

Long term data and measurement error

Published 28 days ago by [Joe Smith](#)

Longer term data

"He's going to get us all killed!" ... "statistically he'll probably only get some of us killed"^[1]

J. K. Penhaul Smith

Introduction

In the [last blog](#) we looked at a mental model which we can use to analyse, understand and talk about the participation hierarchy in a sports club, called the "leaky pipeline". This was a model which was used [successfully by UH1WVC to understand where there were barriers to participation](#), particularly for international students in their club.

What we have not yet talked about is how we and why we must continue this data collection effort, even after we have identified these barriers to participation, and how we might analyse these data to account for measuring something over a longer time period.

The first thing we need to understand is that, despite our best efforts, it is extremely rare to run the same data collection effort twice and achieve exactly the same results. If the method you used was identical then the results will be extremely similar, but even in laboratory controlled systems with single cells organisms, using identical equipment, there are still subtle fluctuations around a set of results. In a sporting context, it would be unusual to be able to run the exact same 5km time twice, let alone three times. Slight changes in things like weather, preparation and mood will all influence the time, even if only by a second. Therefore, we need to be able to account for these differences when we are doing our data collection in our sports club.

This variability is at the heart of modern statistics and it is quite easy to get bogged down in the nitty gritty, if we don't make sure the data we are using is as good as possible and explain our processing and reasoning clearly.

Averaging our data

Because we know we will get slightly different results each time, the normal way to account for this is to either: ask the same set of questions to three or more "randomly selected" groups in a club, or to sample the whole club using the same set of questions, then ask the whole club again a set period of time later (a day, a week or month would be normal times to wait) and do this until we have at least 3 time points.

This allows us to understand the "average" for a given factor within the time we measured it. This is where modern statistics comes into play, which defines which average you should choose and why. For our purposes I suggest that as long as you pick one method, tell your readers which method you have chosen and why, then stick to that method, you will probably not run into too much trouble.

These examples are all taken from the practise data set attached to this blog. This practise data is comparing the number and percentage of people who prefer sausages, burgers and veggie burgers, broken down by participation level. This is so we know what to order when we run a barbeque for the end of the beginners course and how this might be different, if wanted to celebrate the success of an Olympian in the club

You have three choices of average: Mean, Median and Mode. All of these averages can be done in Excel with minimal effort.

Mean: All the values you acquired in your data collection efforts, divided by the number of data points collected. For example: the number of beginners who prefer sausages over the three weekends surveyed: $(200+216+184)/2 = 200$ beginners

Median: the middle of a list of numbers, when sorted into ascending order. The number of beginners who prefer sausages over the three weekends surveyed: (184, 200, 216) 200 is the middle data point, so the median = 200 beginners.

Mode: the most commonly occurring value. So, the mode of the number of Olympians who prefer burgers (0, 0, 0) so the mode = 0 Olympians.

Accounting for measurement error

Once we have a method of averaging our data, then we need to account for the measurement error associated with your data collection method. This is basically a way of saying that "we realise that our data is inherently messy, because this is how the universe works. Therefore, we are saying that the answer is within this set of values." Measurement errors are always reported as being + and - the average you have chosen and are normally presented as error bars on either side of the average, such as by using error bars on a figure.

There are three commonly used choices of metrics to account for measurement error and all can be done easily in Excel.

The first and a commonly applied measurement of data variability is the Standard Deviation which in excel =stdev.s(our data). Formally this is defined as the mean variation of our data about the mean, giving us an indication of how variable our measurement method is. As we can see from the practise data set, the standard deviation of the number of beginners who prefer sausages = 16 beginners. If we were to write about this finding we must report mean = 200 ± 16 beginners (standard deviation)

The second measurement of error and the one which is often used when larger data sets are involved is Standard Error. This accounts for both the standard deviation and the number of times we measured our data. Standard Error = Standard Deviation / \sqrt{V} (number of data points) or, in excel =(stdev.s(our data))/(sqrt(number of data points)). So the standard error of the number of beginners who prefer sausages = $16/\sqrt{3} = 9.24$ and we would report it as mean = 200 ± 9.24 beginners (standard error). This is a time when it might also be appropriate to round our value to the nearest whole number, as 0.24 of a sailor is unlikely to express a preference for colour.

Range: the greatest value of our data set minus the lowest. Most often used when talking about the median or mode of our data. The range of those Olympians who prefer burgers = 0 and would be reported as mode = 0 ± 0 Olympians (range).

There are a number of statistical tests we could apply to these data, however we are not going to do this, as this is beyond the scope of this post.

At its most fundamental level, modern statistics relies upon being able to say that the difference between two or more data points is not due to chance. So, if we were to plot our data into a bar chart (Fig: 1), then we can see that while the mean percentage of people who prefer the different colours is similar, the error bars do not overlap at higher participation levels. This indicates that if we were to run a statistical test upon these data, the differences are unlikely to be due to chance.

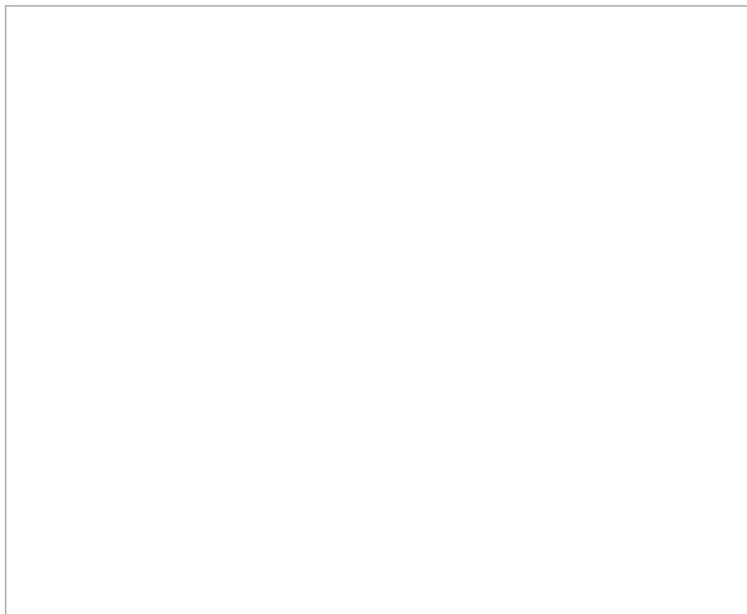


Figure 1: mean percentage of sailors at the defined participation level, who prefer the indicated colour. All bars are ± standard deviation.

What can we learn from these data?

If our error bars do not overlap, then it is unlikely that the results will be due to chance.

This suggests that there is a difference between those who prefer sausages compared to burgers and veggie burgers at higher participation levels. So, if we were planning a barbeque, we could use these data to inform our ordering plans and ensure that we only buy as much food as is going to be eaten.

How might we apply this to real world data? To return to our leaky pipeline model, maybe this method could be used to understand the percentage of the coaches in your club identify as female, compared to male. We might then plot these data against the local population data (available via the Office for National Statistics). If there is a difference in the number or percentage of male or female coaches, across our sampling period and the error bars do not overlap, then we can say with confidence that one group are facing barriers to attaining higher levels and these barriers need to be understood and addressed. Not only that, but we know that the issue that we could see in our data in the previous blog, is consistent across our sampling period and is unlikely to be due to chance. Since we now have these data, we can also say with confidence that, if we plotted before and after our intervention, that our actions are having the desired effect.

Conclusions

In this blog we have seen how we might our data collection process and run it over a set of repeated measurements. This not only improves the accuracy of our data, by allowing us to understand how variable our data collection method might be. Once we have done that, we can then show how the error bars do, or do not, overlap between different samples and how this may be an issue for the development of our club.

If you enjoyed this you can find all my other ConnectedCoaches blogs - including the others in this data series - [here](#).

Login to follow, share, comment and participate. Not a member? [Join for free now](#).

^[1] Inej talks to Jesper. Six of Crows, Leigh Bardugo, 2015, Audible edition, Audible Studios, Webster USA.