A Horse Owner's Guide to Interpreting Research Evidence

Part 5: Statistical analysis

By Tracy Bye

In the previous instalment of this series we talked about the scientific method and testing a hypothesis. We have discussed the idea of using an outcome measure (or **dependant variable**) to test the impact of our intervention, but how do we ultimately decide whether our intervention has made a difference to our dependant variable? Intervention studies are often **quantitative** which means they have an output that is numerical. This is where **statistics** come in.

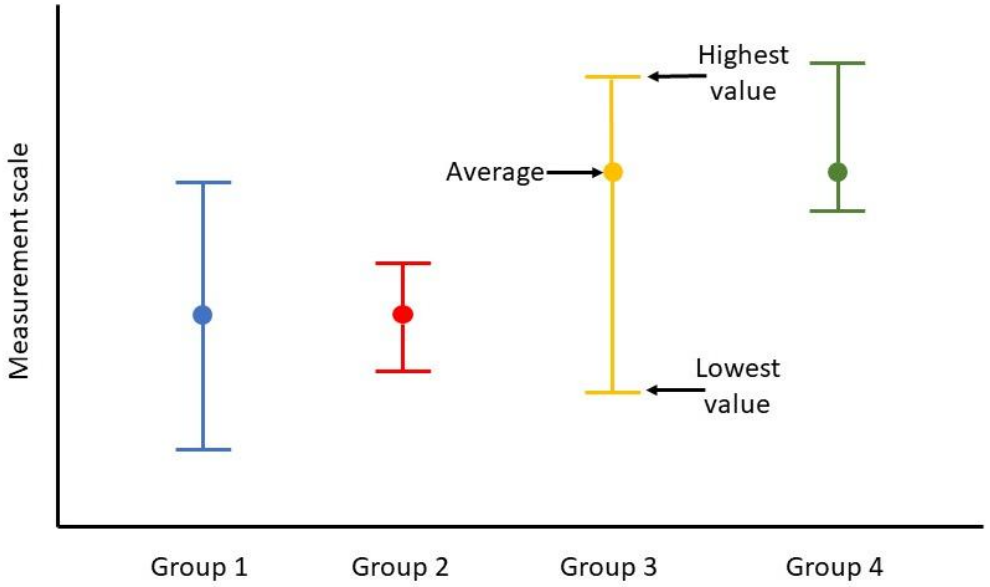
A statistic is quite simply one number that tells us something about a group of numbers. Statistics for intervention studies come in two main types, **descriptive statistics** and **inferential statistics**. Descriptive statistics describe the outcome measure seen in the sample, usually with an average for each group or condition and a measure of how spread out the data are in each group.

Inferential statistics tell you what you can infer from the data, and these are the statistics that we use to work out if our intervention has led to a **significant difference**. When we are doing a scientific study, it is impossible to test every horse in the world, so we conduct our research on a smaller **sample population**. However, we know that all horses are individuals, so inferential statistics tell us whether the difference we have seen is more likely to be due to the intervention, or just a feature of the individual differences of horses in our sample. When we conduct inferential tests, we are aiming to determine whether we will accept or reject our **null hypothesis**. In statistics the null hypothesis is the idea that there is no (i.e. null) difference between the groups.

Scientists will typically do one of a number of inferential tests to come up with a **P value** for the data. 'P' stands for probability, and like the probabilities you will have learnt in school, it can be any value between 0 and 1, with 0 being 'impossible' and 1 being 'certain'. So, the probability of what? P is the probability that we could observe the difference measured (or an even larger difference) due to normal biological variation (individual differences) in a world where the null hypothesis is in fact true. So, if P is large (approaching 1) then it is more likely that any differences are due to chance or biological variation, and if P is low (approaching 0) then it is more likely that the difference is due to something that has been changed between the two groups (i.e. our intervention).

There are lots of different tests which can be used to calculate P for any given study, and the choice of test depends on the study design and sample size, but they all work in a similar way. If there is a large difference between the averages of two groups and the spread of the

data is small then the 'P' for a difference between them goes down (approaches 0) and if the averages of the two groups are close together and the data are more spread out then the 'P' increases (approaches 1). The concept of the P value was developed in the 1920's by British mathematician Ronald Fisher, and he proposed the standard cut off point for determining a **statistically significant difference** between two things as being P=0.05. This is the **level of significance** that most scientists adopt, so if an experiment returns a P value in the range of 0-0.049 this will be referred to a statistically 'significant' result and we would reject our null hypothesis. If the P value is in the range 0.05-1 then this is a 'non-significant' result, so we would accept our null hypothesis, as there is high probability that we could have measured this amount of difference in a world where the null hypothesis is true.



Illustrative graph comparing example measurement for four different groups

The above graph shows a comparison of four different groups for an outcome measure. If we imagine this measure is heart rate, the dots represent the average heart rate value for that group, and the 'whiskers' above and below represent the spread of the data. You can see that Group 2 is much less spread out than Group 1, but the average heart rates of the two groups are the same, so there would not be a statistically significant difference between them. Groups 3 and 4 have a higher average heart rate value than Groups 1 and 2, but their spread is asymmetrical. Group 3 has a very large spread, which overlaps with Groups 1 and 2, and so would not be significantly different from them. Group 4 has a much smaller spread of data. This overlaps with Groups 1 and 3, but not with Group 2. Of all the pairings, Groups 2 and 4 are the ones most likely to have a statistically significant difference between them.

Side note: We do not always use the full range of the data to calculate the spread, we may use another descriptive statistic, such as the standard deviation which shows the middle

two thirds of the data, but you do not need to understand how this works in order to understand the basic concepts of statistics.

Without over-complicating things too much, there are a few extra things to consider when interpreting statistics. Firstly, the difference between statistical significance and biological/clinical significance. We may find a statistically significant effect, but this does not mean that this effect is big enough to have any real impact on the horse.

For example, if we were to study heart rate as a marker of stress in horses housed in different environments like in the images below. If the average heart rate in Group 1 was 42 beats per minute (bpm) and in Group 2 was 38 bpm and this was a statistically significant difference, would we conclude that whatever was happening to Group 1 was particularly stressful? Probably not, as their heart rate is still in the normal range for horses at rest of 38-42 bpm. So here it is important to consider the amount of difference between the averages for the groups (known as **effect size**) as well as the P value when interpreting the results.

Our fictional study compares Group 1: individually stabled horses with average heart rate of 42 bpm (Photo by Anna Kaminova on Unsplash) and Group 2: group housed horses with average heart rate of 38 bpm (Photo by Kyriacos Georgiou on Unsplash)

The other things that are worth thinking about are **statistical power** and **sample size**. These two concepts are linked to each other, so we will discuss them together. **Sample size** means how many horses were in the study. As we know horses come in a variety of shapes, sizes, and temperaments, all of which can influence how they will respond to an intervention in a research study. It is therefore better to have as large a sample of horses in the study as possible. This is why it is not credible to decide whether an intervention makes a difference based on a case study of one horse.

Statistical power is how likely a statistical test is to find a difference if there actually is one. Statistical power increases with sample size. The size of the sample that is needed for a study is worked out based on the size of the difference that we expect to see, and this will differ depending on the outcome measure we are using and what already know about the intervention. Unfortunately, this means that there is no 'ideal' sample size for research, but studies that have not included a power calculation, and have small numbers of horses, may be more likely to falsely accept the null hypothesis (i.e. not find a difference where there really is one).

In order to get the best out of our research, we therefore want to make sure that our sample size is large enough. But on the other side of the coin, we need to be aware of the ethical implications of involving horses in research. We want to use as few animals as we can within research studies, while still having a large enough sample to be able to answer our research question. **In the next instalment of this series we will look at these ideas further as we delve into the world of research ethics.**