannot be explained by y.

*Diagram of linear regression with equation and residuals*

The beta terms are the parameters we speak of when we call this a parametric model. The parameters and this epsilon character are the outputs we referred to earlier which can tell us about the relationship between our response and explanatory variables.

Linear regressions can only be applied to certain datasets as they have a few assumptions associated with them:

* All observations are independent i.e. no [pseudoreplication](https://en.wikipedia.org/wiki/Pseudoreplication)
* The variances of the residuals are the same regardless of the value of the explanatory variable
* The residuals are normally distributed

*Diagram of normally distributed residuals for linear regression*

* The response variable responds in a linear manner to the explanatory variable (residuals are [homoscedastic](https://en.wikipedia.org/wiki/Homoscedasticity) (great word))
* There is no error in the measurement of the explanatory variable
* The response variable varies in a linear manner with respect to the explanatory variable

1. **Testing the Assumptions of Linear Regressions**

We want to quantify the relationship between sand content (continuous) and total base saturation (continuous) so ideally we would run a linear regression but first we need to test that our data meets the assumptions.

The Shapiro-Wilk test is an easy way to check whether your residuals are normally distributed.

Code for Shapiro Wilk Test

*Output of Shapiro-Wilk test*

If the p-value > 0.05 then the residuals are normally distributed and your data meets one of the assumptions. Ours doesn’t but let’s see about the other assumptions.

By first running a linear regression, we can quickly plot four diagnostic plots which we can use to assess whether our dataset can have a linear regression applied to create residuals vs fitted, normal quantile-quantile (Q-Q), scale-location and residuals vs leverage plots.

Code for plots

The residuals vs fitted plot shows whether the relationship between our variables is linear. The line should be pretty straight but it doesn’t need to sit on the dotted line.

*Diagram of residual vs fitted plot*

Our plot doesn’t indicate the relationship between total base saturation and sand content is linear.

If the residuals are normally distributed the points will sit on the dashed line. It is common and acceptable for the points to deviate slightly towards the extremes.

*Diagram of QQ plot*

Ours deviate quite drastically I’m afraid.

With the scale-location plot we are looking for equally (randomly) spread points around a somewhat horizontal line.

*Diagram of scale-location plot*

Once again, we are not finding this pattern.

The residuals vs leverage plot helps us find out which outliers are influential in the linear regression analysis. Values in the upper or lower left of the plot indicate issues if they are shown with Cook’s distance lines. Patterns are not relevant here.

*Diagram of residuals vs leverage plot*

At least this plot shows we don’t have any influential outliers.

But in conclusion, not even close…

*Gif of Blackadder quote*

Don’t give up though.

1. **Transforming Your Data to Match Linear Regression Assumptions**

It is simple to remove outliers from your data because R conveniently displays the row numbers of outliers (though not all are necessarily influential) for you in the residuals vs leverage plot. Use the filter() and row\_number() functions to remove rows containing outliers. Bear in mind that you want to maintain a large enough sample size. The standard sample size in 10 times the number of parameters.

Transform your data based on what you see when you produce a histogram of the distribution of your response variable.

Code for creating histogram of response variable

*Histogram of response variable*

This [link](https://rcompanion.org/handbook/I_12.html) provides more information on how to transform your data based on the distribution it currently shows. Generally you start by only transforming your response variable but you might find that transforming both will give you a closer to normal distribution. Just remember that the outputs of your statistical test will reflect the relationship between the transformed variables and not the ones in your prediction. It might be best to [back-transform](http://www.biostathandbook.com/transformation.html) your data and make a statement in your results based on that so it is immediately clear to your reader whether your predictions were correct and it is easy for them to relate that to the bigger picture significance of your study.

Some datasets just won’t satisfy the assumptions of a linear model no matter how much data transformation you do. Luckily, there is a non-parametric alternative.

*Diagram of flowchart for parametric vs non-parametric*

1. **Non-Parametric Alternatives**

Why not save yourself the hassle of all this data transformation jiggery pokery and go straight for a non-parametric alternative? Well, non-parametric alternatives to linear models do not have parameters.

*Obviously gif from Harry Potter-* [*https://c.tenor.com/ewWvv0\_ZTPMAAAAC/obviously-snape.gif*](https://c.tenor.com/ewWvv0_ZTPMAAAAC/obviously-snape.gif)

This makes them less desirable because you cannot quantify the relationship between your response and explanatory variables as clearly and they have less statistical power. You can visualize your data as a scatter plot and gain an idea that way. One benefit of the lack of parameters means that you don’t have to think about back-transforming if you decided to keep your data log transformed.

Spearman’s rank correlation is a common non-parametric alternative to a linear regression. It ranks all the x and y values separately. It is a measure of the statistical dependence of ranking between two variables e.g. if the lowest value of x always correlates with the lowest value of y then the relationship would be monotonic so the Spearman’s correlation coefficient (rho) would be 1. This is different from a linear relationship because x and y do not need to increase/decrease at a constant rate. We shall now run a linear regression (for the sake of comparison) and a Spearman’s rank correlation on to test the relationship between total base saturation and sand content.

Code for running linear regression and Spearman’s rank correlation

1. **Interpreting Your Results**

These are the results from our liner regression:

*Diagram of annotated results table screen grab for linear regression*

The F statistic is not vital to interpreting results in this context. What really want to know is the p-value.

***What is a p-value?***

Linear models are technically used to disprove the null hypothesis rather than prove your hypothesis. Expected values are generated based on the null hypothesis and then the probability of you gaining the results you did if the null hypothesis were true is represented as a p-value. The arbitrary but standard threshold for significance is p<0.05. If the p-value is less than 0.05, the probability of obtaining the results you did if the null hypothesis were true is less than 5%. The idea being that you are very unlikely to have observed the relationship you did by chance and therefore the relationship you observed is more likely to reflect that of the true relationship between your chosen variables than the null hypothesis is.

*Diagram of what a p-value is*

***What is R2?***

The R2 value is a measure of how much variation in the response variable is due to your explanatory variable. An R2 value of 1 means all observed values fall on the regression line whereas 0 means the opposite. A high R2 value therefore tell you the explanatory power of your model is high and will make predictions that are more accurate.

R2 is less important if you do not want to use your model for predictions and just want to understand the relationship because a significant effect is a significant effect (p-value high enough) but knowing the R2 is high will allow you to be more confident if you want to understand the nature of this relationship because you know there is less error surrounding the parameters you have estimated.

These are the results from our Spearman’s rank correlation:

*Diagram of annotated results table screen grab for Spearman’s rank correlation*

Note that the Spearman’s rank correlation provides you with much less detail as to the nature of the relationship between total base saturation and sand content.

R can also help you interpret the results of your linear regression but not your Spearman’s rank correlation (report says it can but it’s lying).

Code for using report() to reveal results of linear regression

*Output of report()*

1. **Visualising the relationship between Continuous Variables**

Both the linear regression and the Spearman’s rank correlation can be displayed using scatter plots because you are trying to demonstrate the same relationship. Just add the equation of the linear regression to one plot and the Spearman coefficient (rho) and its associated p-value to the other.

Code for producing graphs

*Graphs of linear regression and Spearman’s rank correlation*

If you want to learn how to make your plots more snazzy you can use this link.

1. **Presenting Your Results**

The results of your statistical test should not be presented as a list or table of (frankly bland) numbers that your reader has to go through and interpret. You have already interpreted their meaning so you can make a clear, concise statement about what they show. This statement should be in basic language, there is no need for complicated scientific jargon. It should describe the relationship you found in terms of your response and explanatory variables and whether or not you found that relationship to be significant. The outputs of your statistical test are just put in brackets at the end of the sentence so your reader knows the justification for that statement.

For our example we would say: “We found that the total base saturation (%) of the soil decreased with sand content (%) (Spearman’s rank correlation, rS= 0.58 , p = 0.002)

If we were to present the results of our linear regression (which are not valid!) we would do it like so: (linear regression, F1,23= 17.29, p < 0.001)

Note that the outputs of the statistical analysis are in the format (name of statistical test, test statistic degrees of freedom, p value =…) and when the p value has more than three decimal places we say less than 0.001

Always refer to your plots somewhere in text. Either by saying “Figure 1 shows…” or by referring to the figure in brackets at the end of a sentence “…(Figure 1)”.