Sports club population data: correlation and causation

Published 20/01/21 by Joe Smith

Sports club population data: correlation and causation

"The connections between causes and effects are often much more subtle and complex than we with our rough and ready understanding of the physical world might naturally suppose"[1]

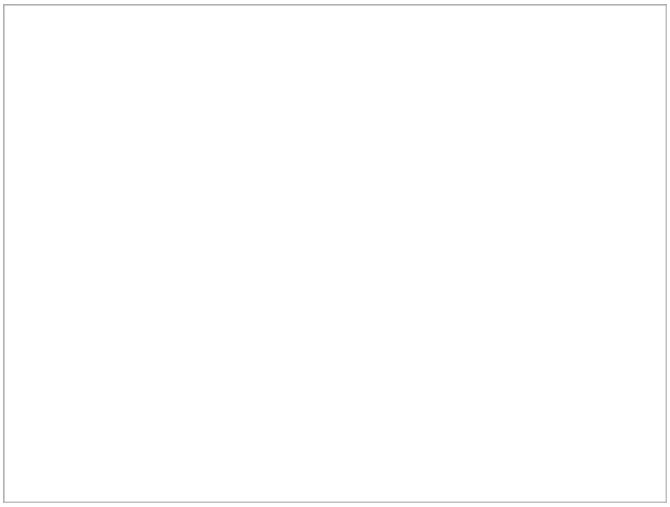
In this month's blog, we start to look at the theory and practise of getting down to business and trying to understand the membership breakdown of your club, team or sport. This process is what UHIWWC built upon, as we saw in a previous case study and requires us to ensure that our data is "accurate, representative, repeatable and answers the question asked"

Broad scale population data

If we were to look at a sports club, then its membership (population) will be broadly drawn from the surrounding area and local population. For example, a rugby club might be made up of members from the local town, or a university sailing club would be made up of students at that university.

Can we use this information, alongside what we already know about the requirements for good data (accurate, representative, repeatable and answering the question asked) to tell us some things about how effective our club is at recruiting members?

Let us imagine that, like UHIWWC, we were to start asking a few basic questions when our members renewed their membership. Please ensure that this is all GDPR/ data protection compliant. We could then compare our "population" to that of the local population, data which is freely available from the Office for National Statistics (ONS. Available here).



Let's take an imaginary sailing club: after we surveyed the entire membership, we found that our club had 200 members, compared to the county which had 200,000 people. Therefore, we can say that, in the year surveyed, our club contained 0.001% of the local population. This information on its own is one number and might not be that helpful, but if we continue this data collection effort, across several years however, then we can confidently say that our club membership may be changing, relative to the local population. Which, if we also check the local county population at the same time, might be "correlated" with changes in the number of people in the local county.

What do we mean if we talk about "correlation"? Does it matter?

The data science that is covered in these blogs is almost entirely correlation data, which contrasts to "causation" data in a number of ways. These differences between correlation and causation are important and must be accounted for when we are looking at, or talking about, population data.

Correlation: A mutual relationship between two variables.

Eg., Clubs with larger membership are more likely to do better at inter-club competitions. One does not necessarily cause the other.

There might be a third factor we haven't measured at play, such as larger clubs also deliberately recruiting the best players to join their club.

Or this could be due to chance, with a number of larger clubs having a good few years of competition at the same time, which was entirely unrelated to their membership recruitment strategies.

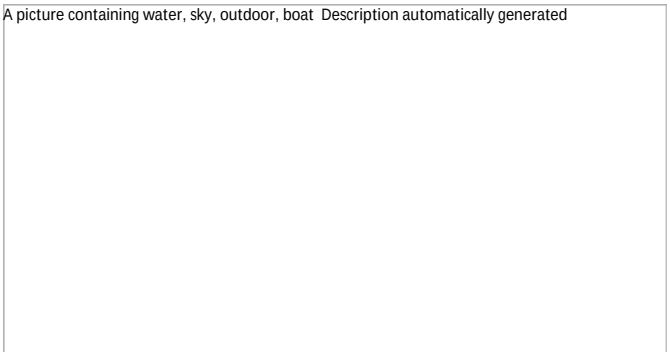
Finally, correlation is not necessarily a linear progression of one factor being correlated to changes in another. While club size may be related to inter-club competitive success, if the clubs are attracting the best participants because of these successes, this will in turn lead to more success. So, the correlation of club success and membership size hides a more complex relationship, in this case a cyclical one.

All of these reasons men that while there may be a relationship between the two factors which we can observe, we cannot say that one causes the other, so there is no "causal" relationship.

Causation: When a change in one variable affects the outcome of another variable. Eg., Clubs which pay for professional coaches will do better at inter-club competitions. Professional coaches directly improve the performance of the club membership, leading to an improvement in its performance in inter-club contests.

The difference between correlation and causation boils down to; if we can prove that changing one factor leads to a change in another factor. This requires the elimination of all other factors which might influence the relationship, alongside a good understanding of the relationship in question, which recurs across time or geography and is not due to chance.

As a result, it is unlikely that anyone measuring the populations of sports clubs will be able to say definitively that there is a causal relationship between two or more variables in the same way an experimental scientist might be able to. This is not a problem because simply knowing that one factor is related to another can help up plan for the future.



How can we use correlation and causation?

To go back to our example, if we track the population of our club and can say that there is a definite correlation between the total membership of our club and the total population of the local county then this will help us plan. In our example this correlation is positive, so when the local population increases our club membership increases, while if the local population decreases our club membership decreases. There may well be a causal relationship between these two factors, but because we cannot prove it, so we cannot treat the relationship between the two as fixed. This will help to remind us to plan for contingencies should this relationship prove to be less stable than we might hope.

If the local county population is increasing, then maybe we should plan into our development strategy a way of absorbing the anticipated increase in our membership. The reverse might also be true, if the local population is declining then maybe our focus should instead shift to retaining our members. We should also bear in mind though that the increases or declines in membership might not appear. We should also have a contingency plan in place for in case nothing or the opposite of what we expect happens. We might plan for an increase in membership, so we put more time and effort into increasing our capacity, while also making sure that we do not dismiss or ignore strategies to retain those members we already have. And vice versa, we might push hard to retain our members, while also understanding that we might need to increase our capacity if we start to recruit large numbers of new members.

By contrast, if the relationship between the local population and our club membership is causal and we know this because of a series of experiments in the area (Scientist *et al*., 2020) then our contingency planning can account for the known relationship between the variables. Therefore, less time and effort can be put into our contingency plans when we are in circumstances in which we know that the relationship holds, while they could be increased if we move beyond the conditions in which our scientists studied. This circumstance may be rare but would be hugely beneficial for the club.

To conclude, in this post we see the theoretical power that understanding the difference between correlation and causation in population data and how this might feed into our future planning strategies. If we know that a relationship is correlation we might put more time and effort into our contingency planning, compared to if we knew that the relationship between these variables was causal when we can be more specific about the planning**.**

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

[1] Dirky Gently to Mrs Rawlinson, Dirk Gently's Holistic detective agency, Douglas Adams, P132, 1987. Pan Macmillian 2012 edition, London.